

# ST451 - Lent term

## Bayesian Machine Learning

Kostas Kalogeropoulos

Bayesian Inference Concepts - Linear Regression

# Outline

- 1 Bayesian Inference Concepts
- 2 Linear Regression
- 3 Bayesian Linear Regression
- 4 Optional: Bayes Estimators and Decision Theory

# Outline

- 1 Bayesian Inference Concepts
- 2 Linear Regression
- 3 Bayesian Linear Regression
- 4 Optional: Bayes Estimators and Decision Theory

# Summary of last lecture

- Define **model (likelihood)**  $f(y|\theta)$  and **prior**  $\pi(\theta)$ .
- Obtain **posterior** via Bayes theorem

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta)$$

- **Bayes Estimators:** posterior mean, median or mode.
- **Bayesian 95% credible intervals:** 2.5-th and 97.5-th percentiles.  $\theta$  is in them with probability 95%
- Bayesian forecasting for future data  $y_n$  via the (posterior-)**predictive distribution**

$$f(y_n|y) = \int f(y_n|\theta)\pi(\theta|y)d\theta.$$

## Example: Normal-Normal

**Likelihood:** Let  $y = (y_1, \dots, y_n)$  be a random sample from a  $N(\theta, \sigma^2)$  -  $\sigma^2$  known. The **likelihood** is given by the joint density of the sample

$$f(y|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right)$$

**Prior:** We assume the **Normal prior**  $N(\mu, \tau^2\sigma^2)$  for  $\theta$ , which gives

$$\pi(\theta) \propto \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2\sigma^2}\right)$$

## Example Normal-Normal - completing the square

**Posterior:** The **posterior** can then be obtained as

$$\begin{aligned}\pi(\theta|y) &\propto f(y|\theta)\pi(\theta) \propto \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2\sigma^2}\right) \\ &\propto \dots \propto \exp\left(-\frac{\theta^2 - 2\theta\frac{\bar{y}\tau^2 + \mu\frac{1}{n}}{(\frac{1}{n} + \tau^2)}}{2\frac{\frac{\sigma^2}{n}\tau^2}{(\frac{1}{n} + \tau^2)}}\right) \stackrel{\mathcal{D}}{=} \mathbf{N}\left(\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\tau^2 + \frac{1}{n}}, \frac{\tau^2\frac{\sigma^2}{n}}{\tau^2 + \frac{1}{n}}\right)\end{aligned}$$

**Bayes Estimators:** Posterior mean, median and mode are **all equal** to

$$\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\tau^2 + \frac{1}{n}} = \left(1 - \frac{\tau^2}{\frac{1}{n} + \tau^2}\right)\mu + \frac{\tau^2}{\frac{1}{n} + \tau^2}\bar{y},$$

a **weighted average** between the prior mean  $\mu$  and the MLE  $\bar{y}$

## Example Normal-Normal (cont'd)

**Bayesian 95% credible intervals:** With  $\mathcal{Z}_{2.5}$  being the 2.5-th percentile of the  $N(0, 1)$ , we get

$$\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\tau^2 + \frac{1}{n}} \pm \mathcal{Z}_{2.5} \sqrt{\frac{\tau^2 \frac{\sigma^2}{n}}{\tau^2 + \frac{1}{n}}}$$

**Predictive distribution:** Given  $\theta$  a future observation  $y_n$  will be  $N(\theta, \sigma^2)$ , so using standard properties of Normal distributions we get that the predictive is also a Normal with mean

$$\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\tau^2 + \frac{1}{n}}$$

and variance

$$\frac{\tau^2 \frac{\sigma^2}{n}}{\tau^2 + \frac{1}{n}} + \sigma^2$$

## Example Normal-Normal: Jeffreys prior

To find **Jeffreys prior** we perform the following steps:

$$\log f(y|\theta) = -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}, \quad \frac{\partial}{\partial \theta} \log f(y|\theta) = \frac{\sum_{i=1}^n (y_i - \theta)}{\sigma^2}$$

$$\mathcal{I}(\theta|y) = -E_Y \left( \frac{\partial^2}{\partial \theta^2} \log f(y|\theta) \right) = -E_Y \left( -\frac{n}{\sigma^2} \right) = \frac{n}{\sigma^2}$$

Hence the Jeffreys prior:  $\pi(\theta) \propto 1$ .



