12. Beyond MLE: Using Prior Information

ECE 830, Spring 2017

Rebecca Willett

The Bayesian Paradigm

Given a parameter θ, we assume observations are generated according to $p(y|\theta)$. In our work so far, we have treated the parameter θ like a fixed, deterministic, but unknown quantity while the observation y is the realization of a random process.

We will now consider **probabilistic models** for θ in addition to our data.

- This allows us to incorporate prior information we have about θ (i.e. information about likely values of θ we have *before collecting any data*).
- It also allows us to make statements about our confidence in different estimates of θ.

Example: Unfair coin

Suppose you toss a single coin 6 times and each time it comes up "heads." It might be reasonable to say that we are 98% sure that the coin is unfair, biased towards heads.

Formally, we can think about this in a hypothesis testing framework using a binomial probabilistic model. Let $y :=$ number of "heads".

$$H_0 : \mathbb{P}(\text{heads}) \equiv \theta > 0.5$$

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$p(\theta > 0.5|y) = ?$$

Example: (cont.)

The problem with this is that

$$p(\theta \in H_0 | x)$$

implies that θ is a random, not deterministic, quantity.

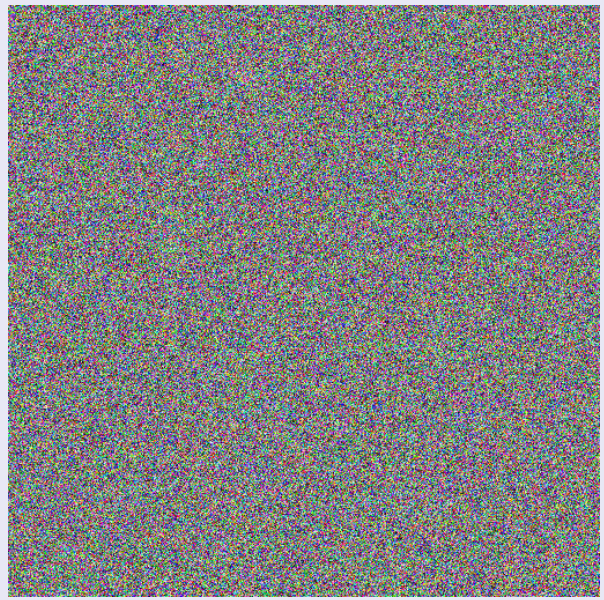
So, while "confidence" statements are very reasonable and in fact a normal part of "everyday thinking," this idea can not be supported from the classical perspective.

All of these "deficiencies" can be circumvented by a change in how we view the parameter θ.

Example: Image processing

In many imaging problems, we have a good sense of what "natural" images should look like.

Likely



Unlikely

This prior information can be exploited to improve image denoising, deblurring, reconstruction, and analysis.

Bayes Rule

If we view θ as the realization of a random variable with density $p(\theta)$, then we can work with the generative (or forward) model

$$\underbrace{p(\theta)}_{\text{prior}} \rightarrow \theta^* \rightarrow \underbrace{p(y|\theta^*)}_{\text{likelihood}} \rightarrow y.$$

We are interested in the inverse problem

$$y \rightarrow p(\theta|y) \rightarrow \hat{\theta}.$$

Bayes Rule (Bayes, 1763) shows that

$$p(\theta|y) = \frac{p(y|\theta)\ p(\theta)}{p(y)} = \frac{p(y|\theta)\ p(\theta)}{\int p(y|\widetilde{\theta})\ p(\widetilde{\theta})\ d\widetilde{\theta}}$$

Once we can compute this posterior distribution, confidence measures such as $p(\theta \in H_0|y)$ are perfectly legitimate quantities to ask for.

Example: Coin toss

Suppose you toss a single coin 6 times and each time it comes up "heads." Mathematically, we can model the problem as follows. Let $\theta = \mathbb{P}(\text{Heads})$. The data (the number of heads y in $n = 6$ tosses) follows a binomial distribution $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. The mathematical equivalent of the question "is the coin probably biased" is the probability $\mathbb{P}(\theta > 0.5|y = 6)$.

Suppose we assume $p(\theta) = \text{Unif}(0, 1)$ (all values of θ are equally probable before we begin to flip the coin, and $\mathbb{P}(\theta > \frac{1}{2}) = \frac{1}{2}$). Now compute

$$p(\theta|y) \;=\; \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \;=\; \frac{\theta^6}{\int \theta^6 d\theta} = \frac{\theta^6}{\frac{1}{7}\theta^7|_0^1} \;=\; 7\,\theta^6 \;.$$

Then

$$\mathbb{P}\left(\theta > \frac{1}{2}\,|\,y = 6\right) \;=\; \int_{\frac{1}{2}}^{1} 7\theta^6 d\theta \;=\; \theta^7|_{\frac{1}{2}}^1 \;=\; 1 - 2^{-7} \;=\; 0.984 \;.$$

(If we chose a different prior we would get a different answer!)

Bayesian statistical models

Definition: Bayesian statistical model

A Bayesian statistical model is composed of a *data generation model*, $p(y|\theta)$, and a *prior* distribution on the parameters, $p(\theta)$.

The prior distribution (or "prior" for short) models the uncertainty in the parameter. More specifically, $p(\theta)$ models our knowledge - or a lack thereof - prior to collecting data.

Notice that

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} \propto p(y|\theta)\, p(\theta)$$

Hence, $p(\theta|y)$ is proportional to the likelihood function multiplied by the prior.

Example: DC level in AWGN

$$
\begin{aligned}
y_i &= A + \nu_i \ , \quad n = 1, \cdots, N \\
\nu_i &\sim \mathcal{N}(0, \sigma^2) \ \ iid \\
\widehat{A} &= \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{MLE estimate}
\end{aligned}
$$

Now suppose that we have prior knowledge that $-A_0 \leq A \leq A_0$. We might incorporate this by forming a new estimator

$$
\widetilde{A} = \begin{cases} -A_0 & , \ \widehat{A} < -A_0 \\ \widehat{A} & , \ -A_0 \leq \widehat{A} \leq A_0 \\ A_0 & , \ \widehat{A} > A_0 \end{cases}
$$

This is called a truncated sample mean estimator of A.

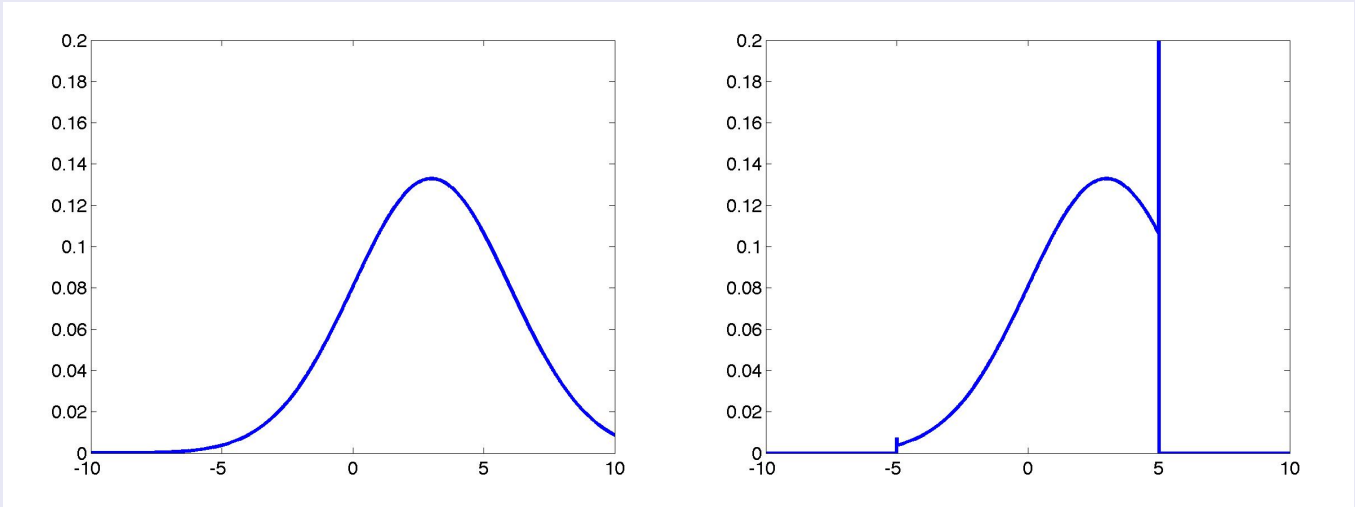
Example: (cont.)

