Nonparametrics & Bayes

Sofia Olhede

EPFL

November 2, 2020

1. Non–parametrics

2. Bayesian Statistics
   - Interval estimates

Some more set–up

- If we'd still like to estimate a parameter, can use the plug-in principle:
- Let $\nu = \nu(F)$ be a parameter of interest.
- We can use $\widehat{\nu} = \nu(\widehat{F}_n)$ as an estimator of $\nu(F)$, i.e. we plug in $\widehat{F}_n$ in $\nu(F)$.
- This is a "flipped" view of viewing ν as a function of F.
- Only sort of parameter we can consider, since no parametric model assumed!
- For the first two moments we get

$$\mu(F) = \int_{-\infty}^{\infty} y \, d\widehat{F}_n(y) = \frac{1}{n}\sum_i Y_i = \overline{Y} \tag{1}$$

$$\begin{aligned} \sigma^2(F) &= \int_{-\infty}^{\infty} \{y - \mu(F)\}^2 \, d\widehat{F}_n(y) = \frac{1}{n}\sum_i \{Y_i - \mu(F)\}^2 \\ &= \frac{1}{n}\sum_i \{Y_i - \overline{Y}\}^2. \end{aligned} \tag{2}$$

Non–Parametric Statistics Cont'd

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - $\hookrightarrow$ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - $\hookrightarrow$ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - $\hookrightarrow$ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - $\hookrightarrow$ Parameters usually interpretable, CDFs are harder to appreciate visually.
- Densities are more easily interpreted – also defined as functional of CDF!
- The density f (when it exists) at $x_0 \in \mathbb{R}$ is $\nu(F) := \frac{d}{dx}F(x)\Big|_{x=x_0}$
- Caution: mapping $F \mapsto \nu(F)$ not a "well behaved" mapping in general...

Non–Parametric Statistics Cont'd

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\Big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a "well-behaved" mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, ..., Y_n\}$ estimator is undefined!
- Problem is that estimator requires differentiation of a function $\hat{F}_n$ with jumps
- We will need a "smoother" estimate of F to plug in instead of $\hat{F}_n$, e.g.

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} \Phi\left(\frac{x-y}{h}\right) d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x-Y_i}{h}\right)$$

  for Φ a standard normal CDF and $h > 0$ a smoothing parameter.
- Transforms flat steps with hard corners to inclined steps with smooth corners (buffs the edges)

Non–Parametric Statistics Cont'd

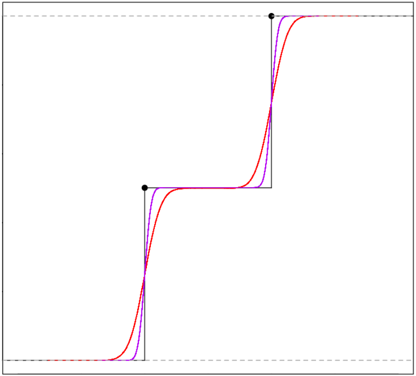

Figure: Empirical distribution function (black) for a size $n = 2$ sample, and "smoothed" approximations by convolution with $\Phi\left(\frac{u}{h}\right)$ for $h = 0.3$ (red) and $h = 0.2$ (purple).

Non–Parametric Statistics Cont'd

At the level of density, this yields the "smoothed plug-in estimator"

$$\hat{f}(x) = \frac{d}{dx}\tilde{F}_n(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-Y_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\varphi\left(\frac{x-Y_i}{h}\right)$$

for $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ the standard normal density.

- Nothing special about choice of φ – can choose any smooth unimodal probability density K that is symmetric about zero and has variance 1.
  ↪ Call such a K a kernel.
- Much more important is the choice of $h > 0$ called a *bandwidth* or *smoothing parameter*.

Definition (Kernel Density Estimator)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, where f is a probability density function. A *Kernel Density Estimator* (KDE) $\hat{f}$ of f is a random density function defined as

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

for $K : \mathbb{R} \to \mathbb{R}$ a *kernel* and $h > 0$ a *bandwidth* or *smoothing parameter*.

