Exercise 1: Bayesian inference

07/03/2023 – 14/03/2023

Group D: Clara Kümpel, Damola Agbelese, Yitong Li

1.1 Maximum likelihood and overfitting

Polynomial model of order P:

$$y = \sum_{k=0}^{P} \theta_k x^k + \epsilon, \qquad i.i.d.: \epsilon \sim \mathcal{N}(0, \sigma^2)$$

a) Let $f(x, \boldsymbol{\theta}) = \sum_{k=0}^{P} \theta_k x^k$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_P)^T$

Likelihood:
For a single data point y:

$$p(y|\boldsymbol{\theta}) \sim \mathcal{N}(f(x, \boldsymbol{\theta}), \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - f(x, \boldsymbol{\theta})\right)^2}{2\sigma^2}\right)$$

For **y** as a vector of data points $\boldsymbol{y} = (y_1, y_2, \dots, y_N)$:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = p(y_1, \dots y_N|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta})$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{\sum_{n=1}^{N}\left(y_n - f(x_n, \boldsymbol{\theta})\right)^2}{2\sigma^2}\right)$$

Log-likelihood:

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -N \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y_n - f(x_n, \boldsymbol{\theta})\right)^2$$

b) Maximum likelihood (ML)

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{\theta})$$

$$= \arg\min \sum_{n=1}^{N} \left(y_n - f(x_n, \boldsymbol{\theta})\right)^2$$

The ML estimator $\hat{\theta}$ turns out to be the Least Square Error (LSE) estimator.
Written with matrices:

$$\widehat{\boldsymbol{\theta}} = \arg\min \left|\left|\boldsymbol{y} - X\widehat{\boldsymbol{\theta}}\right|\right|^2$$

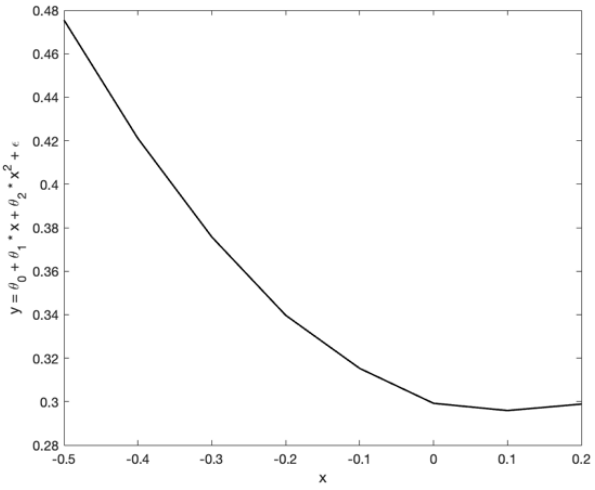
Where $y = (y_1, \dots, y_N)^T, X = \begin{pmatrix} 1 & x & x^2 & \dots & x^p \\ 1 & \dots & \dots & \dots & x^p \\ 1 & x & x^2 & \dots & x^p \end{pmatrix}$

$$\Rightarrow \frac{\partial}{\partial \theta}\left(\left|\left|\boldsymbol{y} - X\widehat{\boldsymbol{\theta}}\right|\right|^2\right) = 0$$

$$\frac{\partial}{\partial \theta}\left(\left(\boldsymbol{y} - X\widehat{\boldsymbol{\theta}}\right)^T\left(\boldsymbol{y} - X\widehat{\boldsymbol{\theta}}\right)\right) = 0$$

$$\Rightarrow \widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

c)



d) Considering the analytic solution is the same as the *polyfit* method in MATLAB:

```
%% 1.1 (d)
% P = 2:
p2 = polyfit(x,y,2);

% P = 1:
p1 = polyfit(x,y,1);

% P = 7:
p7 = polyfit(x,y,7);
```

```
p2 =

    0.5074   -0.0990    0.2996


p1 =

   -0.2512    0.3148


p7 =

  108.4740   83.2556    3.8034   -6.7386   -0.3359    0.6841   -0.1009    0.2987
```

When P = 2, the estimated parameters are much close to the true values.

```
% Log-likelihood:
N = length(x);
LL2 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p2,x)).^2)
LL1 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p1,x)).^2)
LL7 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p7,x)).^2)
```

```
LL2 =

   20.2780


LL1 =

   18.1221


LL7 =

   20.2795
```

In this case, estimators with higher P result in higher log-likelihood. All estimators perform not much differently.