Is $\widetilde{A}$ a better estimator of A than the sample mean $\widehat{A}$? Let p_{MLE} denote the density of $\widehat{A}$. Since $\widehat{A} = \frac{1}{n}\sum y_i$,

$$p_{\mathrm{MLE}} = \mathcal{N}(A, \sigma^2/n).$$

The density of the truncated sample mean (TSM) $\widetilde{A}$ is given by

$$\begin{aligned} p_{\mathrm{TSM}} =& \mathbb{P}(\widehat{A} \le -A_0)\ \delta(a + A_0) + p_{\mathrm{MLE}} I_{\{-A_0 \le a \le A_0\}} \\ &+ \mathbb{P}(\widehat{A} \ge A_0)\ \delta(a - A_0) \end{aligned}$$



p_{MLE} p_{TSM}

Now consider the MSE of the sample mean $\widehat{A}$:

$$
\begin{aligned}
\mathsf{MSE}(\widehat{A}) &= \int_{-\infty}^{\infty} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&= \int_{-\infty}^{-A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + \int_{A_0}^{\infty} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&> \int_{-\infty}^{-A_0} (-A_0-A)^2 \, p_{\mathrm{MLE}}(a) \, da + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + \int_{A_0}^{\infty} (A_0-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&= (-A_0-A)^2 \, \mathbb{P}(\widehat{A} \le -A_0) + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + (A-A_0)^2 \, \mathbb{P}(\widehat{A} \ge A_0) \\
&= \mathsf{MSE}(\widetilde{A})
\end{aligned}
$$

$$\mathsf{MSE}(\widehat{A}) > \mathsf{MSE}(\widetilde{A})$$

Note

1. $\widetilde{A}$ is biased
2. Although $\widehat{A}$ is the MLE, $\widetilde{A}$ is better in the MSE sense
3. Prior information is aptly described by regarding A as a random variable with a prior distribution.

$$\text{uniform}(-A_0, A_0)$$

$\Rightarrow$ We know $-A_0 \leq A \leq A_0$, but otherwise A is arbitrary.

Elements of Bayesian Analysis

(a) Joint distribution

$$p(y, \theta) = p(y|\theta)p(\theta)$$

(b) Marginal distributions

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

$$p(\theta) = \int p(y|\theta)p(\theta)dy \text{ (“prior”)}$$

(c) Posterior distribution

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)dy}$$

Example: Binomial + Beta

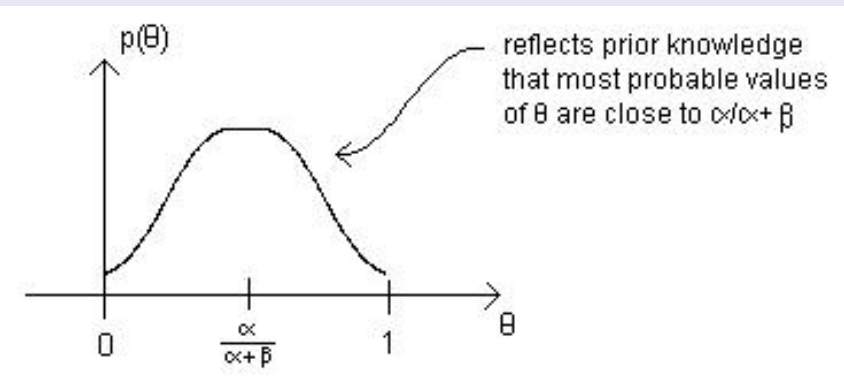
$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, 0 \leq \theta \leq 1$$

$$= \text{binomial likelihood}$$

$$p(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$= \text{Beta prior distribution}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ is the Gamma function



Example: (cont.)

- Joint Density

$$p(y, \theta) =$$

- Marginal Density

$$p(y) =$$

- Posterior Density

$$p(\theta|y) =$$

Bayesian Estimation

We are interested in estimating θ given the observation y within a Bayesian framework. Naturally, then, any estimation strategy will be based on the posterior distribution $p(\theta|y)$.

However, we need a criterion for assessing the quality of potential estimators.

Loss

Definition: Loss

The quality of an *estimate* $\widehat{\theta}$ is measured by a real-valued *loss (or cost) function*

$$L(\theta, \widehat{\theta}).$$

For example, *squared error* or quadratic loss is simply

$$L(\theta, \widehat{\theta}(y)) = (\theta - \widehat{\theta}(y))^\top (\theta - \widehat{\theta}(y)) = \|\theta - \widehat{\theta}(y)\|^2.$$

Definition: Bayes Risk

The quality of an estimator is measured by the expected loss, known as the Bayes risk:

$$R(\widehat{\theta}) := \mathbb{E}_{y,\theta}\left[L(\theta, \widehat{\theta}(y))\right].$$

(When write $R(\widehat{\theta})$, we mean risk relative to a particular *strategy* (i.e. way of choosing $\widehat{\theta}$), not a particular *value* of $\widehat{\theta}$)

Note that the expectation is with respect to both y and θ.

For example, if y and θ are jointly continuous, then

$$\begin{aligned} R(\widehat{\theta}) &= \iint L(\theta, \widehat{\theta}(y)) p(\theta, y) dy d\theta \\ &= \iint L(\theta, \widehat{\theta}(y)) p(y|\theta) p(\theta) dy d\theta \\ &= \mathbb{E}_y \left[\mathbb{E}_{\theta|y} \left[L(\theta, \widehat{\theta}(y)) | y = y \right] \right] \end{aligned}$$

In general, Bayesian estimation seeks the "Bayes estimator"

$$\begin{aligned}
\widehat{\theta} &= \arg\min_{\widetilde{\theta}} R(\widetilde{\theta}) \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{y,\theta} \left\{ L\left(\theta, \widetilde{\theta}(y)\right) \right\} \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{y} \left\{ \mathbb{E}_{\theta|y} \left\{ L\left(\theta, \widetilde{\theta}(y)\right) | y = y \right\} \right\}
\end{aligned}$$

that minimizes the Bayes risk. Thus, given the data y, the "best" or optimal estimator under a given loss function is given by

$$\widehat{\theta}(y) = \arg\min_{\widetilde{\theta}} \mathbb{E}\left[L\left(\theta, \widetilde{\theta}\right) | y \right].$$

This is called the "posterior expected loss"; it depends only on the loss function and the posterior distribution.

Bayesian estimation with squared error loss

Measure the loss as $L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|^2$. Now note

$$\begin{aligned}
&\mathbb{E}_{\theta|y=y}[(\theta - \widehat{\theta}(y))^\top(\theta - \widehat{\theta}(y))] \\
&= \mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y] + \mathbb{E}[\theta|y] - \widehat{\theta}(y))^\top(\theta - \mathbb{E}[\theta|y] + \mathbb{E}[\theta|y] - \widehat{\theta}(y))] \\
&= \mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y])^\top(\theta - \mathbb{E}[\theta|y])] + 2\mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y])^\top(\mathbb{E}[\theta|y] - \widehat{\theta}(y))] \\
&\quad + \mathbb{E}_{\theta|y}[(\mathbb{E}[\theta|y] - \widehat{\theta}(y))^\top(\mathbb{E}[\theta|y] - \widehat{\theta}(y))]
\end{aligned}$$

The first term is independent of $\widehat{\theta}(y)$ and the second term is 0. The third term can be minimized by taking

$$\widehat{\theta}_{\mathrm{PM}}(y) = \mathbb{E}[\theta|y] = \int \theta p(\theta|y) d\theta$$

which is the posterior mean. (In signal processing this is also called the "Bayesian minimim mean squared error estimator".)