Non–Parametric Statistics Cont'd

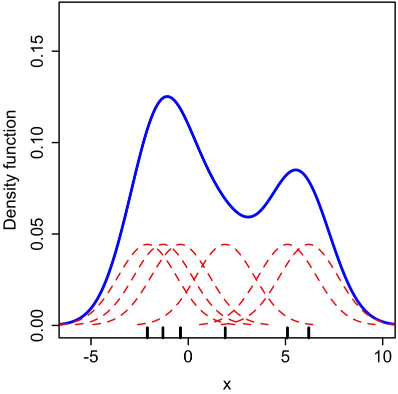

Figure: Schematic Illustration of a kernel density estimator

Non–Parametric Statistics Cont'd

Only problem: how should we choose arbitrary tuning parameter $h > 0$?

$\hookrightarrow$ Can have decisive effect on quality of estimator.

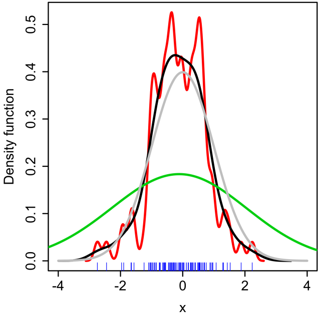

Figure: Effect of bandwidth choice on KDE of standard normal density, $n = 100$. True density in gray.KDE with: $h = 0.05$ in red, $h = 0.337$ in black, $h = 2$ in green.

Non–Parametric Statistics Cont'd

To select h, need to understand its effect on KDE.

In short, it regulates the bias-variance tradeoff:

- Large h: gives "flattened" estimator (higher bias) but quite stable to small perturbations of the sample values (low variance).
- Small h: gives "wiggly" estimator (lower bias) but overly sensitive to small perturbations of the sample values (high variance).

What bias and variance? Those corresponding to integrated mean squared error:

$$\mathrm{IMSE}(\hat{f}, f) = \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}(x) - f(x)\Big)^2 dx.$$

$$\mathrm{IMSE}(\hat{f}, f) = \underbrace{\int_{\mathbb{R}} \left(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\right)^2 dx}_{\text{integrated squared bias}} + \underbrace{\int_{\mathbb{R}} \mathbb{E}\left\{\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right\}^2 dx}_{\text{integrated variance}}$$

To get a useful expression for this we resort to asymptotics.

Non–Parametric Statistics Cont'd

Theorem (Asymptotic Risk of KDE)

Let $f \in C^3$ be a probability density and $K \in C^2$ a kernel function satisfying

$$\int_{\mathbb{R}} \left(f''(x)\right)^2 dx < \infty \quad \int_{\mathbb{R}} \left|f'''(x)\right| dx < \infty \quad \& \quad \int_{\mathbb{R}} \left(K''(x)\right)^2 dx < \infty.$$

If $\hat{f}_n$ is the KDE of f with iid sample size n, kernel K and bandwidth h,

$$\mathrm{IMSE}(\hat{f}, f) = \frac{h^4}{4} \int_{\mathbb{R}} \left(f''(x)\right)^2 dx + \frac{1}{nh} \int_{\mathbb{R}} K^2(x) dx + o\left(h^4 + \frac{1}{nh}\right).$$

as $h \to 0$.

Conclusions:

- For consistency, need $h \to 0$ but $nh \to \infty$ as $n \to \infty$.
- Optimal choice of h will unfortunatelydepend on (unknown) f''
- For the record, optimal h is given (after some calculations) by

$$h^* = \left\{ \frac{1}{n} \int_{\mathbb{R}} K^2(x) dx \Big/ \int_{\mathbb{R}} \left(f''(x)\right)^2 dx \right\}^{1/5}$$

- Plugging in the optimal bandwidth yields the a risk of asymptotic order $n^{-4/5}$
- Compare this to parametric model optimal rate of n^{-1}
- Asymptotic bias proportional to curvature of f.

Non–Parametric Statistics Cont'd

Proof (∗).

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right]=\frac{1}{h}\int_{\mathbb{R}}K\left(\frac{x-t}{h}\right)f(t)dt=\int_{\mathbb{R}}K(y)f(x-hy)dy$$

by change of variables $y=(x-t)/h$. Now Taylor expanding f yields

$$f(x-hy)=f(x)-hyf'(x)+\frac{1}{2}h^2y^2f''(x)+o(h^2)\quad \text{as } h\to 0.$$

Plugging into the equation for the expectation, we get that $\mathbb{E}[\hat{f}_n(x)]$ equals

$$f(x)\underbrace{\int_{\mathbb{R}}K(y)dy}_{=1}-hf'(x)\underbrace{\int_{\mathbb{R}}yK(y)dy}_{=0}+\frac{1}{2}h^2f''(x)\underbrace{\int_{\mathbb{R}}y^2K(y)dy}_{=1}+o(h^2)$$

as $h\to 0$ by the kernel properties of K. In summary the pointwise bias is

$$\mathbb{E}\left[\hat{f}_n(x)\right]-f(x)=\frac{1}{2}h^2f''(x)+o(h^2),\quad \text{as } h\to 0.$$

