STATS205: Introduction to Nonparametric Statistics

Lecturer: Tengyu Ma
Scribe: Nicholas K. Branigan and Andrew Kirjner

Lecture # 8
May 21, 2021

8.1 Review and overview

In the last lecture, we began our treatment of nonparametric density estimation with a discussion of the histogram algorithm. This intuitive approach involves constructing a histogram from our observations and normalizing it so that it constitutes a valid density function. In particular, we discuss about bias-variance trade-off:: as the bandwidth increases, the variance decreases while variance increases, and as we consider more examples, variance decreases but the bias, given its dependence on the expectation of the algorithm's predictions, does not change with the number of examples.

In this lecture, we move to kernel density estimation, a more sophisticated technique for this problem. Then, we briefly touch parametric and nonparametric mixture models and begin our discussion of Bayesian nonparametric methods.

8.2 Kernel density estimation

8.2.1 Introduction

We define the kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{8.1}$$

where h is the bandwidth and K is the kernel function. Recall from Lecture 1 that we have defined the kernel function to be any smooth, non-negative function K such that

$$\int_{\mathbb{R}} K(x)dx = 1, \quad \int_{\mathbb{R}} xK(x)dx = 0, \quad \text{and} \quad \int_{\mathbb{R}} x^2 K(x)dx > 0.$$

Two kernel functions we have seen are the boxcar and Gaussian kernels. For the former, we now show that kernel density estimation is very similar to the histogram approach. Recall that the boxcar kernel $K(x) = \frac{1}{2}\mathbf{1}\{|x| \leq 1\}$. Thus, using the boxcar, our kernel density estimator is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2}\mathbf{1}\left\{\left|\frac{x - x_i}{h}\right| \leq 1\right\}. \tag{8.2}$$

Define $B_x = \{i : |x_i - x| \leq h\}$ and let $|B_x|$ be the cardinality of this set, i.e., the number of points in B_x. Then we can write,

$$\begin{aligned}\hat{f}(x) &= \frac{1}{nh} \sum_{i \in B_x} \frac{1}{2} \mathbf{1} \left\{ \left| \frac{x - x_i}{h} \right| \leq 1 \right\} \\ &= \frac{1}{nh} \sum_{i \in B_x} \frac{1}{2} \\ &= \frac{|B_x|}{2nh}.\end{aligned}$$

To see the similarity with the histogram algorithm, recall that for the histogram,

$$\hat{f}(x) = \frac{\hat{p}_j}{h} = \frac{|B_j|}{nh}, \tag{8.3}$$

for $x \in B_j$. Moreover, note that for the histogram, h corresponds to the bin width, whereas for our boxcar density estimator, h is half of the bin width. The characteristic difference between the approaches is that for kernel density estimators, our bins are not fixed but moving with and centered at x.

Of course, we require that our kernel density estimator constitutes a valid density. There are two approaches for verifying that (8.1) coheres with the definition of the probability density function. The first is to check directly that (8.1) integrates to 1:

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}(x) dx &= \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right) dx \\ &= \frac{1}{nh} \sum_{i=1}^{n} \int_{\mathbb{R}} K\left(\frac{x - x_i}{h} \right) dx \\ &= \frac{1}{nh} \sum_{i=1}^{n} \int_{\mathbb{R}} K\left(\frac{x}{h} \right) dx \qquad (8.4) \\ &= \frac{1}{nh} \sum_{i=1}^{n} h \int_{\mathbb{R}} K(z)\, dz \qquad (8.5) \\ &= \frac{1}{nh} \sum_{i=1}^{n} h \\ &= 1.\end{aligned}$$

To reach (8.4), we used that shifting a function being integrated over the continuum by a constant has no effect on the value of the integral. In (8.5) we made a change of variables.[1]

The second approach for sanity checking our definition in (8.1) is to view $K(x)$ as a probability density function. Examining (8.2) confirms that this is a valid move. Then, we find that $\hat{f}$ is identical to a density function as desired. The following theorem formalizes this.

[1] Let $\mathcal{R}[a, b]$ denote the set of functions that are Riemann integrable on $[a, b]$. Then, let $f \in \mathcal{R}[a, b]$ and let g be a strictly increasing function from $[c, d]$ onto $[a, b]$ such that g is differentiable on $[c, d]$ and $g' \in \mathcal{R}[c, d]$. Then $(f \circ g) \cdot g' \in \mathcal{R}[c, d]$ and $\int_a^b f(x)dx = \int_c^d f(g(t))g'(t)dt$ [JP10].

