STAT 425: Introduction to Nonparametric Statistics **Winter 2018**

Lecture 8: Density Estimation: Parametric Approach

Instructor: Yen-Chi Chen

8.1 Parametric Method

So far, we have learned several nonparametrc methods for density estimation. In fact, we can use a simple parametric method for density estimation.

We will start with a simple example by assuming the data is from a Gaussian (Normal) distribution. Recall that we observe

$$X_1, \cdots, X_n \sim P,$$

where P is the underlying population CDF and it has a PDF p. If we fit a Gaussian distribution to the data, we need to find the two parameters of Gaussian: the mean μ and the variance σ^2. While there are many approaches for estimating them (e.g., method of moments, or maximum likelihood method), we use a very simple estimator: the sample mean and sample variance.

Let

$$\widehat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \widehat{\sigma}_n^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

be the sample mean and sample variance. Then our density estimator is

$$\widehat{p}_n(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma}_n^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2}.$$

8.1.1 Analysis of Parametric Method

Is the parametric approach a good one? It depends. If the true PDF p is close to a Gaussian distribution, then probably the parametric approach is a good one. But if p is very far away from being a Gaussian, this method is going to give us a huge bias. Now we analyze the quality of estimation in the parametric approach. The goal is to quantify $\widehat{p}_n(x) - p(x)$.

Because the sample mean $\widehat{\mu}_n \xrightarrow{P} \bar{\mu} = \mathbb{E}(X_1)$ and the sample variance $\widehat{\sigma}_n^2 \xrightarrow{P} \bar{\sigma}^2 = \mathsf{Var}(X_1)$, we define another density

$$\bar{p}(x) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\bar{\mu})^2}.$$

This is the density generated by the Normal distribution with the mean being the population mean and the variance being the population variance. *It is NOT the actual population PDF.* Namely, $\bar{p}(x) \neq p(x)$ in general.

Using $\bar{p}(x)$, we obtain

$$\widehat{p}_n(x) - p(x) = \widehat{p}_n(x) - \bar{p}(x) + \bar{p}(x) - p(x). \qquad (8.1)$$

The first difference $\widehat{p}_n(x) - \bar{p}(x)$ is something that converges to 0 because the sample mean and variance converges to their population counterparts. Namely, we have

$$\widehat{p}_n(x) \xrightarrow{P} \bar{p}(x).$$

Note that this is due to the continuous mapping theorem. However, the second difference $\bar{p}(x) - p(x)$ never goes to 0 unless the the true PDF is Gaussian.

In what follows we study the convergence rate of $\widehat{p}_n(x) - \bar{p}(x)$. This will help us understand when a parametric approach may be better than a nonparametric approach. Recall two facts in parameter estimation:

$$\begin{aligned} \widehat{\mu}_n - \bar{\mu} &= O_P(1/\sqrt{n}) \\ \widehat{\sigma}_n^2 - \bar{\sigma}^2 &= O_P(1/\sqrt{n}). \end{aligned}$$

$$\begin{aligned} \widehat{p}_n(x) - \bar{p}(x) &= \frac{1}{\sqrt{2\pi\widehat{\sigma}_n^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\bar{\mu})^2} \\ &= \underbrace{\frac{1}{\sqrt{2\pi\widehat{\sigma}_n^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2}}_{(A)} + \underbrace{\frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\widehat{\mu}_n)^2}}_{(B)} \\ &\quad + \underbrace{\frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\bar{\mu})^2}}_{(C)} \end{aligned}$$

For the part (A),

$$\begin{aligned} (A) &= \frac{1}{\sqrt{2\pi\widehat{\sigma}_n^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} \left(\frac{1}{\widehat{\sigma}_n^2} - \frac{1}{\bar{\sigma}^2} \right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} \left(\frac{\bar{\sigma}^2 - \widehat{\sigma}_n^2}{\widehat{\sigma}_n^2 \bar{\sigma}^2} \right) \\ &= O_P(1/\sqrt{n}). \end{aligned}$$