Non–Parametric Statistics Cont'd

The pointwise variance $\mathrm{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left[K\left(\frac{x-Y_i}{h}\right)\right]=\frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right]-\mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

and by similar manipulations as earlier, and the expression for $\mathbb{E}\left[\hat{f}_n(x)\right]$, we get

$$\mathrm{var}[\hat{f}_n(x)]=\underbrace{\frac{1}{nh}\int_{\mathbb{R}}K^2(y)f(x-hy)dy}_{A}-\underbrace{\frac{1}{nh^2}\,\mathbb{E}^2[\hat{f}_n(x)]}_{B}$$

Now observe that as as $h\to 0$, we have

$$B=\frac{1}{nh^2}(f(x)+\frac{1}{2}h^2f''(x)+o(h^2))^2=\frac{1}{nh^2}[f(x)+o(h)]^2=o\left(\frac{1}{n}\right).$$

On the other hand, Taylor expanding $f(x-hy)=f(x)+o(1)$ as $h\to 0$, we have

$$A=\tfrac{1}{nh}\int_{\mathbb{R}}K^2(y)[f(x)+o(1)]dy=\tfrac{1}{nh}f(x)\int_{\mathbb{R}}K^2(y)dy+o\left(\tfrac{1}{nh}\right)$$

since $\frac{1}{nh}o(1)=o\left(\frac{1}{nh}\right)$

Non–Parametric Statistics Cont'd

Putting A and B together gives

$$\mathrm{var}[\hat{f}_n(x)] = \frac{1}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{f(x)}{nh}\right) - o\left(\frac{1}{n}\right) = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{1}{nh}\right)$$

Summing pointwise squared-bias and variance, the pointwise MSE is given by

$$MSE(\hat{f}_n(x), f(x)) = \frac{1}{4}h^4(f''(x))^2 + \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(h^4 + \frac{1}{nh}\right)$$

Finally, integrating over $\mathbb{R}$ and re-arranging yields the sought form

$$\mathrm{IMSE}(\hat{f}, f) = \frac{1}{nh}\int_{\mathbb{R}} K^2(x)dx + \frac{h^4}{4}\int_{\mathbb{R}} \left(f''(x)\right)^2 dx + o\left(h^4 + \frac{1}{nh}\right).$$

$\square$

Non–Parametric Statistics Cont'd

Can we do better than $n^{-4/5}$ by more smoothness assumptions?

Theorem (Minimax Optimal Rates for KDE)

Let $\mathcal{F}(m, r)$ be the subset of m-differentiable densities with mth derivative in an L^2 ball of radius r,

$$\int_{\mathbb{R}} \left(f^{(m)}(x)\right)^2 dx \leq r^2.$$

Then, given any KDE $\hat{f}_n$,

$$\sup_{f \in \mathcal{F}(m,r)} \mathbb{E}\left\{\int_{\mathbb{R}} \left(\hat{f}_n(x) - f(x)\right)^2 dx\right\} \geq C n^{-\frac{2m}{2m+1}},$$

where the constant $C > 0$ depends only on m and c.

- The smoother the density the better the worst case rate.
- Can never beat n^{-1}, though.
- The price to pay for flexibility!

Non–Parametric Statistics Cont'd

So how do we choose h in practice? Here's a couple of approches:

- Pilot estimator: use a parametric family (e.g. normal, or mixture) to obtain a preliminary estimator $\check{f}$, and plug this into the optimal bandwidth expression to select a bandwidth.

- Least squares cross-validation: try to construct an unbiased estimator of the IMSE after all, it is an expectation. Then choose h to minimise the estimated IMSE. Also known as unbiased risk estimation.