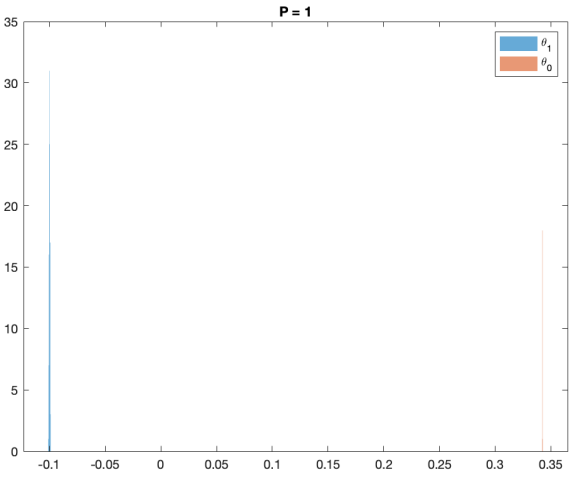
e) Using new values of x:

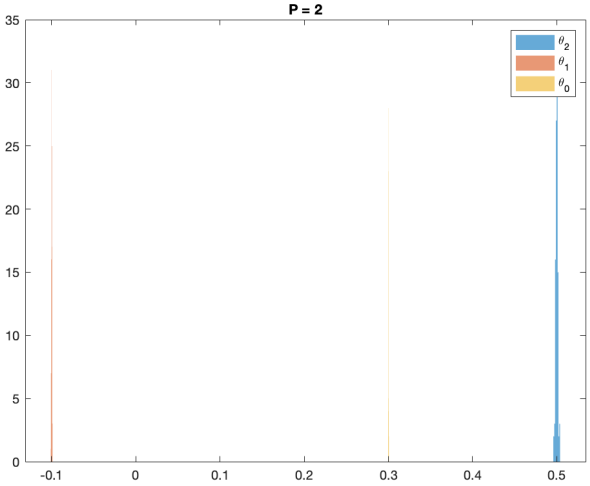
```
%% 1.1 (e)
% increase x
x = -0.5:0.01:0.5;
epsilon = sigma2 * randn(size(x));
y = theta0 + theta1 * x + theta2 * x.^2 + epsilon;

p2 = polyfit(x,y,2);
p1 = polyfit(x,y,1);
p7 = polyfit(x,y,7);

LL2 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p2,x)).^2)
LL1 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p1,x)).^2)
LL7 = -N*log(sqrt(2*pi*sigma2)) - (1/2/sigma2)*sum((y-polyval(p7,x)).^2)

```

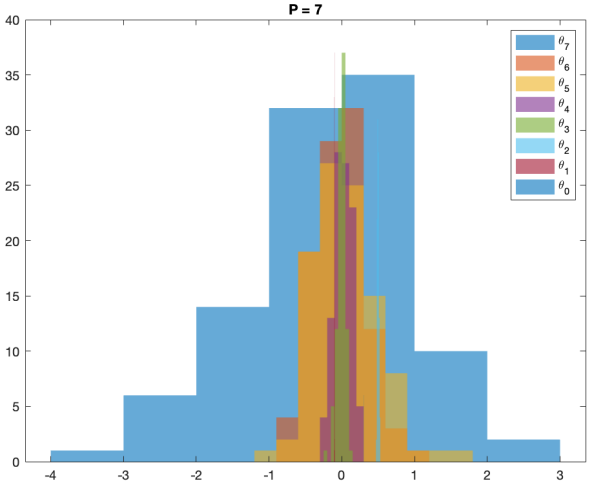
```
LL2 =

   20.2090


LL1 =

 -189.5529


LL7 =

 -246.4703
```

When fitting a new set of data with parameters generated with the other set of data, performances of estimators with P=1 and P=7 drop drastically, which means they do not generalize well when encountering different data sets.
Performance of estimator with P=2 does not change much.

f)







As the order of polynomial term increases, the estimator's consistency decreases and thus spreads on the histogram across repetitions where different error signals keep being generated.