## Example Normal-Normal: Inference with Jeffreys prior

**Posterior:** The **posterior** from Jeffreys prior is

$$\begin{aligned}\pi(\theta|y) &\propto \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{n\theta^2 - 2\theta \sum_{i=1}^n y_i}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\theta^2 - 2\theta\bar{y}}{2\frac{\sigma^2}{n}}\right) \stackrel{\mathcal{D}}{=} N\left(\bar{y}, \frac{\sigma^2}{n}\right)\end{aligned}$$

**Bayes Estimators:** All are the **same** as the MLE, i.e.  $\bar{y}$ .

**Bayesian 95% credible intervals:**  $\bar{y} \pm \mathcal{Z}_{2.5}\sqrt{\frac{\sigma^2}{n}}$ .

**Same** as in the frequentist case but with different interpretation.

**Predictive distribution:** We get  $N(\bar{y}, \frac{\sigma^2}{n} + \sigma^2)$ .

# Outline

- 1 Bayesian Inference Concepts
- 2 Linear Regression**
- 3 Bayesian Linear Regression
- 4 Optional: Bayes Estimators and Decision Theory

# Motivating Example: Prostate Cancer

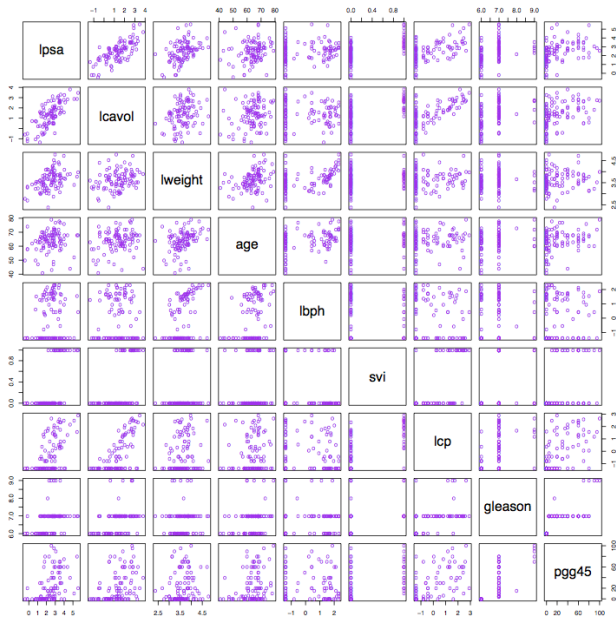
Data from the following study on prostate cancer

*Stamey, T., et al. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, Journal of Urology 16: 1076-1083.*

Examines association between the level of **prostate-specific antigen** (PSA) and a number of **clinical measures** in men who were about to receive a radical prostatectomy.

The variables are cancer volume (lcavol), prostate weight (lweight), age, amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

# Motivating Example: Data



## Motivating Example: Aims of the analysis

- Determine the level of PSA on a **future** patient based on the clinical measurements. Otherwise detailed histological and morphometric analysis is required.
- How is each of these variables **associated** with PSA? Is the association **linear**?
- Are any of these variables **redundant** in the presence of the others? Which are the **most important**?
- Are there any **synergies** between these variables?

## Data setup

Data consist of measurements on all these variables on several individuals.

We typically denote value of the **response** variable, in this case log-PSA, on the individual  $i$  with  $Y_i$ . The vector  $Y = (Y_1, \dots, Y_n)$  is assumed to be a  $n$ -dimensional random variable.

The remaining variables  $X_1, \dots, X_p$  contain the clinical measurements.  $X_{ji}$  refers to the value of the clinical measurement  $j$  of the individual  $i$ .

The  $X$ 's are **not assumed to be random** they are treated as fixed inputs, with  $Y$  being regarded as the **output**.

# Linear Regression

The linear regression model is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, \dots, n.$$

It is more convenient to use matrix algebra. Define  $y = (Y_1, \dots, Y_n)$ ,  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  the **error terms**,  $\beta = (\beta_0, \dots, \beta_p)^T$  denoting the **regression coefficients** and the **design matrix**

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{pmatrix}.$$

Then we can rewrite the model in matrix notation as

$$y = X\beta + \epsilon,$$

## Example: Prostate Cancer Regression Coefficients

Test your understanding on interpreting coefficients in the table below:

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	-0.14	0.10	-1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	-0.29	0.15	-1.87
gleason	-0.02	0.15	-0.15
pgg45	0.27	0.15	1.74



# MLE of Linear Regression

The **MLE** and the **least squares** estimators can be shown to be:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The **variance** of the MLE is given by

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1},$$

where  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2$  or  $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2$ .