Let's consider the second approach in more detail. Notice that we can write

$$\begin{aligned} IMSE(\hat{f}_h, f) &= \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}_h(x) - f(x)\Big)^2 dx = \mathbb{E}\left[\int_{\mathbb{R}} \Big(\hat{f}_h(x) - f(x)\Big)^2 dx\right] \\ &= \underbrace{\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)\,dx\right] - 2\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h(x) f(x)\,dx\right]}_{H(\hat{f}_h)} + \mathbb{E}\left[\int_{\mathbb{R}} f^2(x)\,dx\right]. \end{aligned}$$

where the last term does not vary with h.

Non–Parametric Statistics Cont'd

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)}\sum_{j\neq i} K\left(\frac{x-Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\left[\hat{f}_{h,-i}(Y_i)\right] = \frac{1}{n-1}\sum_{j\neq i}\mathbb{E}\left[\frac{1}{h}K\left(\frac{Y_i-Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h}K\left(\frac{Y_1-Y_2}{h}\right)\right] =$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{u-v}{h}\right)f(u)f(v)dudv = \int_{\mathbb{R}}\mathbb{E}\left[\frac{1}{h}K\left(\frac{Y_1-v}{h}\right)\right]f(v)dv =$$

$$= \int_{\mathbb{R}}\mathbb{E}\left[\frac{1}{nh}\sum_{k=1}^{n}K\left(\frac{Y_k-v}{h}\right)\right]f(v)dv = \mathbb{E}\left[\int_{\mathbb{R}}\underbrace{\frac{1}{nh}\sum_{k=1}^{n}K\left(\frac{Y_k-v}{h}\right)}_{=\hat{f}_h(v)}f(v)dv\right]$$

Thus $\{\hat{f}_{h,-i}(Y_i)\}_{i=1}^n$ are n variables with mean $\mathbb{E}\left[\int_{\mathbb{R}}\hat{f}_h(x)f(x)dx\right]$!

Non–Parametric Statistics Cont'd

Motivates definition of leave-one-out cross validation estimator

$$LSCV(h) = \int_{\mathbb{R}} \hat{f}_h^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(Y_i)$$

which by construction satisfies

$$\mathbb{E}[LSCV(h)] = H(\hat{f}_h).$$

Strategy: choose h by minimising $LSCV(h)$. Does it work?

Theorem (Stone's Theorem)

In the same context, and under the same assumptions, let h_{CV} denote the bandwidth selected by cross-validation. Then,

$$\frac{\int_{\mathbb{R}} \left(\hat{f}_{h_{CV}}(x) - f(x)\right)^2 dx}{\inf_{h>0} \int_{\mathbb{R}} \left(\hat{f}_h(x) - f(x)\right)^2 dx} \xrightarrow{a.s.} 1,$$

provided that the true density f is bounded.

Non–Parametric Statistics Cont'd

Conceptually, can generalise KDE very easily to higher dimensions.

- Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{iid}{\sim} f(\boldsymbol{y})$ be a sample in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to [0, +\infty)$
- Let $\boldsymbol{H} \succeq 0$ be a $d \times d$ symmetric positive-definite bandwidth matrix.
- Let K be a probability density on $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{I}_{d\times d}$.
  - ↪ E.g. $K(x_1, \ldots, x_n) = \prod_{j=1}^{d} \varphi(x_j)$ for φ the $N(0,1)$ density.

We can define a d–dimensional KDE as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{Y}_i)\right), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Once again choice of kernel is secondary but choice of $\boldsymbol{H}$ is paramount.

- Considerably harder: need to choose $d(d+1)/2 \sim d^2$ bandwidth parameters.
- Intuitively: $\boldsymbol{H} = \boldsymbol{U}\mathrm{diag}\{h_1, \ldots, h_d\}\boldsymbol{U}^\top$ for $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{d\times d}$ and $h_j > 0$.
  - ↪ Choose d smoothing directions, and a bandwidth for each such direction.
- LSCV-type solutions exist for d moderate (computationally intensive).
- Visualisation challenging for $d > 3$.

Some more set–up

- We have assumed that there is some parameter θ with some unknown constant value.
- We could think of the unknown parameter θ as being a realisation from random variable Θ where Θ has some supposed distribution $p(\Theta = \theta)$.
- The previous approach is a special case of this method with $p(\Theta = \theta_0) = 1$ and $p(\Theta \neq \theta) = 0$.
- We write

$$p(\mathcal{D}, \theta) = p(\mathcal{D}|\theta)p(\theta) = p(\theta|\mathcal{D})p(\mathcal{D}),$$

where $\mathcal{D} = (X_1, \ldots, X_n)$ and $p(\cdot)$ is either a pmf or pdf.