For the part (B), we will need to use an approximation: for any $C > 0$ and $\epsilon_n \to 0$,

$$1 - C^{\epsilon_n} = 1 - e^{\epsilon_n \cdot \log C} = \epsilon_n \cdot \log C + \frac{(\epsilon_n \cdot \log C)^2}{2!} + \cdots = O(\epsilon_n). \tag{8.2}$$

$$\begin{aligned} (B) &= \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{1}{2\bar{\sigma}^2}(x-\widehat{\mu}_n)^2} \\ &= \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \left\{ e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} - e^{-\frac{1}{2\bar{\sigma}^2}(x-\widehat{\mu}_n)^2} \right\} \\ &= \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \cdot e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} \left\{ 1 - e^{-[\frac{1}{2\bar{\sigma}^2}(x-\widehat{\mu}_n)^2 - \frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2]} \right\} \\ &= \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \cdot e^{-\frac{1}{2\widehat{\sigma}_n^2}(x-\widehat{\mu}_n)^2} \underbrace{\left\{ 1 - e^{-[\frac{1}{2}(x-\widehat{\mu}_n)^2][\frac{1}{\bar{\sigma}^2} - \frac{1}{\widehat{\sigma}_n^2}]} \right\}}_{= O\left(\frac{1}{\bar{\sigma}^2} - \frac{1}{\widehat{\sigma}_n^2}\right) \text{ by equation (8.2)}} \\ &= O\left(\frac{1}{\bar{\sigma}^2} - \frac{1}{\widehat{\sigma}_n^2} \right) \\ &= O_P(1/\sqrt{n}). \end{aligned}$$

Note that in the above analysis, we treat $\widehat{\mu}_n$ as nonrandom but it is actually a random quantity. To take into account this, we can just replace the O's by O_P's.

Similarly, we can expand part (C) and this will lead us the same rate, i.e,

$$(C) = O_P(1/\sqrt{n}).$$

Thus, we conclude

$$\widehat{p}_n(x) - \bar{p}(x) = O_P(1/\sqrt{n}).$$

8.2 Mixture Model

A problem of parametric model is that the bias $\bar{p}(x) - p(x)$ is unavoidable so even we have huge amount of observations from the same population, we are still unable to recover the original PDF. However, the parametric model has an advantage that each parameter has its own meaning so it is very easy to interpret the result. In addition, as our analysis has shown, a parametric model has a convergence rate $O_P(1/\sqrt{n})$, which is often faster than a nonparametric estimator (the optimal rate is $O_P(1/n^{2/5})$). Thus, in many situations we would like to stick with parametric models.

If we want to use a parametric model, how can we resolve the problem of unavoidable bias? Here is a method that can alleviate this bias – mixture of distributions.

The mixture of distributions is using a mixture of parametric distribution to model the underlying population PDF. For instance, the famous Gaussian mixture model (GMM) uses

$$p_{\mathsf{GMM}}(x) = \sum_{\ell=1}^{K} \pi_\ell \cdot \phi(x; \mu_\ell, \sigma_\ell^2),$$

where $\pi_\ell \geq 0$ are weights with $\sum_{\ell=1}^{K} \pi_\ell = 1$ and

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

is the PDF of a standard normal distribution. Namely, GMM uses a mixture of K Gaussians to model the population PDF. Here the number K is a tuning parameter that specifies the number of Gaussians in our model.