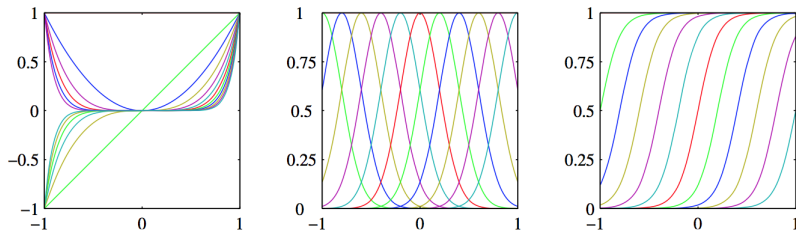
The distribution of the MLE is the **t-distribution** with  $n - p$  degrees of freedom.

# Linear Basis Functions

The model is linear its parameters  $\beta$  not  $X$  so we can replace each  $X_i$  with  $\phi(X_i)$ . We can then write the model as

$$y = \phi(X)\beta + \epsilon$$

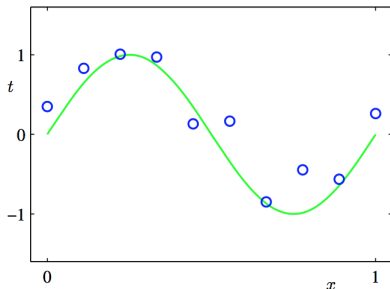
Examples include dummy variables, polynomial terms, Gaussian kernels, sigmoid functions etc.



**Figure 3.1** Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

## Example: Polynomial Curve Fitting

Let's consider the following example on simulated data. The generating process is  $\sin(2\pi x)$  and we observe this function on 10 different points in  $[0, 1]$  with independent Gaussian error.

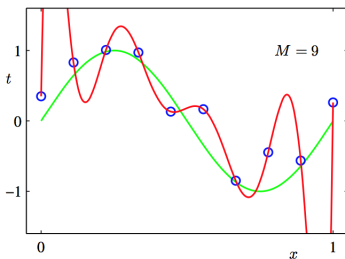
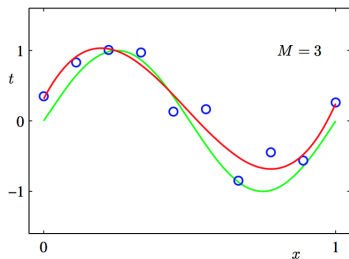
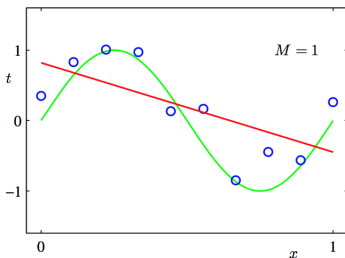
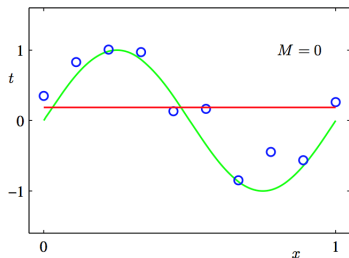


We fit a linear regression model with polynomial basis functions

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_M x_i^M + \epsilon_i$$

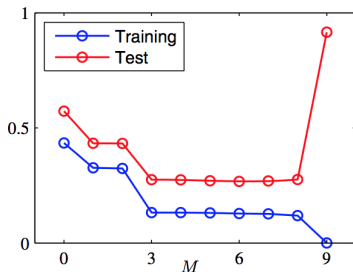
## Example: Fit of different polynomials

For each order of polynomial we find the MLE and plot the corresponding function to assess its fit



# Training and Test Error - Overfitting

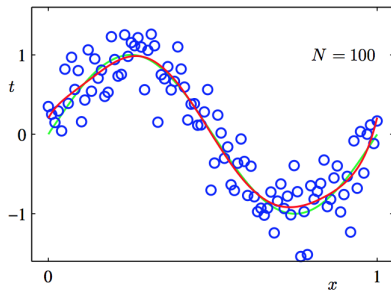
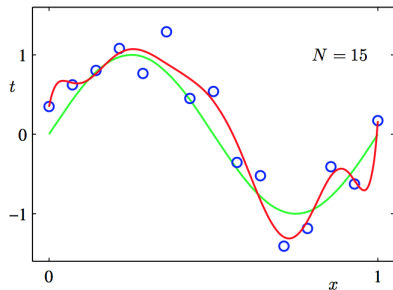
In addition to the training data set of 10 observations we simulate 100 more points in the same way and assess the training and test error.