1.2 Maximum-A-Posteriori Estimation

Parameters:

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_P)^T$$

Gaussian prior:

$$p(\boldsymbol{\theta}) = \frac{1}{\sqrt{|2\pi\Sigma_0|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu_0})^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu_0})\right)$$

With prior covariance Σ_0 and prior mean $\boldsymbol{\mu_0}$
In this exercise, $\Sigma_0 = I, \boldsymbol{\mu_0} = \mathbf{0}$

a) Posterior:

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f(x,\boldsymbol{\theta}))^2}{2\sigma^2}} \frac{1}{\sqrt{|2\pi\Sigma_0|}} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu_0})^T \Sigma_0^{-1} (\boldsymbol{\theta}-\boldsymbol{\mu_0})} / p(y)$$

$$\propto \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f(x,\boldsymbol{\theta}))^2}{2\sigma^2}} \frac{1}{\sqrt{|2\pi\Sigma_0|}} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu_0})^T \Sigma_0^{-1} (\boldsymbol{\theta}-\boldsymbol{\mu_0})}$$

Log-posterior:

$$\log\big(p(\boldsymbol{\theta}|y)\big) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{\big(y - f(x,\boldsymbol{\theta})\big)^2}{\sigma^2} - \frac{1}{2}\log(|2\pi\Sigma_0|) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu_0})^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu_0}) - \log\big(p(y)\big)$$

b) Maximum-A-Posteriori Estimation:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\big(p(\boldsymbol{\theta}|y_1, \dots, y_N)\big)$$

$$= \arg\min_{\theta} \frac{\sum_{n=1}^{N}\big(y_n - f(x_n,\boldsymbol{\theta})\big)^2}{\sigma^2} + (\boldsymbol{\theta} - \boldsymbol{\mu_0})^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu_0})$$

In our case, $\Sigma_0 = I, \boldsymbol{\mu_0} = \mathbf{0}$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta}\left(\sum_{n=1}^{N}\big(y_n - f(x_n,\boldsymbol{\theta})\big)^2 + \boldsymbol{\theta}^T\boldsymbol{\theta}\right)$$

The MAP estimator turns out to be ridge regression (LSE with L2 regularization)

Written with matrices:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta}\left(\left|\left|\boldsymbol{y} - X\boldsymbol{\theta}\right|\right|^2 + \boldsymbol{\theta}^T\boldsymbol{\theta}\right)$$

Where $y = (y_1, \dots, y_N)^T, X = \begin{pmatrix} 1\ x\ x^2 \dots x^p \\ 1 \dots \dots \dots\ x^p \\ 1\ x\ x^2 \dots x^p \end{pmatrix}$

$$\therefore \frac{\partial}{\partial\theta}\left(\left|\left|\boldsymbol{y} - X\hat{\boldsymbol{\theta}}\right|\right|^2 + \hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\theta}}\right) = 0$$

$$\frac{\partial}{\partial\theta}\left(\big(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}\big)^T\big(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}\big) + \hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\theta}}\right) = 0$$

$$\therefore \hat{\boldsymbol{\theta}} = (X^TX + I)^{-1}X^T\boldsymbol{y}$$

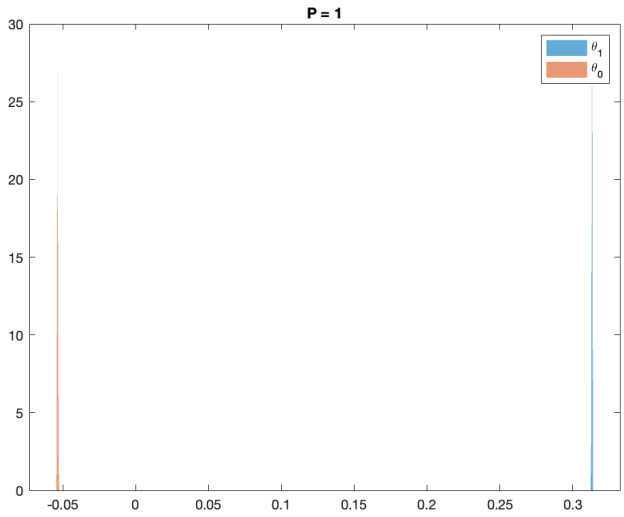
c)