Example: DC Level in AWGN

$$y_i = A + \nu_i$$

$i = 1, \cdots, n$, $\nu_i \sim \mathcal{N}(0, \sigma^2)$. Prior for unknown parameter A:

$$p(A) = \text{Unif}[-A_0, A_0]$$

$$p(y|A) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - A)^2 \right\}$$

$$p(A|y) = \begin{cases} \dfrac{\frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - A)^2\right\}}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right\} da} & \text{if } |A| \le A_0 \\ 0 & \text{if } |A| > A_0 \end{cases}$$

Example: (cont.)

Bayes Minimum MSE Estimator:

$$\begin{aligned}\widehat{A} &= \mathbb{E}[A|y] = \int_{-\infty}^{\infty} A p(A|y) dA \\ &= \frac{\int_{-A_0}^{A_0} A \cdot \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - A)^2\right\} dA}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right\} da}\end{aligned}$$

Notes:

1. No closed-form estimator
2. As $A_0 \to \infty$, $\widehat{A} \to \frac{1}{n}\sum_{i=1}^{n} y_i$
3. For smaller A_0, truncated integral produces an $\widehat{A}$ that is a function of y, σ^2, and A_0
4. As n increases σ^2/n decreases and posterior $p(A|y)$ becomes tightly clustered about $\frac{1}{n}\sum_i y_i$
   $\Rightarrow \widehat{A} \to \frac{1}{n}\sum y_i$ as $n \to \infty$ (the data "swamps out" the prior)

Other Common Loss Functions

Absolute Error Loss (Laplace, 1773):

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_1 := \sum_{i=1}^{p} |\theta_i - \widehat{\theta}_i|$$

Scalar case:

$$\begin{aligned} \mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] &= \int_{-\infty}^{\infty} |\theta - \widehat{\theta}| p(\theta|y) d\theta \\ &= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) p(\theta|y) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) p(\theta|y) d\theta \end{aligned}$$

The optimal estimator under this loss is referred to the "minimum mean absolute error" (MMAE) estimator.

To see what estimator minimises this loss, we differentiate $\mathbb{E}\left[L(\theta, \widehat{\theta})|y\right]$ with respect to $\widehat{\theta}$ (using Leibnitz's rule) to get

$$\frac{\partial}{\partial \widehat{\theta}} \mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] = P(\widehat{\theta}(y)|y) - (1 - P(\widehat{\theta}(y)|y)),$$

where $P(\theta|y)$ is the posterior cumulative distribution function of θ given y. Setting this equal to zero, this implies $P(\widehat{\theta}(y)|y) = 1/2$ or

$$\mathbb{P}(\theta < \widehat{\theta}|y) = \mathbb{P}(\theta > \widehat{\theta}|y).$$

The optimal $\widehat{\theta}$ under absolute error loss is the posterior median.

Uniform Loss:

$$L(\theta, \widehat{\theta}) = I_{\left\{\|\widehat{\theta}-\theta\|>\epsilon\right\}} = \begin{cases} 1 & \text{if } \|\theta - \widehat{\theta}\| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

where $\epsilon > 0$ is small. The posterior expected loss is

$$\mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] = \mathbb{E}\left[I_{\left\{\|\widehat{\theta}-\theta\|>\epsilon\right\}}|y\right] = \mathbb{P}(\|\widehat{\theta} - \theta\| > \epsilon|y)$$

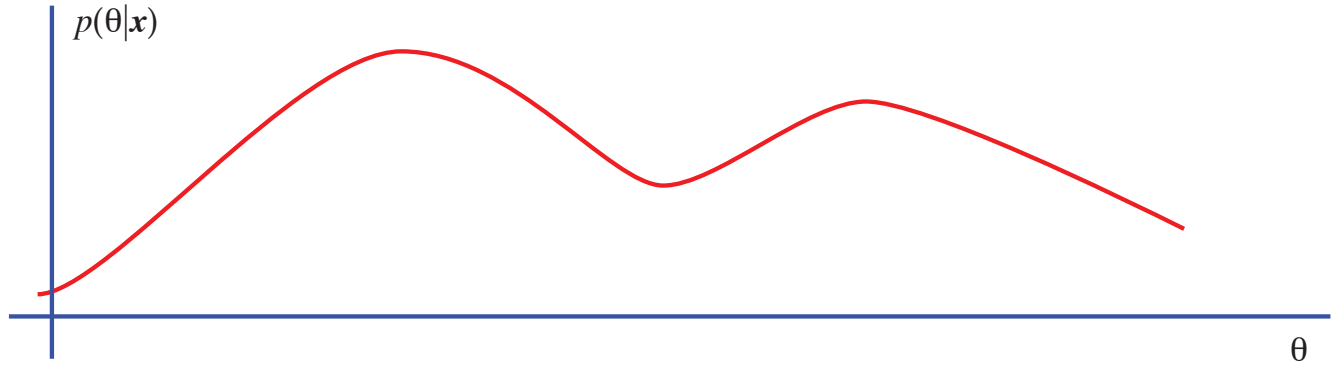
which is the posterior probability that θ deviates from $\widehat{\theta}(y)$ by more then ϵ. To minimize this uniform loss we must choose $\widehat{\theta}$ to be the value of θ with highest posterior probability. Taking the limit as $\epsilon \longrightarrow 0$ gives:

The optimal estimator $\widehat{\theta}$ under uniform loss is the posterior mode.

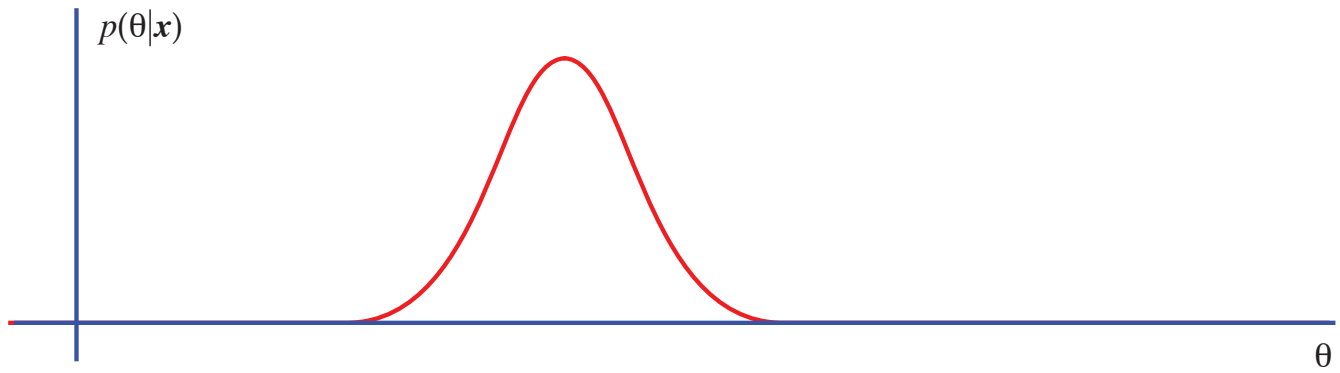
Definition

Maximum A Posteriori (MAP) estimator - the value of θ where $p(\theta|y)$ is maximized:

$$\widehat{\theta}_{\mathrm{MAP}}(y) = \arg\max_{\widetilde{\theta}} p(\widetilde{\theta}|y) = \arg\max_{\widetilde{\theta}} p(y|\widetilde{\theta})p(\widetilde{\theta})$$



(a) General posterior PDF



(b) Gaussian posterior PDF

If the posterior is symmetric and unimodal, then

$$\widehat{\theta}_{\mathrm{MMSE}} = \widehat{\theta}_{\mathrm{MMAE}} = \widehat{\theta}_{\mathrm{MAP}}$$

Computation

Both $\widehat{\theta}_{\mathrm{PM}}$ and $\widehat{\theta}_{\mathrm{MMAE}}$ require integrating with respect to $p(\theta|y)$. Often this calculation will be intractable. How can we approximate these estimators numerically?
One common approach: if we can simulate $\theta_1, \ldots, \theta_M$ from $p(\theta|y)$, then we can apply the following Monte Carlo estimates:

$$\begin{aligned} \widehat{\theta}_{\mathrm{PM}}(y) &\approx \frac{1}{M}\sum_{i=1}^{M}\theta_i \\ \widehat{\theta}_{\mathrm{MMAE}}(y) &\approx \operatorname{median}\{\theta_1, \ldots, \theta_{\mathrm{M}}\} \end{aligned}$$

If the posterior mode cannot be determined analytically, then numerical approaches can be applied.
Which of the three loss functions is used is often dictated by computational considerations.