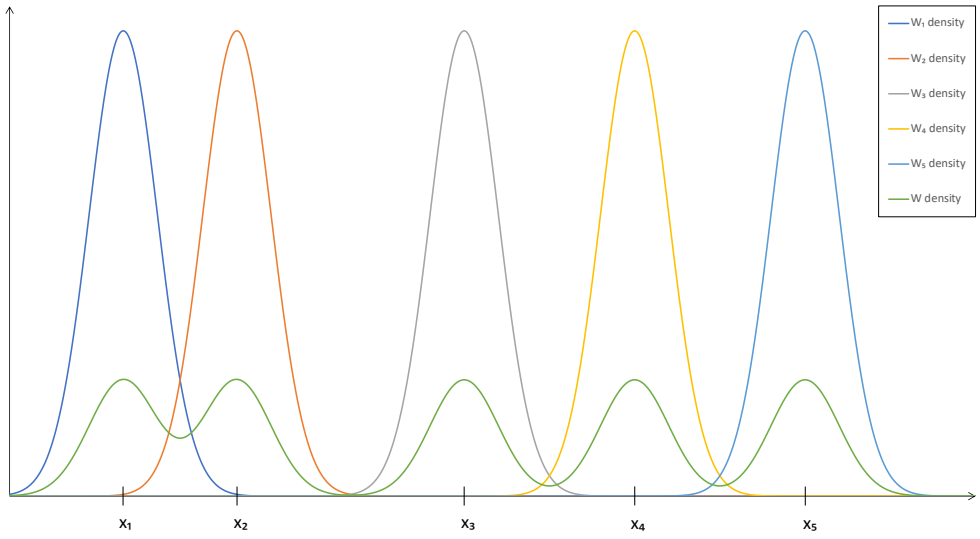


Figure 8.1: Example of Gaussian mixture model. In the figure above, $W = \frac{1}{5}\sum_{i=1}^{5} W_i$, where $W_i \sim \mathcal{N}(x_i, h^2)$. Per Theorem 8.1, the Gaussian kernel density estimator assumes that the density $\hat{f}$ is an equally weighted mixture of Gaussians centered on the observations $\{x_i\}_{i=1}^n$ with variance h^2.

Theorem 8.1. *Let $\xi \sim K(x)$, $Z \sim \text{Unif}\{x_1, \ldots, x_n\}$. Further, define $W = Z + \xi h$. Then, the density function of W is $\hat{f}$.*

Proof. Let $W_i = x_i + \xi h$ for each $i \in \{1, \ldots, n\}$. Finding the density of W_i is straightforward. To do this, we first find the distribution function of W_i:

$$\begin{aligned} F_{W_i}(x) &= P\{W_i \leq x\} \\ &= P\{x_i + \xi h \leq x\} \\ &= P\left\{\xi \leq \frac{x - x_i}{h}\right\} \\ &= F_\xi\left(\frac{x - x_i}{h}\right). \end{aligned}$$

Then, we differentiate to find the density:

$$\begin{aligned} f_{W_i}(x) &= \frac{d}{dx} F_{W_i}(x) \\ &= \frac{d}{dx} F_\xi\left(\frac{x - x_i}{h}\right) \\ &= f_\xi\left(\frac{x - x_i}{h}\right)\frac{1}{h} \\ &= \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \end{aligned}$$

Now, since $W = W_i$ with probability $\frac{1}{n}$, we can easily find the density of W. We again appeal to distribution functions to show this.

$$\begin{aligned} F_W(x) &= P\{W \leq x\} \\ &= \tfrac{1}{n} P\{W_1 \leq x\} + \cdots + \tfrac{1}{n} P\{W_n \leq x\} \\ &= \tfrac{1}{n} F_{W_1}(x) + \cdots + \tfrac{1}{n} F_{W_n}(x). \end{aligned} \tag{8.6}$$

Differentiating (8.6) gives that

$$\begin{aligned} f_W(x) &= \frac{1}{n} \sum_{i=1}^{n} f_{W_i}(x) \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \\ &= \hat{f}(x). \end{aligned}$$

(If W were not a mixture of random variables but a sum of them, computing its density would be far more complicated.) □

8.2.2 Integrated risk

Now, we turn to the risk associated with the Gaussian kernel density estimator.