The test error decreases until  $M = 6$  and increase afterwards. The training error keeps decreasing and drops to 0 for  $M = 9$ . MLE leads to **overfit**.

# Model Complexity

In order to identify the data generating process our model should not be too **complex** compared to the data we are training it. Increasing the data improves learning for the model with  $M = 9$ .



# Parameter Estimates

One way to reduce **model complexity** is to reduce the number of predictors. Not necessarily the best way.

More insight is obtained by looking at the parameter estimates for polynomials of different order.

$M = 0$	$M = 1$	$M = 6$	$M = 9$
0.19	0.82	0.31	0.35
	-1.27	7.99	232.37
		-25.43	-5321.83
		17.37	48568.31
			-231639.30
			640042.26
			-1061800.52
			1042400.18
			-557682.99
			125201.43

# Regularisation

Instead of removing predictors we could instead **restrict** them closer to 0. In the least squares criterion

$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_i + \cdots + \beta_M x_i^M)^2,$$

we add a **penalty term**

$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_i + \cdots + \beta_M x_i^M)^2 + \lambda \sum_{i=1}^M \beta_i^2.$$

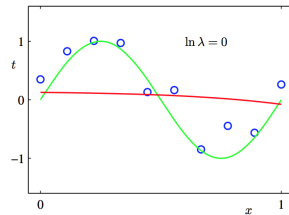
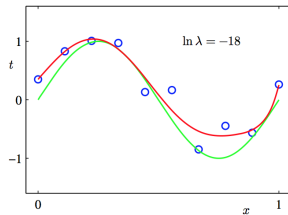
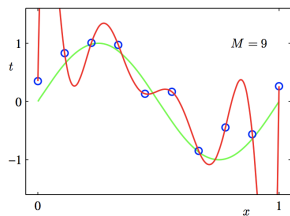
In the general case of a linear regression model  $y = X\beta + \epsilon$  the above is minimised at the point

$$\hat{\beta}^\lambda = (X^T X + \lambda^2 I_p)^{-1} X^T y$$



# Output From Regularised Approach

$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
0.35	0.35	0.13
232.37	4.74	-0.05
-5321.83	-0.77	-0.06
48568.31	-31.97	-0.05
-231639.30	-3.89	-0.03
640042.26	55.28	-0.02
-1061800.52	41.32	-0.01
1042400.18	-45.95	-0.00
-557682.99	-91.53	0.00
125201.43	72.68	0.01



# Outline

- 1 Bayesian Inference Concepts
- 2 Linear Regression
- 3 Bayesian Linear Regression**
- 4 Optional: Bayes Estimators and Decision Theory

# Bayesian Linear Regression

Techniques like this come under the names **shrinkage**, **ridge regression** or **weight decay** in the context of neural networks.

But how can it be justified from a Statistical Inference point of view?

Adjusting for overfit is automatic under the Bayesian framework and comes in a natural way.

It turns out that this is a special case of **Bayesian Linear Regression**.

## Case of known $\sigma^2$

We will first assume that  $\sigma^2$  is **known**.

The **likelihood**  $f(y|\beta)$  is the  $N(X\beta, \sigma^2 I_n)$ , where  $I_n$  is the identity matrix of dimension  $n$ .

The **prior** on  $\beta$  can be set to  $N(\mu_0, \sigma^2 \Omega_0)$

The **posterior** is then the  $N(\mu_n, \sigma^2 \Omega_n)$ , where

$$\begin{aligned}\Omega_n &= (X^T X + \Omega_0^{-1})^{-1}, \\ \mu_n &= (X^T X + \Omega_0^{-1})^{-1}(\Omega_0^{-1} \mu_0 + X^T y),\end{aligned}$$

If we set  $\mu_0 = 0$  and  $\Omega_0 = g^2 I_p$  the **Bayes Estimator** coincides with the **ridge regression estimator**.

# Notes on Bayesian Linear Regression

The Bayes estimate is a weighted average between the prior mean and the MLE.

The prior  $N(\mu_0, \sigma^2 \Omega_0)$  **shrinks** the parameters to  $\mu_0$ . This can be interpreted as **prior information**.

The amount of force is determined by  $\Omega_0$ . Prior can be viewed as a **tuning parameter** in the Machine Learning context.

The Bayes/ridge regression estimator is **biased** but has smaller variance due to the 'shrinking effect'.