Choosing a Prior

Two approaches:

1. Informative (or "subjective") priors:
   - design/choose priors that are compatible with prior knowledge of unknown parameters
   - can be impractical in complicated problems with many parameters
   - injecting subjective opinion into analysis contrary to making scientific analysis as objective as possible.
2. Non-informative priors:
   - attempt to remove subjectiveness from Bayesian procedures
   - designs are often based on invariance arguments

Selecting an Informative Prior

Clearly, the most important objective is to choose the prior $p(\theta)$ that best reflects the prior knowledge available to us.

In general, however, our prior knowledge is imprecise and any number of prior densities may aptly capture this information.

Moreover, usually the optimal estimator can't be obtained in closed-form.

Therefore, sometimes it is desirable to choose a prior density that models prior knowledge and is nicely matched in functional form to $p(y|\theta)$ so that the optimal estimator (and posterior density) can be expressed in a simple fashion.

Conjugate Priors

Idea: Given $p(y|\theta)$, choose $p(\theta)$ so that $p(\theta|y) \propto p(y|\theta)\, p(\theta)$ has a simple functional form.

Conjugate priors: choose $p(\theta) \in \mathcal{P}$, where $\mathcal{P}$ is a family of densities (e.g., Gaussian family) so that the posterior density also belongs to that family.

Definition

$p(\theta)$ is a *conjugate prior* for $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

Example: Conjugate priors for exponential random variables

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} \text{exponential}(\theta)$$

$$p(y|\theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i} = \theta^n e^{-\theta t}$$

where $t := \sum y_i$.

Example: (cont.)

Let $\theta \sim \text{Gamma}(\alpha, \beta)$, so that $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$ for $\theta \in [0, \infty)$. Then

$$p(y, \theta) =$$

$$p(y) =$$

$$=$$

$$=$$

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)}$$

$$=$$

$$=$$

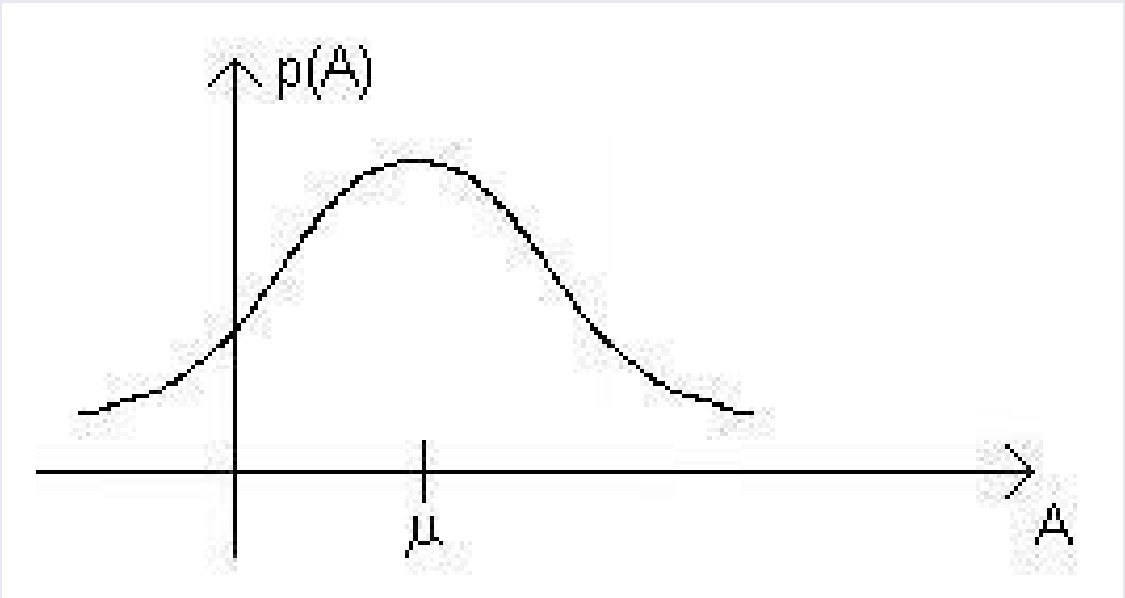
Thus the Gamma prior is conjugate for the exponential distribution!

Example: Constant in AWGN

$$y_i = A + \nu_i \ , \ i = 1, \cdots, n; \qquad \nu_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Rather than modeling $A \sim \text{Uniform}(-A_0, A_0)$ (which did not yield a closed-form estimator) consider

$$p(A) = \frac{1}{\sqrt{2\pi {\sigma_A}^2}} \exp\left\{ -\frac{1}{2{\sigma_A}^2}(A - \mu)^2 \right\}$$



Example: (cont.)

With $\mu = 0$ and $\sigma_A = \frac{1}{3}A_0$ this Gaussian prior also reflects prior knowledge that it is unlikely for $|A| \geq A_0$.
The Gaussian prior is also conjugate to the Gaussian likelihood

$$p(y|A) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - A)^2\right]$$

so that the resulting posterior density is also a simple Gaussian, as shown next. First note that

$$p(y|A) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\left(nA^2 - 2nA\bar{y}\right)\right]$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Example: (cont.)

$$
\begin{aligned}
p(A|y) &= \frac{p(y|A)\, p(A)}{\int p(y|a)\, p(a)\, da} \\
&= \frac{\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(nA^2 - 2nA\bar{y}) + \frac{1}{{\sigma_A}^2}(A-\mu)^2\right)\right]}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(na^2 - 2na\bar{y}) + \frac{1}{{\sigma_A}^2}(a-\mu)^2\right)\right] da} \\
&= \frac{e^{-\frac{1}{2}Q(A)}}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}Q(a)} da}
\end{aligned}
$$

where

$$
Q(A) = \frac{n}{\sigma^2}A^2 - \frac{2nA\bar{y}}{\sigma^2} + \frac{A^2}{{\sigma_A}^2} - \frac{2\mu A}{{\sigma_A}^2} + \frac{\mu^2}{{\sigma_A}^2}
$$

Now let

$$
\begin{aligned}
\sigma^2_{A|y} &:= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\
\mu^2_{A|y} &:= \left(\frac{n}{\sigma^2}\bar{y} + \frac{\mu}{{\sigma_A}^2}\right)\sigma^2_{A|y}
\end{aligned}
$$

Example: (cont.)

Then by "completing the square" we have

$$\begin{aligned} Q(A) &= \frac{1}{\sigma^2_{A|y}}\left(A^2 - 2\mu_{A|y}A + \mu^2_{A|y}\right) - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}} + \frac{\mu^2}{{\sigma_A}^2} \\ &= \frac{1}{\sigma^2_{A|y}}\left(A^2 - \mu_{A|y}\right)^2 - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}} + \frac{\mu^2}{{\sigma_A}^2} \end{aligned}$$

Hence

$$p(A|y) = \frac{\exp\left[-\frac{1}{2\sigma^2_{A|y}}(A-\mu_{A|y})^2\right]\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}}\right)\right]}{\int_{-\infty}^{\infty} \underbrace{\exp\left[-\frac{1}{2\sigma^2_{A|y}}(a-\mu_{A|y})^2\right]}_{\text{"unnormalized" Gaussian density}} \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}}\right)\right]}_{\text{Constant, indep. of } a} da}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2_{A|y}}}\exp\left[-\frac{1}{2\sigma^2_{A|y}}(A-\mu_{A|y})^2\right]$$

$$A|y \sim \mathcal{N}(\mu_{A|y}, \sigma^2_{A|y})$$

Example: (cont.)