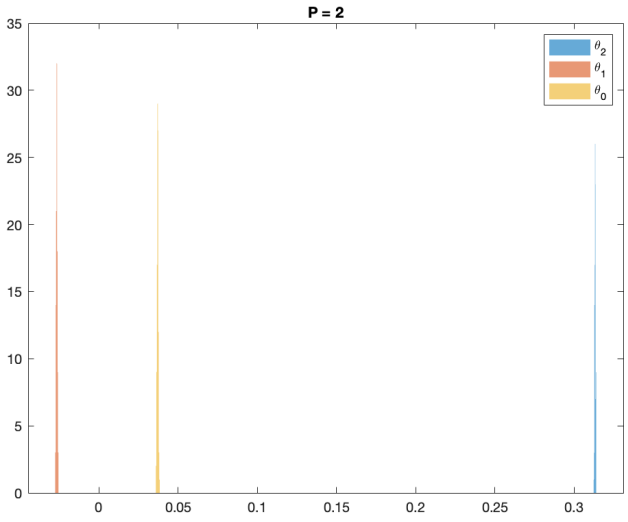
```
93      %% 1.2 (c)
94      rng(1)          % seed
95      x = -0.5:0.1:0.2;
96      epsilon = sigma2 * randn(size(x));
97      x = x';
98      X1 = [ones(size(x)),x];
99      X2 = [ones(size(x)),x,x.^2];
100     X7 = [ones(size(x)),x,x.^2, x.^3, x.^4, x.^5, x.^6, x.^7];
101     y = theta0 * ones(size(x)) + theta1 * x + theta2 * x.^2 + epsilon * ones(size(x));
102     B1 = ridge(y,X1,1);
103     B2 = ridge(y,X2,1);
104     B7 = ridge(y,X7,1);
105
```

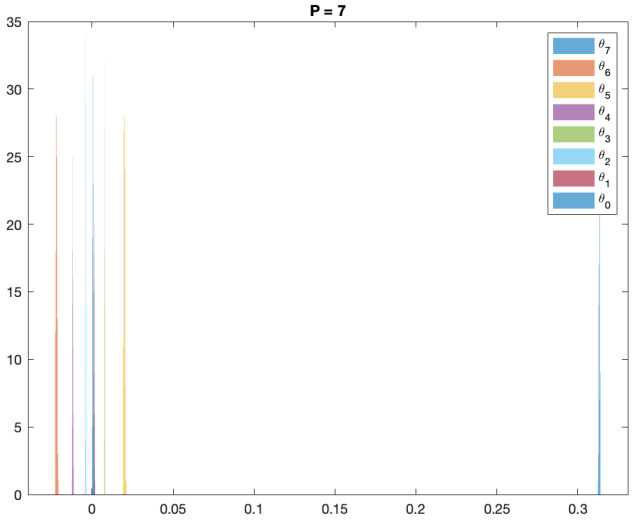
```
B1 =
    0.3105
   -0.0536

B2 =
    0.3105
   -0.0265
    0.0372

B7 =
    0.3105
   -0.0219
    0.0199
   -0.0119
    0.0077
   -0.0039
    0.0014
    0.0006
```

Compared to 1.1 (c), estimators of the first three orders keep relatively invariant, yet deviate from the true value more than 1.1 (c).
The norm of the estimator of P=7 is much smaller than 1.1 (c) where the optimization function does not have the regularization term.

d)







MAP estimators keep good consistency across repetitions.