Theorem 8.2. *For the Gaussian kernel density estimator the risk is*

$$R\left(f, \hat{f}\right) = \underbrace{\frac{1}{4}\sigma_k^2 h^4 \int f''(x)^2 dx}_{bias} + \underbrace{\frac{\beta_k^2}{nh}}_{\sigma^2} + \mathcal{O}\left(h^6\right) + \mathcal{O}\left(n^{-1}\right), \tag{8.7}$$

where $\sigma_k^2 = \int x^2 K(x) dx$ *and* $\beta_k^2 = \int K(x)^2 dx$. *The first term in* (8.7) *is the bias and the second is the variance.*

Recall that for the histogram density estimator,

$$R\left(\hat{f}, f\right) = \underbrace{\frac{h^2}{12} \int f'(x)^2 dx}_{\text{bias}} + \underbrace{\frac{1}{nh}}_{\sigma^2} + \mathcal{O}(h^2) + \mathcal{O}(n^{-1}). \tag{8.8}$$

Comparing (8.7) and (8.8), we see that for the Gaussian kernel density estimator, we want $f''(x)$ to be small rather than $f'(x)$. More importantly, for values of $h < 1$, the bias of the Gaussian kernel density estimator will be lower than for the histogram estimator. We encounter the usual bias-variance tradeoff here: increasing h results in more smoothing which boosts bias and depresses variance whereas decreasing h results in less smoothing which depresses bias and boosts variance.

By minimizing the risk with respect to h, we find the optimal bandwidth

$$h^* = \left(\frac{\beta_k^2}{\sigma_k^2 A(f) n}\right)^{1/5}, \quad \text{where} \quad A(f) = \int f''(x)^2 dx. \tag{8.9}$$

As usual, the optimal bandwidth is inversely dependent on n. Plugging h^* into (8.7), we find that as a function of n, $R(f, \hat{f}) \propto \mathcal{O}(n^{-4/5})$. We observe that this is an improvement over the histogram estimator where as a function of n, $R(f, \hat{f}) \propto \mathcal{O}(n^{-2/3})$.

Now, we show that for the boxcar kernel density estimator, the bias is as claimed in (8.7). From (8.3), $\hat{f}(x) = \frac{|B_x|}{2nh}$, so

$$\begin{aligned}
\mathbb{E}[\hat{f}(x)] &= \frac{1}{2nh}\,\mathbb{E}[|B_x|] \\
&= \frac{n}{2nh}\int_{x-h}^{x+h} f(u)du \\
&= \frac{1}{2h}\int_{x-h}^{x+h} f(u)du \\
&= \frac{1}{2h}\int_{x-h}^{x+h} \left(f(x) + (u-x)f'(x) + \tfrac{1}{2}(u-x)^2 f''(x)\right) du + \text{higher order terms} \qquad (8.10) \\
&= \frac{1}{2h}\left(2hf(x) + f'(x)\int_{x-h}^{x+h}(u-x)du + f''(x)\int_{x-h}^{x+h}\tfrac{1}{2}(u-x)^2 du\right) + \text{higher order terms} \\
&= f(x) + \mathcal{O}(h^2)f''(x).
\end{aligned}$$

In (8.10), we have carried out a degree 2 Taylor expansion for f at x.[2] Thus, we find that the bias at x is

$$\left(f(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2 = \mathcal{O}(h^2)^2 f''(x)^2 = \mathcal{O}(h^4)f''(x)^2.$$

So, the total bias is $\mathcal{O}(h^4)\int f''(x)^2$, which agrees with (8.7) as desired.

8.2.3 Choosing h empirically

There are two approaches for selecting the optimal bandwidth parameter h^* for kernel density estimators: normal references and cross-validation.

Unfortunately, we cannot use (8.9) to directly pick h^*. Though we will know β_k^2, σ_k^2, and n, we will not know $A(f)$, which depends on the object of our estimation. Normal references assumes for the purpose of finding h^* that $f \sim \mathcal{N}(\mu, \tau^2)$. Then, when K is a Gaussian kernel, $h^* = 1.06\tau n^{-1/5}$. In practice, τ is commonly chosen to be $\min\{\hat{\tau}, \frac{Q}{1.34}\}$, where $\hat{\tau}$ is the sample standard deviation and Q is the interquartile range.[3] Note that we only make this normality assumption to choose h. If we believed that f were actually normally distributed, we would be better suited with a parametric density estimation approach.