Some more Bayesian set–up

- This gives us

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta) \\ \text{Posterior} &= \text{Likelihood} \times \text{Prior}. \end{aligned}$$

- By Bayes' Theorem and note that $p(D)$ is **not** a function of θ and it is given by

$$p(D) = \int_{\Theta} p(D|\theta)p(\theta)d\theta.$$

- We write the likelihood with a conditional sign rather than a semi-colon to reflect the fact that θ is a random variable rather than a constant.
- Using Bayes theorem allows us to determine a posterior distribution for Θ which gives us all the available information about it after we have seen the data, D.

Some more Bayesian set–up

- We may want to report a single plausible value for Θ which summarises its posterior distribution.
- The "best" summary of the posterior, $\tilde{\theta}$, needs a loss function, $L(\Theta, \tilde{\theta})$, which ensures that the posterior density is concentrated near the point estimate, $\widetilde{\theta}$.
- We take $\widetilde{\theta}$ as the value that minimises the expected posterior loss so that

$$E_{\Theta|D}\{L(\Theta, \tilde{\theta})|D\} = \int_{-\infty}^{\infty} L(\theta, \tilde{\theta}) p(\theta|D) d\theta.$$

  is a minimum.
- If $L(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$ we take $\widetilde{\theta}$ to be the mean of the posterior distribution, $p(\theta|D)$.

Some Bayesian statistics

- If $L(\theta, \widetilde{\theta}) = |\theta - \widetilde{\theta}|$ we find that $\widetilde{\theta}$ is the posterior median.
- If we take $L(\theta, \widetilde{\theta}) = 1$ if $\widetilde{\theta} \neq \theta$ and zero otherwise we take $\widetilde{\theta}$ as the point at the maximum of the density, i.e. the posterior mode.
- In Bayesian statistics all the information is contained in the posterior pdf $p(\theta|D)$.
- We may want an interval which will contain Θ with probability that include the most concentrated areas of $p(\theta|D)$.
- We can do this by determining the $100\gamma\%$ credible interval which is an interval which contains $100\gamma\%$ of the total density in the posterior distribution.

Some Bayesian statistics

- Let $\ell(\mathrm{x})$ and $u(\mathrm{x})$ be some functions of the observed data then a $100\gamma\%$ credible interval satisfies

$$\begin{aligned} P(\ell(\mathrm{x}) < \Theta < u(\mathrm{x})|D) &= \int_{\ell x}^{u(x)} p(\theta|D)d\theta \\ &= \gamma. \end{aligned}$$

Further, we define the $100\gamma\%$ Highest Posterior Density (HPD) region to be the credible interval for which $u(\mathrm{x}) - \ell(\mathrm{x})$ is a minimum.

Gaussian example

- Assume that
$$Y_i|\mu \sim N(\mu, \sigma_0^2).$$
- We also put a prior distribution on μ of
$$\mu\,|\mu_0, \tau \sim N(\mu_0, \tau^2).$$

We can now from this calculate

$$\begin{aligned}
p(\mu|Y) &= \frac{p(Y|\mu)p(\mu)}{p(Y)} \propto p(Y|\mu)p(\mu) \\
&\propto \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(Y_i-\mu)^2} \frac{1}{\left(2\pi\tau^2\right)^{1/2}} e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2} \\
&\propto \exp\{-\frac{1}{2}\{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\}\mu^2 + \{\frac{n\bar{Y}}{\sigma_0^2}+\frac{\mu_0}{\tau^2}\}\mu\}. \qquad (3)
\end{aligned}$$

Gaussian example

- This is a Gaussian distribution on μ with a variance of

$$\frac{1}{\sigma_*^2} = \frac{n}{\sigma_0^2} + \frac{1}{\tau^2} = \frac{n\tau^2 + \sigma_0^2}{\sigma_0^2 \tau^2}.$$

This implies the posterior $\mu | Y \sim N(\mu_*, \sigma_*^2)$ with

$$\sigma_*^2 = \frac{\sigma_0^2 \tau^2}{n\tau^2 + \sigma_0^2},$$

and

$$\mu_* = \frac{\frac{n\bar{Y}}{\sigma_0^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}} = \frac{\frac{n}{\sigma_0^2}\bar{Y} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}}.$$

- The latter is a convex combination of μ_0 and $\bar{Y}$. As $n \to \infty$ it becomes the sample mean.