In GMM, each $(\pi_\ell, \mu_\ell, \sigma_\ell^2)$ is a set of 3 parameters and we have one constraint: $\sum_{\ell=1}^{K} \pi_\ell = 1$, so there will be totally $3K - 1$ parameters. Estimation of these parameters are often done by the MLE (maximum likelihood estimator), namely,

$$\widehat{\pi}_1, \widehat{\mu}_1, \widehat{\sigma}_1^2, \cdots, \widehat{\pi}_K, \widehat{\mu}_K, \widehat{\sigma}_K^2 = \underset{\pi_\ell, \mu_\ell, \sigma_\ell^2 : \ell = 1, \cdots, K}{\mathsf{argmax}} \sum_{i=1}^{n} \log \left(\sum_{\ell=1}^{K} \pi_\ell \cdot \phi(X_i; \mu_\ell, \sigma_\ell^2) \right)$$

And the final density estimator is

$$\widehat{p}_{\mathsf{GMM}}(x) = \sum_{\ell=1}^{K} \widehat{\pi}_\ell \cdot \phi(x; \widehat{\mu}_\ell, \widehat{\sigma}_\ell^2),$$

GMM is easy to interpret – each component is like a hidden *ideal* distribution. So our data can be viewed as coming from a population with K hidden sub-populations. The parameter π_ℓ is the proportion of ℓ-th sub-population and μ_ℓ and σ^2_ℓ are the center and variation of the ℓ-th sub-population. Moreover, a GMM can well-approximate a complicated distribution when K is large. Namely, under good conditions, the bias $\mathbb{E}(\widehat{p}_{\mathsf{GMM}}(x)) - p(x)$ converges to 0 when $K \to \infty$.

Although the GMM or other mixture model have these good advantages, there are three serious issues about them.

- **Identifiability problem.** When $K \geq 2$, we may have non-unique MLE. Namely, there might be different sets of parameters that lead to the same distribution. Consider a simple case where $p(x) = 0.3\phi(x; 0, 1) + 0.7\phi(x; 2, 1)$. There are two equivalent representation for the same PDF:

$$(\pi_1, \mu_1, \sigma^2_1, \pi_2, \mu_2, \sigma^2_2) = (0.7, 0, 1, 0.3, 2, 1) \text{ or } (0.3, 2, 1, 0.7, 0, 1).$$

- **Computation problem.** Even we do not have identifiability issue, the MLE often does not have a closed-form solution so we need to use a numerical method such as a gradient descent/ascent approach[1] or the EM algorithm[2] to find the MLE. However, the likelihood function being optimized is often non-convex and has many local optima. Thus, there is no simple approach that guarantees that what we obtained from a computational algorithm is the actual MLE.
- **Choice of K.** The quantity K plays a key role in a mixture model and it acts as the tuning parameter in our model. However, unlike the tuning parameters in nonparametric estimation (e.g., smoothing bandwidth, number of nearest neighbor, number of basis) that we have theories about the optimal choice, the effect of K on the quality of estimation is very complicated and there is no simple form of it. Thus, choosing K turns out to be a more difficult task than the tuning parameter selection problem in nonparametric method. One may use a model selection technique[3] to choose it; we will discuss this at the end of this quarter.

8.3 Density Estimation: Final Comments

Here is a comparison among all the density estimators we have introduced so far:

Type	Method	Convergence rate	Tuning parameter	Limitation
Parametric	Parametric model	$O\left(\frac{1}{\sqrt{n}}\right)$	None	Unavoidable bias
	Mixture model	$O\left(\frac{1}{\sqrt{n}}\right)$	K, number of mixture	Hard to compute
Nonparametric	Histogram	$O\left(\frac{1}{n^{1/3}}\right)$	b, bin size	Lower convergence rate
	Kernel density estimator	$O\left(\frac{1}{n^{2/5}}\right)$	h, smoothing bandwidth	
	K-nearest neighbor	$O\left(\frac{1}{n^{2/5}}\right)$	k, number of neighbor	
	Basis approach	$O\left(\frac{1}{n^{2/5}}\right)$	M, number of basis	

Note that there are far more other density estimators but due to the time constraint, we cannot cover others. To understand the performance of a density estimator, we often analyze its mean integrated square error (MISE). The MISE can often be written as a bias and a variance part and we can often get a good sense on how to choose the tuning parameter by optimizing the MISE.

[1] https://en.wikipedia.org/wiki/Gradient_descent

[2] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

[3] https://en.wikipedia.org/wiki/Model_selection