Next, we examine how to use cross-validation to choose h. Cross-validation is somewhat trickier in the unsupervised setting, since we do not have labels to evaluate our performance against on

[2] A real-valued function f is said to be of class C^n on (a, b) if $f^{(n)}(x)$ exists and is continuous for all $x \in (a, b)$. Define $P_n(x) = f(c) + f^{(1)}(c)(x-c) + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n$. Let $f \in C^{n+1}$ on (a, b), and let c and d be any points in (a, b). Then Taylor's Theorem says that there exists a point t between c and d such that $f(d) = P_n(d) + \frac{f^{n+1}(t)}{(n+1)!}(d-c)^{n+1}$ [JP10].

[3] For some data, the interquartile range is the data's 75th percentile minus its 25th percentile.

held-out data. To address this challenge, we rewrite our integrated risk

$$\begin{aligned} R(f, \hat{f}) &= \int \left(f(x) - \hat{f}(x)\right)^2 \\ &= \int f(x)^2 dx - 2\int f(x)\hat{f}(x)dx + \int \hat{f}(x)dx. \end{aligned} \tag{8.11}$$

In (8.11), the first term is constant with respect to $\hat{f}$, so we are not concerned about it when choosing $\hat{f}$. The second term can be rewritten $-2\,\mathbb{E}_{X\sim f}[\hat{f}(x)]$, and with held-out data $x'_1, \ldots, x'_m \sim f$, we can compute a Monte Carlo estimator of the expectation:

$$\mathop{\mathbb{E}}_{X\sim f}\left[\hat{f}(x)\right] \approx \frac{1}{m}\sum_{i=1}^{m} \hat{f}(x'_i).$$

If we have insufficient data for a hold-out set, we can use cross-validation. Under leave-one-out and Monte Carlo we have

$$\mathop{\mathbb{E}}_{X\sim f}\left[\hat{f}(x)\right] \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i),$$

where $\hat{f}_{-i}$ denotes the estimator obtained using $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$. Finally, the third term can be directly computed. Thus, the leave-one-out cross validation score is defined

$$\hat{J}\left(\hat{f}\right) = \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i).$$

We would like an efficient way to find the leave-one-out loss. A naive approach to computing $\hat{J}$ could be quite expensive since it would require that we fit $\hat{f}$ n times. Fortunately, we can do better.

Theorem 8.3. *We can compute the leave-one-out cross validation score for a kernel density estimator $\hat{f}$ as*

$$\hat{J}\left(\hat{f}\right) = \frac{1}{hn^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K^*\left(\frac{x_i - x_j}{h}\right) + \frac{2}{nh}K(0) + \mathcal{O}\left(n^{-2}\right), \tag{8.12}$$

where $K^*(x) = \int K(x-y)K(y)dy - 2K(x)$.

8.3 Mixture models

8.3.1 Introduction

From Theorem 8.1, we see that kernel density estimators recover the density of a random variable that is a mixture of n distributions of the same form as the kernel centered on the data points. In this section, we present a more parametric approach by examining mixture models where the number of distributions in the mixture is $k < n$. We assume that our density

$$f(x) = \frac{1}{k}\sum_{i=1}^{k} f_i(x) \tag{8.13}$$

for some distributions $\{f_i \mid i \in \{1, \ldots, k\}\}$. There is an equivalent generative specification of this model:

1. Draw i from some distribution over $\{1, \ldots, k\}$. A simple choice that we have used in (8.13) and that we will use going forward is $i \sim \text{unif}\{1, \ldots, k\}$.
2. Draw $x \sim f_i$.

8.3.2 Gaussian mixtures and model fitting

A popular mixture model is the Gaussian mixture. Under this model, $f_i(x) = \mathcal{N}(\mu_i, \Sigma_i)$ for $\mu_i \in \mathbb{R}^d$, $\Sigma_i \in \mathbb{R}^{d \times d}$.

$$f(x; \mu_1 \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1} (x-\mu_i)}. \tag{8.14}$$

This Gaussian mixture model is a fully parametric approach since k is fixed. We can make it less parametric by letting k grow with n in some way. There are three algorithms commonly used for fitting mixtures: maximum likelihood estimation (MLE), the Expectation Maximization algorithm (EM), and the method of moments.