Now

$$\begin{aligned}\widehat{A} &= \mathbb{E}[A|y] = \mu_{A|y} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{\mu}{{\sigma_A}^2}}{\frac{n}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\ &= \left(\frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/n}\right)\bar{y} + \left(\frac{\sigma^2/n}{{\sigma_A}^2 + \sigma^2/n}\right)\mu \\ &= \alpha\bar{y} + (1-\alpha)\mu\end{aligned}$$

where

$$0 < \alpha = \frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/n} < 1$$

Interpretation

1. When there is little data ($\sigma_A^2 \ll \frac{\sigma^2}{n}$), α is small and $\widehat{A} \approx \mu$.
2. When there is a lot of data ($\sigma_A^2 \gg \frac{\sigma^2}{n}$), $\alpha \approx 1$ and $\widehat{A} \approx \bar{y}$.

Interplay between data and prior knowledge: small $n \Leftrightarrow \widehat{A}$ favors prior, large $n \Leftrightarrow \widehat{A}$ favors data

Overview

So far...

- Bayesian methods assume the unknown parameter θ is stochastic and we have a prior probabilistic model for θ
- Given the prior $p(\theta)$ and data $y \sim p(y|\theta)$, we can compute the posterior

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta).$$

- Given the posterior, we can estimate θ multiple different ways. Popular examples:
  - Posterior mean: $\hat{\theta} = \mathbb{E}[\theta|y] = \int \theta p(\theta|y) d\theta$.
  - Maximum a posteriori (MAP): $\hat{\theta} = \arg\max_\theta p(\theta|y)$.
- Choosing conjugate priors ensures the posterior has a nice functional form.

Ahead

The multivariate Gaussian linear model...

- ... with a multivariate Gaussian prior $\Longrightarrow$ ridge regression
- ... with a multivariate Laplace prior $\Longrightarrow$ LASSO (least absolute shrinkage and selection operator) regression

These models and methods appear in a wide variety of modern machine learning and signal processing settings.

The Multivariate Gaussian Model

We consider the following linear statistical model

$$y = X\theta + \nu$$

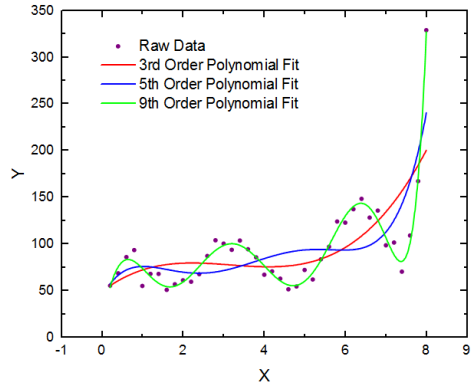
where

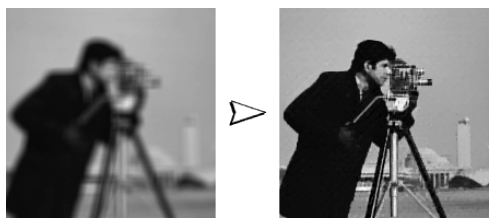
y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known

Examples

This linear model appears throughout science, engineering, and machine learning.

- y measures the US census count over n years. X is a Vandermonde matrix representing an order-p polynomial approximation, and θ contains the p polynomial coefficients.
- y is an n-pixel blurry image we take with our camera, X models the blurring process, and θ is the desired blur-free image (here $p = n$).
- Each element of y is your heart rate at n different times of the day. Each of the p columns of X measures one of your activities (e.g. alcohol consumption, nap time(s), exercise, proximity to attractive people) at the same times in the day. θ characterizes how much each of these activities contributes to your heart rate.





A Gaussian prior

Consider the following Bayesian linear statistical model

$$y = X\theta + \nu$$

where

y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known and full-rank
θ	$\sim$	$\mathcal{N}(\mu_\theta, \Sigma_\theta)$
θ	is	unknown, $p \times 1$ (p unknown parameters)
μ_θ	is	known, $p \times 1$
Σ_θ	$\in$	$\mathcal{S}_p$ is known and full-rank
θ and ν	are	independent

This model amounts to a Gaussian prior on θ and a Gaussian conditional distribution of y given θ.

What is the posterior?

First, note the y and θ are jointly Gaussian:

$$\begin{bmatrix} \theta \\ y \end{bmatrix} = \begin{bmatrix} 0 & I_p \\ I_p & X \end{bmatrix} \begin{bmatrix} \nu \\ \theta \end{bmatrix}.$$

Since

$$\begin{bmatrix} \nu \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} \Sigma_\nu & 0 \\ 0 & \Sigma_\theta \end{bmatrix} \right),$$

we have

$$\begin{bmatrix} \theta \\ y \end{bmatrix} \sim \mathcal{N}\left(\qquad\qquad , \qquad\qquad \right).$$

Lemma

If

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

then

$$Z_1 | Z_2 = z_2 \sim \mathcal{N}\left(\mu', \Sigma'\right)$$

where

$$\mu' := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)$$
$$\Sigma' := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

We next apply this lemma to $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \theta \\ y \end{bmatrix}$.

Gauss-Markov Theorem

The posterior distribution of $\theta|y$ is

$$\theta|y \sim \mathcal{N}(\mu_{\theta|y}, \Sigma_{\theta|y})$$

where

$$\begin{aligned}
\mu_{\theta|y} &= \mu_\theta + \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1}(y - X\mu_\theta) \\
&= \mu_\theta + \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1} X^\top \Sigma_\nu^{-1}(y - X\mu_\theta) \\
\Sigma_{\theta|y} &= \Sigma_\theta - \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1} X\Sigma_\theta \\
&= \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1}
\end{aligned}$$

The second version of each expression is a result of the following:

Matrix Inversion Lemma

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Specifically, first apply the matrix inversion lemma to $\Sigma_{\theta|y} = \Sigma_\theta - \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1} X\Sigma_\theta$ to get $\Sigma_{\theta|y} = \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1}$.

Now let $G := \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1}$, so that

$$\begin{aligned}\mu_{\theta|y} &= \mu_\theta + G(y - X\mu_\theta)\\ \Sigma_{\theta|y} &= \Sigma_\theta - GX\Sigma_\theta.\end{aligned}$$

Now using the formula above $\Sigma_{\theta|y} = \Sigma_\theta - GX\Sigma_\theta$, we have

$$\begin{aligned} GX\Sigma_\theta =& \Sigma_\theta - \Sigma_{\theta|y} = \Sigma_{\theta|y}(\Sigma_{\theta|y}^{-1}\Sigma_\theta - I) \\ =& \Sigma_{\theta|y}\left[\left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)\Sigma_\theta - I\right] \\ =& \Sigma_{\theta|y} X^\top \Sigma_\nu^{-1} X \Sigma_\theta. \end{aligned}$$

That last gives the identity

$$G = \Sigma_{\theta|y} X^\top \Sigma_\nu^{-1} = \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1} X^\top \Sigma_\nu^{-1}$$

as desired.

Observations

- The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the posterior mean and MAP estimators are both

$$\begin{aligned}\widehat{\theta}(y) = \mu_{\theta|y} &= \mu_\theta + \Sigma_\theta X^\top (X\Sigma_\theta X^\top + \Sigma_\nu)^{-1}(y - X\mu_\theta) \\ &= \mu_\theta + (X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1})^{-1} X^\top \Sigma_\nu^{-1}(y - X\mu_\theta)\end{aligned}$$

- $\widehat{\theta}(y)$ is an affine function of y.
- $\widehat{\theta}(y)$ is itself multivariate Gaussian.