1.3 Bayesian Inference in the Univariate Gaussian Case

$$y = x\theta + \epsilon$$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{\epsilon^2}{2\sigma_\epsilon^2}}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{(\theta-\mu_p)^2}{2\sigma_p^2}}$$

$$x \text{ is constant}$$

a) Likelihood:
Given 1 fixed data point (x, y):

$$p(y|\theta) = p(x\theta + \epsilon|\theta) \sim \mathcal{N}(\theta x, \sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{(y-\theta x)^2}{2\sigma_\epsilon^2}}$$

b) Posterior of model parameters θ

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

$$\log\big(p(\theta|y)\big) = \log\big(p(y|\theta)\big) + \log\big(p(\theta)\big) - \log\big(p(y)\big)$$

$$= -\frac{1}{2}\log(2\pi\sigma_\epsilon^2) - \frac{(y-\theta x)^2}{2\sigma_\epsilon^2} - \frac{1}{2}\log(2\pi\sigma_p^2) - \frac{(\theta-\mu_p)^2}{2\sigma_p^2} - \log\big(p(y)\big)$$

$$= -\log\left(2\pi\sigma_\epsilon\sigma_p \cdot p(y)\right) - \frac{(y-\theta x)^2}{2\sigma_\epsilon^2} - \frac{(\theta-\mu_p)^2}{2\sigma_p^2}$$

$$\therefore p(\theta|y) \propto \exp\left(-\frac{(y-\theta x)^2}{2\sigma_\epsilon^2} - \frac{(\theta-\mu_p)^2}{2\sigma_p^2}\right)$$

Precision: $\lambda_\epsilon = \frac{1}{\sigma_\epsilon^2}, \lambda_p = \frac{1}{\sigma_p^2}$

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(x^2\lambda_\epsilon\left(\theta - \frac{y}{x}\right)^2 + \lambda_p(\theta-\mu_p)^2\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(x^2\lambda_\epsilon\left(\theta^2 - \frac{2y}{x}\theta + \frac{y^2}{x^2}\right) + \lambda_p(\theta^2 - 2\mu_p\theta + \mu_p^2)\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left((x^2\lambda_\epsilon + \lambda_p)\theta^2 - 2(\lambda_\epsilon xy + \lambda_p\mu_p)\theta + \lambda_\epsilon y^2 + \lambda_p\mu_p^2\right)\right)$$

$$= \exp\left(-\frac{1}{2}(x^2\lambda_\epsilon + \lambda_p)\left(\theta^2 - \frac{2(\lambda_\epsilon xy + \lambda_p\mu_p)}{x^2\lambda_\epsilon + \lambda_p}\theta + \frac{\lambda_\epsilon y^2 + \lambda_p\mu_p^2}{x^2\lambda_\epsilon + \lambda_p}\right)\right)$$

$$= \exp\left(-\frac{1}{2}(x^2\lambda_\epsilon + \lambda_p)\left(\theta^2 - \frac{2(\lambda_\epsilon xy + \lambda_p\mu_p)}{x^2\lambda_\epsilon + \lambda_p}\theta + \frac{(x^2\lambda_\epsilon + \lambda_p)(\lambda_\epsilon y^2 + \lambda_p\mu_p^2)}{(x^2\lambda_\epsilon + \lambda_p)^2}\right)\right)$$

$$= \exp\left(-\frac{1}{2}(x^2\lambda_\epsilon + \lambda_p)\left(\theta^2 - \frac{2(\lambda_\epsilon xy + \lambda_p\mu_p)}{x^2\lambda_\epsilon + \lambda_p}\theta + \frac{(\lambda_\epsilon xy + \lambda_p\mu_p)^2}{(x^2\lambda_\epsilon + \lambda_p)^2}\right)\right)$$

Define:

$$\lambda_{post} = x^2\lambda_\epsilon + \lambda_p$$

We get:

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\lambda_{post}\left(\theta^2 - 2\lambda_{post}(\lambda_\epsilon xy + \lambda_p\mu_p)\theta + \lambda_{post}^2(\lambda_\epsilon xy + \lambda_p\mu_p)^2\right)\right)$$
$$= \exp\left(-\frac{1}{2}\lambda_{post}\left(\theta - \lambda_{post}(\lambda_\epsilon xy + \lambda_p\mu_p)\right)^2\right)$$

Define:

$$\mu_{post} = \lambda_{post}(\lambda_\epsilon xy + \lambda_p\mu_p)$$

We get:

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\lambda_{post}(\theta - \mu_{post})^2\right) = \exp\left(-\frac{(\theta - \mu_{post})^2}{\sigma_{post}^2}\right)$$

Where:

$$\frac{1}{\sigma_{post}^2} = \lambda_{post} = \frac{x^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_p^2}$$