Let $\theta = (\mu_1 \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k, z_1, \ldots, z_n)$, where the z_i's are in $\{1, \ldots, k\}$ and denote the Gaussians to which the observations are assigned. MLE amounts to solving the optimization problem

$$\max_{\theta} \frac{1}{n} \sum_{j=1}^{n} \log f(x_j; \mu_{z_j}, \Sigma_{z_j}). \tag{8.15}$$

This is often impossible to do analytically, so numerical methods are frequently required. EM is beyond the scope of the class, but it can be applied to fit mixtures under the MLE approach or even the more general Bayesian framework. The method of moments involves relating model parameters to the moments of random variables. Recall that for random variables x_i, $i \in \{1, \ldots, d\}$, the first moments are

$$\mathbb{E}[x_i] \text{ for each } i,$$

the second moments are

$$\mathbb{E}[x_i x_j] \text{ for each } i, j,$$

and the third moments are

$$\mathbb{E}[x_i x_j x_k] \text{ for each } i, j, k.$$

We can estimate these using empirical moments. For example, for observations $x^{(1)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$, the empirical first moment for the i'th dimension of x is

$$\frac{1}{n} \sum_{j=1}^{n} x_i^{(j)} \approx \mathbb{E}[x_i].$$

If the moments are functions of the model parameters, we can exploit this to fit our model. For example, in the Gaussian mixture case, $\mathbb{E}[x_i] = \frac{1}{k}\sum_{j=1}^{k}(\mu_j)_i$. In general, suppose

$$\begin{aligned}\mathbb{E}[x_i] &= q_i(\mu, \Sigma)\\ \mathbb{E}[x_i x_j] &= q_{ij}(\mu, \Sigma)\\ &\vdots\end{aligned}$$

Then we can construct loss functions to minimize with respect to θ:

$$\left(\frac{1}{n}\sum_{j=1}^{n} x_i^{(j)} - q_i(\mu, \Sigma)\right)^2 \tag{8.16}$$

$$\left(\frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)} - q_{ij}(\mu, \Sigma)\right)^2 \tag{8.17}$$

$$\vdots$$

8.4 Bayesian nonparametric statistics

8.4.1 Review of the Bayesian approach

Under the Bayesian take on statistics, we treat our model parameter θ as a random variable, and we express our beliefs regarding θ prior to our statistical analysis through a distribution over θ called a prior. In the unsupervised setting the Bayesian approach assumes the following hierarchical model:

1. Draw $\theta \sim p(\theta)$.
2. Draw data $x^{(1)}, \ldots, x^{(n)} \overset{iid}{\sim} p(x \mid \theta)$.[4]

Our goal is to infer the posterior distribution $p\left(\theta \mid x^{(1)}, \ldots, x^{(n)}\right)$. To accomplish this, we use Bayes' rule:

$$\begin{aligned}p\left(\theta \mid x^{(1)}, \ldots, x^{(n)}\right) &= \frac{p\left(\theta, x^{(1)}, \ldots, x^{(n)}\right)}{p\left(x^{(1)}, \ldots, x^{(n)}\right)}\\ &= \frac{p\left(x^{(1)}, \ldots, x^{(n)} \mid \theta\right) p\left(\theta\right)}{p\left(x^{(1)}, \ldots, x^{(n)}\right)}\\ &= \frac{\prod_{i=1}^{n} p\left(x^{(i)} \mid \theta\right) p\left(\theta\right)}{\int \prod_{i=1}^{n} p\left(x^{(i)} \mid \theta\right) p\left(\theta\right) d\theta}.\end{aligned}$$

Now, in the supervised setting, suppose we have a dataset $S = \left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$, where the $x^{(i)}$'s are fixed. Then, our generative story takes the following form:

[4]Note that in this section, we drop density function subscripts for notational elegance. For example, to be as clear as possible, we would write $f_\theta(\theta)$ rather than $p(\theta)$ and $f_{x|\theta}(x \mid \theta)$ not $p(x \mid \theta)$. $p(\theta)$ and $p(x \mid \theta)$ are not the same function p. "Think of them as living things that look inside their own parentheses before deciding what function to be" [Owe18].