Special case 1: The noncommittal prior

Consider the case where $\mu_\theta = 0$, $\Sigma_\theta = \sigma^2 I_p$ and $\sigma^2 \longrightarrow \infty$. This can be thought of as a "noncommittal" prior. Then $\Sigma_\theta^{-1} \longrightarrow 0_p$ and

$$\widehat{\theta}(y) = \mu_{\theta|y} \quad =$$

Furthermore, if $\Sigma_\nu = \sigma^2 I_n$, then

$$\widehat{\theta}(y) =$$

which is

Special case 2: Uncorrelated prior

Consider the case where $\mu_\theta = 0$, $\Sigma_\theta = \sigma_\theta^2 I_p$, and $\Sigma_\nu = \sigma_\nu^2 I_n$. Then

$$\widehat{\theta}(y) = \mu_{\theta|y} =$$

$$=$$

$$=$$

This is referred to as ridge regression.

Note that even when $X^\top X$ is poorly conditioned or not invertible, the sum $X^\top X + \frac{\sigma_\nu^2}{\sigma_\theta^2} I_p$ can be well conditioned even for small values of $\frac{\sigma_\nu^2}{\sigma_\theta^2}$. As a result, this estimator, while biased, can have far less variance than the least squares estimator.

Example:

$y = \theta + \nu \in \mathbb{R}^n \;\; , \;\; \nu \sim \mathcal{N}(0, \sigma^2 I_n)$

$$
\begin{aligned}
p(\theta) &= \mathcal{N}(0, \Sigma_{\theta\theta}) \text{ indep. of } \nu \\
\mathbb{E}[y] &= \\
\mathbb{E}\left[yy^\top\right] &= \\
&= \\
\mathbb{E}\left[y\theta^\top\right] &= \mathbb{E}\left[\theta\theta^\top\right] + \mathbb{E}\left[\nu\theta^\top\right] \\
&= \\
\begin{bmatrix} y \\ \theta \end{bmatrix} &\sim \mathcal{N}
\end{aligned}
$$

We can invoke the Gauss-Markov theorem to get

$$\widehat{\theta} =$$

Example: (cont.)

(Alternative derivation:) From our Bayesian perspective, we are interested in $p(\theta|y)$.

$$p(\theta|y) = \frac{(2\pi)^{-N/2}\,(2\pi)^{-N/2}\,|\Sigma|^{-1/2}\,\exp\left\{-\frac{1}{2}[y^\top\theta^\top]\Sigma^{-1}\begin{bmatrix} y \\ \theta \end{bmatrix}\right\}}{(2\pi)^{-N/2}\,|\Sigma_{yy}|^{-1/2}\,\exp\left\{-\frac{1}{2}y^\top\Sigma_{yy}^{-1}y\right\}}$$

In this formula we are faced with

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{y\theta} \\ \Sigma_{\theta y} & \Sigma_{\theta\theta} \end{bmatrix}^{-1}$$

The inverse of this covariance matrix can be written as

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{y\theta} \\ \Sigma_{\theta y} & \Sigma_{\theta\theta} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{yy}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\Sigma_{yy}^{-1}\Sigma_{y\theta} \\ I \end{bmatrix} Q^{-1} \begin{bmatrix} -\Sigma_{\theta y}\Sigma_{yy}^{-1} & I \end{bmatrix}$$

where $Q := \Sigma_{\theta\theta} - \Sigma_{\theta y}\,\Sigma_{yy}^{-1}\,\Sigma_{y\theta}$ is the Schur complement of $\Sigma_{\theta\theta}$. (Verify this formula by applying RHS above to Σ to get I.)

Example: (cont.)

Furthermore,

$$\det \Sigma = \det \Sigma_{yy} \det Q.$$

Substituting this expression into $p(\theta|y)$ we get

$$\begin{aligned} p(\theta|y) &= (2\pi)^{-N/2} \, |Q|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} (\theta - \Sigma_{\theta y} \Sigma_{yy}^{-1} y)^\top Q^{-1} (\theta - \Sigma_{\theta y} \Sigma_{yy}^{-1} y) \right\} \\ \theta|y &\sim \mathcal{N}\left(\Sigma_{\theta y} \Sigma_{yy}^{-1} y, Q\right) \end{aligned}$$

Thus the posterior mean of θ is

$$\widehat{\theta} = \Sigma_{\theta y} \Sigma_{yy}^{-1} y$$

and the posterior variance is

$$Q = \Sigma_{\theta\theta} - \Sigma_{\theta y} \Sigma_{yy}^{-1} \Sigma_{y\theta}$$

Example: DC Level in AWGN

$$y_i = A + \nu_i,\ i = 1, \ldots, n$$

where A is an unknown scalar to be estimated and

$$\begin{aligned} A &\sim \mathcal{N}(\mu_A, {\sigma_A}^2) \\ \nu_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, {\sigma_\nu}^2)\ , \text{ indep. of } y \end{aligned}$$

This problem falls within the Gaussian linear model with

Example: (cont.)

Using the second formula for $\mu_{A|x}$, we obtain

$$\widehat{A}(y) = \mu_{A|x} =$$

$$=$$

$$=$$

$$=$$

Example: (cont.)

In other words,

$$\widehat{A}(y) = (1-\alpha)\mu_A + \alpha\overline{y}$$

where

$$\alpha =$$

controls the tradeoff between prior knowledge and data. Limiting cases:

$$\begin{aligned}
n \to \infty &\implies \alpha \to &&\implies \widehat{A} \to \\
n = 0 &\implies \alpha = &&\implies \widehat{A} = \\
\sigma_A^2 \to \infty &\implies \alpha \to &&\implies \widehat{A} \to \\
\sigma_A^2 = 0 &\implies \alpha = &&\implies \widehat{A} =
\end{aligned}$$

Tikhinov regularization

We can also compute the MAP estimator directly. Assume here that $\Sigma_\nu = \sigma_\nu^2 I_n$ and $\mu_\theta = 0$, and define Γ such that $\Sigma_\theta^{-1} = \Gamma^\top \Gamma$. (If we have the eigendecomposition $\Sigma_\theta = V \Lambda V^\top$, then $\Sigma_\theta = V \Lambda^{-1} V^\top = V \Lambda^{-1/2} \underbrace{\Lambda^{-1/2} V^\top}_{\Gamma}$.) Then

$$
\begin{aligned}
-\log p(y|\theta) &\propto \frac{1}{2\sigma_\nu^2} \|y - X\theta\|_2^2 \\
-\log p(\theta) &\propto \frac{1}{2} \|\theta^\top \Sigma_\theta^{-1} \theta\|_2^2 \\
\hat{\theta}_{\text{MAP}} &= \arg\min_\theta \{-\log p(y|\theta) - \log p(\theta)\} \\
&= \\
&=
\end{aligned}
$$

(If $\Gamma = \sigma_\theta I_p$ we arrive at the ridge regression expression.)

Tikhinov regularization (cont.)

Consider again the estimate

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2}{2}\|\Gamma\theta\|_2^2$$

The first term measures how well θ fits our observed data. The second term reflects our prior knowledge – essentially we seek θ for which $\Gamma\theta$ has a small norm.

As σ_ν increases and we must cope with more noise in our data, we increase our dependence on this prior.

Furthermore, our choice of Γ (and hence Σ_θ) can determine which θs our estimator will be biased towards.

Example: Smooth θ

If we think θ should vary smoothly from one element to the next, we might choose Γ so that $(\Gamma\theta)_i = \theta_i - \theta_{i-1}$.