# Bayesian Lasso

If we use the **Laplace** prior  $\text{La}(0, 1/\gamma)$  for  $\beta$  the Bayes estimator corresponding to the **posterior mode** is the Lasso Regression estimator.

The Laplace prior can also be written as a hierarchical Normal-Exponential prior:

$$\beta_i \sim N\left(0, \sigma^2 \tau_i^2\right), \quad \tau_i^2 \sim \text{Exponential}\left(\frac{\gamma^2}{2}\right),$$

where  $\lambda = \gamma/\sigma$ .

Note however that the posterior mean and median provide **different Bayes estimators**.

## Example: Results

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	-0.141		-0.046		-0.152	-0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	-0.288		0.000		-0.051	0.079
gleason	-0.021		0.040		0.232	0.011
pgg45	0.267		0.133		-0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

## Predictive distribution

For the predictive distribution of a **new** observation  $y_*$  given a **new** set of covariates  $X_*$ , we need to derive

$$f(y_*|y) = \int_0^\infty f(y_*|X_*, \beta, \sigma^2) \pi(\beta|y, X, \sigma^2) d\beta,$$

where the first term is  $N(X_*\beta, \sigma^2)$  and the second term is  $N(\mu_n, \sigma^2\Omega_n)$ .

Using standard properties of the Normal distribution we get that the **predictive** distribution is

$$N\left(\mu_n X_*, \sigma^2 + X_*^T \sigma^2 \Omega_n X_*\right)$$



# Outline

- 1 Bayesian Inference Concepts
- 2 Linear Regression
- 3 Bayesian Linear Regression
- 4 Optional: Bayes Estimators and Decision Theory

# Detour: Statistical Decision Theory

Given  $f(y|\theta)$ , a **statistical decision problem** consists of

- 1 The parameter space  $\Theta$ .
- 2 A set  $\mathcal{A}$  of all possible actions  $a$ , e.g.  $\hat{\theta}$ , choice of  $H_0$  vs  $H_1$ .
- 3 A **loss function**  $L(a, \theta) : \mathcal{A} \times \Theta \rightarrow \mathcal{R}$ , reflecting the loss for action  $a$  and true parameter value  $\theta$ .
- 4 One or more **decision rules**, i.e. functions  $\delta(y) : \mathcal{R} \rightarrow \mathcal{A}$  that indicate the action  $a$  based on  $y$ .

# Frequentist Risk

The aim is to minimise some kind of risk, e.g. the **frequentist risk**.

$$R(\delta(y), \theta) = E_{Y|\theta} (L(\delta(y), \theta)) = \int L(\delta(y), \theta) f(y|\theta) dy.$$

**Example:** Let  $\delta(y)$  be a point estimator  $\hat{\theta}$  and  $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ . Then  $R(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$ , i.e. the Mean Squared Error. of  $\hat{\theta}$ .

In frequentist inference we aim to minimising the frequentist risk. Not always easy because it **depends on unknown  $\theta$** , hence we want the optimal  $\delta(y)$  for all  $\theta$ .

Often rely on **conservative** approaches such as the minimax or ad-hoc criteria such as minimum variance unbiased estimators or most powerful tests.

# Bayes Risk

In Bayesian inference we minimise the **Bayes Risk** given a prior  $\pi(\theta)$

$$r(\delta(y), \pi(\theta)) = E_{\theta} [R(\delta(y), \theta)] = \int R(\delta(y), \theta) \pi(\theta) d\theta.$$

Easier as it doesn't depend on  $\theta$ , essentially we are 'averaging' the frequentist risk over  $\theta$  according to  $\pi(\theta)$ .

So **averaging** over potential  $y$  still takes place in Bayesian inference but only under the Bayesian design where  $\pi(\theta)$  is also involved.

The decision rule that minimises Bayes risk is called **Bayes rule**.

# Bayes Estimators

- Bayes estimators minimise the Bayes risk: posterior mean, median and mode corresponds to quadratic, absolute and 0 – 1 error loss functions respectively.
- They are typically **biased** (in case of proper priors) but they tend to have lower variance and are typically **admissible** estimators (no other estimator has smaller frequentist risk for all  $\theta$ ).
- Asymptotically they **converge** to the MLEs.

# Today's lecture - Reading

Murphy: 1.7, 5.3.1, 5.3.3 ,5.7.1, 7.5, 7.6.1 and 7.6.2

Bishop: 1.1, 3.1.1, 3.1.4, 3.3.1 and 3.3.2