1. Draw $\theta \sim p(\theta)$.

2. Draw a label $y^{(i)} \sim p\left(y^{(i)} \mid x^{(i)}, \theta\right)$ for each $i \in \{1, \ldots, n\}$.

Given a test example x^*, we want to find $p(y^* \mid x^*, S)$, where y^* denotes the (unknown) label associated with x^*. We can do this if we can first infer the posterior $p(\theta \mid S)$. Why? Observe that

$$\begin{aligned} p(y^* \mid x^*, S) &= \int p(y^* \mid x^*, \theta, S) p(\theta \mid x^*, S) d\theta \\ &= \int p(y^* \mid \theta, x^*) p(\theta \mid S) d\theta, \end{aligned}$$

where we've used that y^* is independent of $y^{(1)}, \ldots, y^{(n)}$ conditional on θ. As in the supervised setting, we can use Bayes' rule to find an expression for the posterior that we can work with:

$$\begin{aligned} p(\theta \mid S) &= \frac{p(\theta, S)}{p(S)} \\ &= \frac{p(S \mid \theta) p(\theta)}{\int p(S \mid \theta) p(\theta) d\theta} \\ &= \frac{p\left(y^{(1)}, \ldots, y^{(n)} \mid \theta\right) p(\theta)}{\int p\left(y^{(1)}, \ldots, y^{(n)} \mid \theta\right) p(\theta) d\theta} \\ &= \frac{\prod_{i=1}^n p\left(y^{(i)} \mid \theta, x^{(i)}\right) p(\theta)}{\int \prod_{i=1}^n p\left(y^{(i)} \mid \theta, x^{(i)}\right) p(\theta) d\theta}. \end{aligned}$$

8.4.2 Bayesian linear regression

Let $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$, $\theta \in \mathbb{R}^d$ with $\theta \sim \mathcal{N}(0, \tau^2 I_d)$. We can simplify the density of θ

$$p(\theta) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-||\theta||_2^2/(2\tau^2)}. \tag{8.18}$$

We assume that $y^{(i)} = {x^{(i)}}^\top \theta + \epsilon^{(i)}$ where $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Our generative model then is

1. Draw $\theta \sim \mathcal{N}(0, \tau^2 I_d)$.

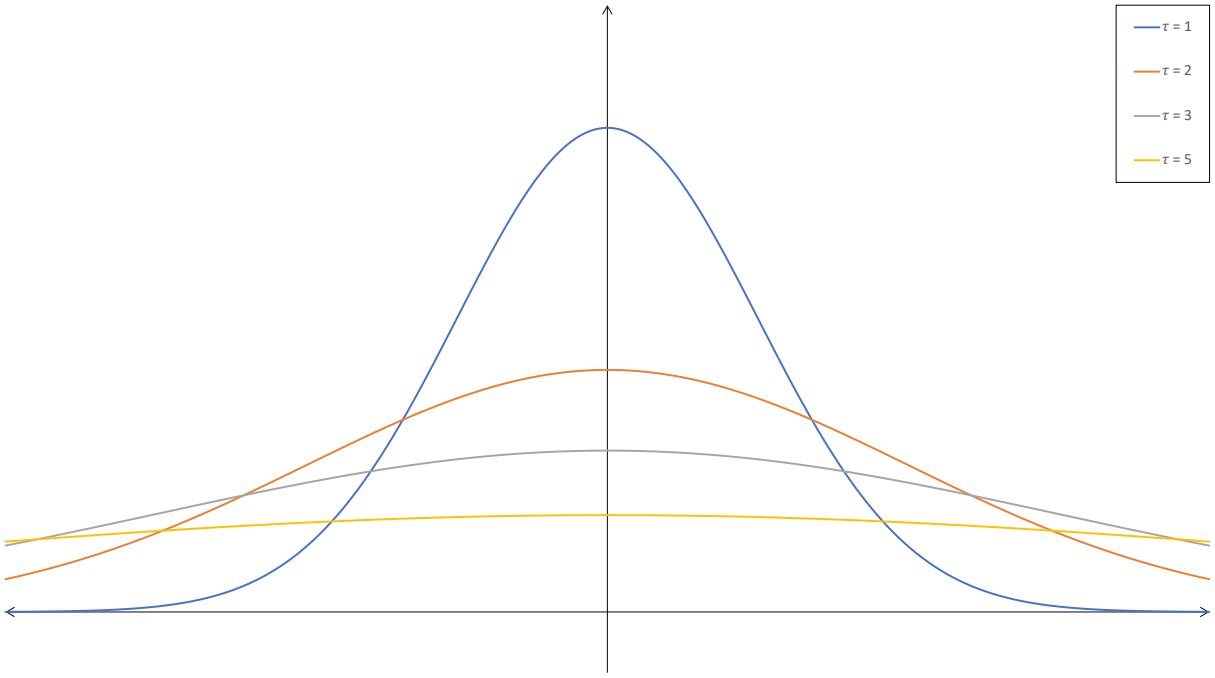
2. Draw $y^{(i)} \sim \mathcal{N}({x^{(i)}}^\top \theta, \sigma^2)$ for each $i \in \{1, \ldots, n\}$.