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in AWGN:

$$y = \theta + \nu$$

where y is the observed signal, θ is the clean signal, and ν is the noise. This can be modeled using a general linear model using $\theta = \theta$ and $X = I_n$. We can adopt a Gaussian prior for θ:

$$\theta \sim \mathcal{N}(0, \Sigma_{\theta\theta}).$$

The Bayesian estimate of θ is then

$$\widehat{\theta} = \Sigma_{\theta\theta} \left(\Sigma_{\theta\theta} + \Sigma_{\nu\nu}\right)^{-1} y.$$

Now suppose that $\Sigma_{\theta\theta}$ and $\Sigma_{\nu\nu}$ are simultaneously diagonalizable, meaning there exists an orthogonal matrix U such that

$$\Sigma_{\theta\theta} = U \Lambda_\theta U^\top$$

$$\Sigma_{\nu\nu} = U \Lambda_\nu U^\top$$

with $\Lambda_\theta, \Lambda_\nu$ diagonal. For example, consider $\Sigma_{\nu\nu} = \sigma^2 I$ and $\Sigma_{\theta\theta}$ arbitrary.

Then the estimator becomes

$$\begin{aligned}
\widehat{\theta} &= \Sigma_{\theta\theta}\left(\Sigma_{\theta\theta}+\Sigma_{\nu\nu}\right)^{-1} y \\
&= U\Lambda_\theta U^\top \left(U\Lambda_\theta U^\top + U\Lambda_\nu U^\top\right)^{-1} y \\
&= U\Lambda_\theta U^\top \left(U(\Lambda_\theta+\Lambda_\nu)U^\top\right)^{-1} y \\
&= U \underbrace{\left[\Lambda_\theta(\Lambda_\theta+\Lambda_\nu)^{-1}\right]}_{\Lambda} U^\top y
\end{aligned}$$

where

$$\Lambda = \begin{bmatrix}
\frac{\lambda_1^{(\theta)}}{\lambda_1^{(\theta)}+\lambda_1^{(\nu)}} & 0 & \cdots & 0 \\
0 & \frac{\lambda_2^{(\theta)}}{\lambda_2^{(\theta)}+\lambda_2^{(\nu)}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \\
0 & 0 & \cdots & \frac{\lambda_n^{(\theta)}}{\lambda_n^{(\theta)}+\lambda_n^{(\nu)}}
\end{bmatrix}.$$

Interpretation:

- U is a change of basis matrix
- $\theta = U^\top y$ are coefficients of y in new basis
- $z = \Lambda\theta$ is a coordinate-wise rescaling of θ
- $\widehat{\theta} = Uz$ is a reconstruction of θ from z.

How should we interpret the weights

$$\lambda_i := \frac{\lambda_i^{(\theta)}}{\lambda_i^{(\theta)} + \lambda_i^{(\nu)}}?$$

Notice that

$$\begin{aligned}
U^\top y &= U^\top \theta + U^\top \nu \\
U^\top \theta &\sim \mathcal{N}(0, U^\top \Sigma_{\theta\theta} U) = \mathcal{N}(0, \Lambda_\theta) \\
U^\top \nu &\sim \mathcal{N}(0, U^\top \Sigma_{\nu\nu} U) = \mathcal{N}(0, \Lambda_\nu).
\end{aligned}$$

Writing

$$U = \left[\begin{array}{cccc} u_1 & u_2 & \cdots & u_n \end{array} \right]$$

we have

$$u_i^\top \theta \sim \mathcal{N}(0, \lambda_i^{(\theta)})$$
$$u_i^\top \nu \sim \mathcal{N}(0, \lambda_i^{(\nu)})$$

Thus, λ_i reflects the proportion of the projection onto u_i that is due to the signal.

A Laplacian prior

Consider the following Bayesian linear statistical model

$$y = X\theta + \nu$$

where

y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known and full-rank
θ	is	unknown, $p \times 1$ (p unknown parameters)
θ	$\sim$	Laplace(λ)
$p(\theta)$	$=$	$\prod_{i=1}^{p} \frac{\lambda}{2} \exp(-\lambda \lvert\theta_i\rvert)$
λ	is	known scalar
θ and ν	are	independent

This model amounts to a Laplacian prior on θ and a Gaussian conditional distribution of y given θ.

Using the Laplacian prior

With the Laplacian prior we do not get the same simple expression for the posterior distribution.

However, we can still examine the MAP estimate where $\Sigma_\nu = \sigma_\nu^2 I_n$:

$$-\log p(y|\theta) \propto \frac{1}{2\sigma_\nu^2}\|y - X\theta\|_2^2$$

$$-\log p(\theta) \propto \lambda \sum_{i=1}^{p} |\theta_i| \equiv \lambda \|\theta\|_1$$

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \{-\log p(y|\theta) - \log p(\theta)\}$$

$$=$$

This estimate is called the LASSO (least absolute shrinkage and selection operator) estimate.

Ridge vs. LASSO

$$\hat{\theta}_{\text{Ridge}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2}{2\sigma_\theta^2}\|\theta\|_2^2 \right\}$$

$$\hat{\theta}_{\text{LASSO}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2\lambda}{2}\|\theta\|_1 \right\}$$

In both cases, we attempt to find a θ which (a) is a good fit to our data and (b) adheres to prior information captured by either the ℓ_2 or ℓ_1 norm of θ.

When should we use one vs. the other?

In general, the LASSO estimator favors *sparser* θ – i.e., θ with more zero-valued elements. There is no closed-form expression for the LASSO estimate.

Example: Deblurring

θ is a p-pixel image. X is a blur operator. $y = X\theta + \nu$ is a p-pixel blurry, noisy image. Some of the eigenvalues of X are very close to zero, so

$$(X^\top X)^{-1}$$

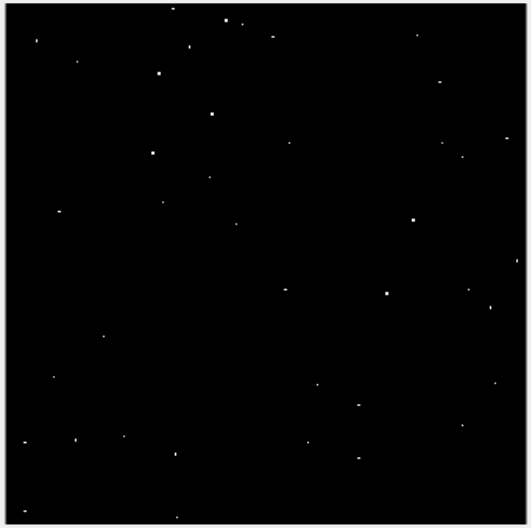
has some huge elements, meaning the least squares estimate

$$\hat{\theta} = (X^\top X)^{-1} X^\top y = (X^\top X)^{-1} X^\top (X\theta + \nu) = \theta + (X^\top X)^{-1} X^\top \nu$$

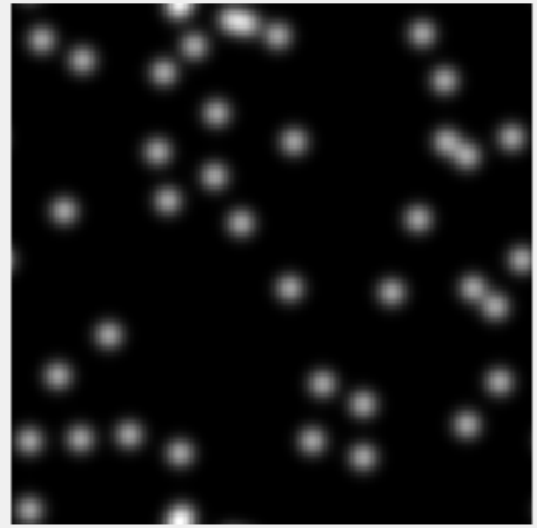
will contain θ plus amplified noise.
We will compare the least-squares estimate, the ridge estimate, and the LASSO estimate.