Theorem 8.4. *Define the design matrix*

$$X = \begin{bmatrix} {x^{(1)}}^\top \\ \vdots \\ {x^{(n)}}^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n, \quad \textit{and} \quad A = \frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I_d.$$

Then $\theta \mid S \sim \mathcal{N}(\frac{1}{\sigma^2} A^{-1} X^\top \vec{y}, A^{-1})$, *and* $y^* \mid x^*, S \sim \mathcal{N}(\frac{1}{\sigma^2} {x^*}^\top A^{-1} X^\top \vec{y}, {x^*}^\top A^{-1} x^* + \sigma^2)$.

Figure 8.2: Densities for $\mathcal{N}(0, \tau^2)$



Let's sanity check Theorem 8.4. We can rewrite A as

$$\underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n} x^{(i)}x^{(i)\top}}_{\text{influence of data}} + \underbrace{\frac{1}{\tau^2}I_d}_{\text{influence of prior}}. \tag{8.19}$$

First, as $n \to \infty$, the first term in (8.19) dominates the second term. As we would hope, as the size of our dataset grows the influence of the prior on the posterior of θ diminishes and, at the limit, vanishes. For this reason, Bayesian methods are less useful under a large data regime. Second, as $\tau \to \infty$, our Gaussian prior becomes increasingly flat and uninformative; see Figure 8.2. Accordingly, in (8.19), τ is inversely related to the influence of the prior on the posterior. Third, the variance of our posterior predictive distribution $y^* \mid x^*, S$ is at least σ^2.

Proof. First, we show that A is positive definite. For non-zero $z \in \mathbb{R}^d$,

$$\begin{aligned} z^\top \frac{1}{\sigma^2} X^\top X z &= \frac{1}{\sigma^2}(Xz)^\top Xz \\ &= \frac{1}{\sigma^2}\langle Xz, Xz, \rangle. \end{aligned}$$

Since X is full-rank (our predictors cannot be a linear combination of each other), X's null space is trivial and $Xz \neq 0$. Because σ^2 is positive and the norm is positive for all non-zero vectors, (**??**) is positive and A is positive definite. Then, A^{-1} is also positive definite since its eigenvalues are

the reciprocals of A's eigenvalues.[5] Thus,

$$x^{*\top} A^{-1} x^* + \sigma^2 \geq \sigma^2.$$

□

As expected, our uncertainty regarding our predictions is bounded below by σ^2, which is the uncertainty intrinsic to the problem. Moreover, as our dataset grows in observations, we approach this lower bound and achieve it at the infinite limit. As $n \to \infty$, $x^{*\top} A^{-1} x^* \to 0$, since as $n \to \infty$, $A \to \infty$.

[5] For any invertible matrix M, M^{-1}'s eigenvalues are the eigenvalues of M inverted. A matrix is positive definite if and only if all of its eigenvalues are positive [Axl14].

Bibliography

[Axl14] Sheldon Axler, *Linear algebra done right*, Springer, New York, 2014.

[JP10] Richard Johnsonbaugh and W. E. Pfaffenberger, *Foundations of mathematical analysis*, dover ed ed., Dover books on mathematics, Dover Publications, Mineola, N.Y, 2010, OCLC: ocn463454165.

[Owe18] Art Owen, *Lecture 6: Bayesian estimation*, October 2018, Unpublished lecture notes from STATS 200.