Example: Deblurring

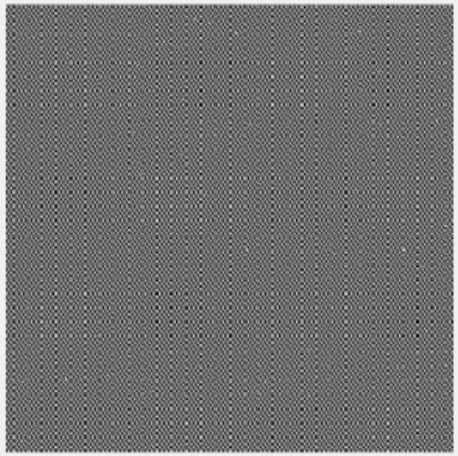
original



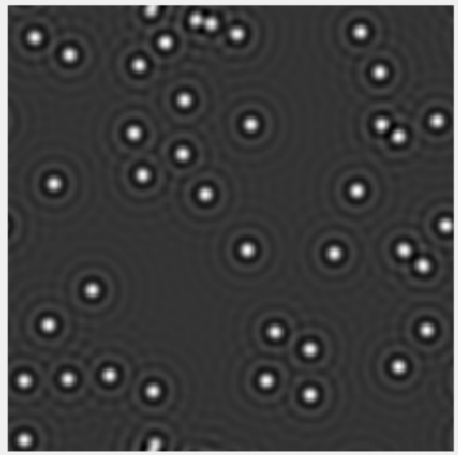
observed



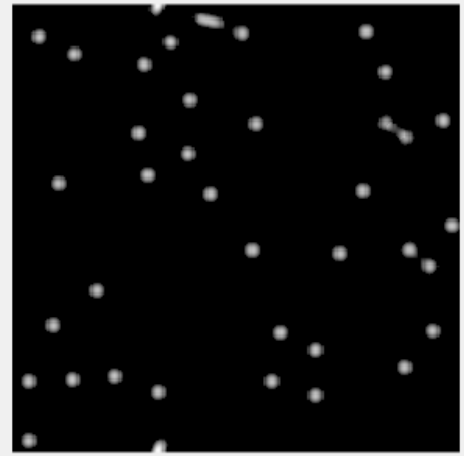
LS, MSE = 17495.1339



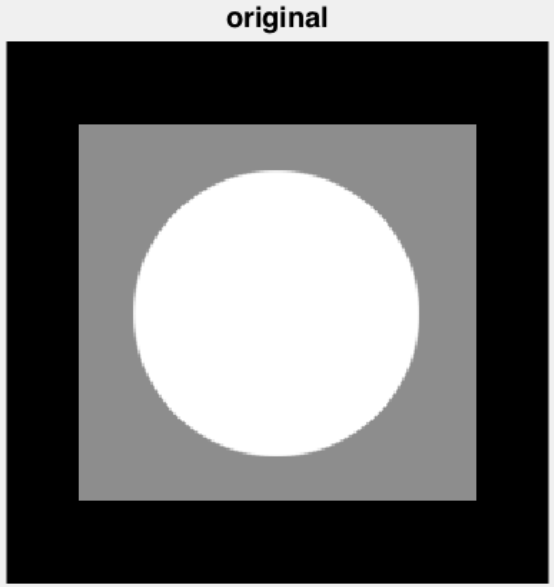
Tikhonov, MSE = 37.9969

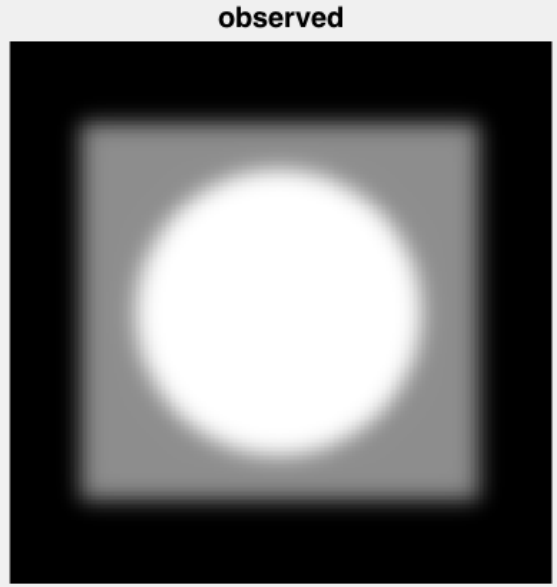


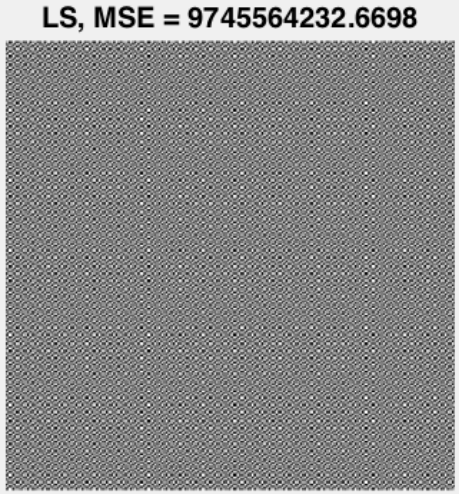
LASSO, MSE = 38.612

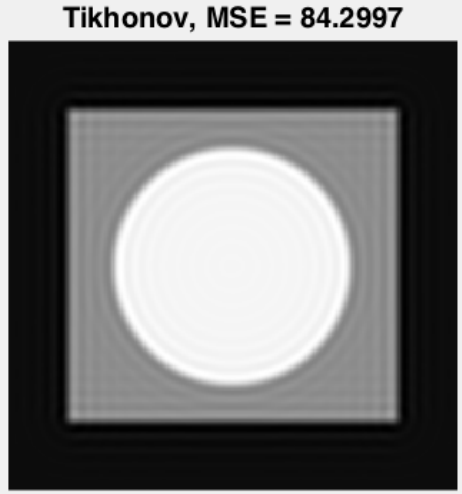


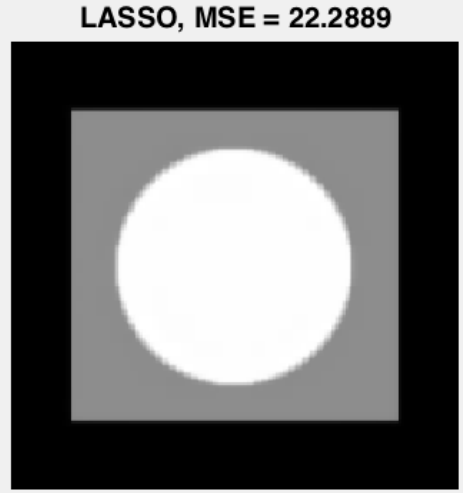
Example: Deblurring











Proof of lemma

Note that the conditional distribution must be a Gaussian, so we just need to calculate the mean and covariance of this Gaussian to fully characterize the distribution.
Let $A := -\Sigma_{12}\Sigma_{22}^{-1}$ and $t := z_1 + Az_2$. Then

$$\text{cov}[t, z_2] = \text{cov}(z_1 + Az_2, z_2) = \Sigma_{12} + A\Sigma_{22} = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$$

which means that t and z_2 are uncorrelated, and since they're Gaussian this implies that they are also independent. Thus

$$\begin{aligned}\mathbb{E}[z_1|z_2] =& \mathbb{E}[t - Az_2|z_2] = \mathbb{E}[t|z_2] - A\mathbb{E}[z_2|z_2] \\ =& \mathbb{E}[t] - Az_2 = \mu_1 + A(\mu_2 - z_2) \\ =& \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)\end{aligned}$$

Proof of lemma (cont.)

We now use the fact that for two random vectors x and y,

$$\mathsf{var}(x+y) = \mathsf{var}(x) + \mathsf{var}(y) + \mathsf{cov}(x,y) + \mathsf{cov}(y,x).$$

Thus

$$\begin{aligned}
\mathsf{var}(z_1|z_2) =& \mathsf{var}(t - Az_2|z_2) \\
=& \mathsf{var}(t|z_2) + A\mathsf{var}(z_2|z_2)A^\top - \mathsf{cov}(t, Az_2) - \mathsf{cov}(Az_2, t) \\
=& \mathsf{var}(t) = \mathsf{var}(z_1 + Az_2) \\
=& \mathsf{var}(z_1) + A\mathsf{var}(z_2)A^\top + A\mathsf{cov}(z_1, z_2) + \mathsf{cov}(z_2, z_1)A^\top \\
=& \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
=& \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\end{aligned}$$