ST308 - Lent term
Bayesian Inference

Kostas Kalogeropoulos

Bayesian Linear Regression

Outline

Topics covered: Interpretation of coefficients, linear Basis functions, Least Squares / Maximum Likelihood estimator, Overfit, Ridge and Lasso Regression, Prediction, Model Choice.

Outline

Motivating Example: Prostate Cancer

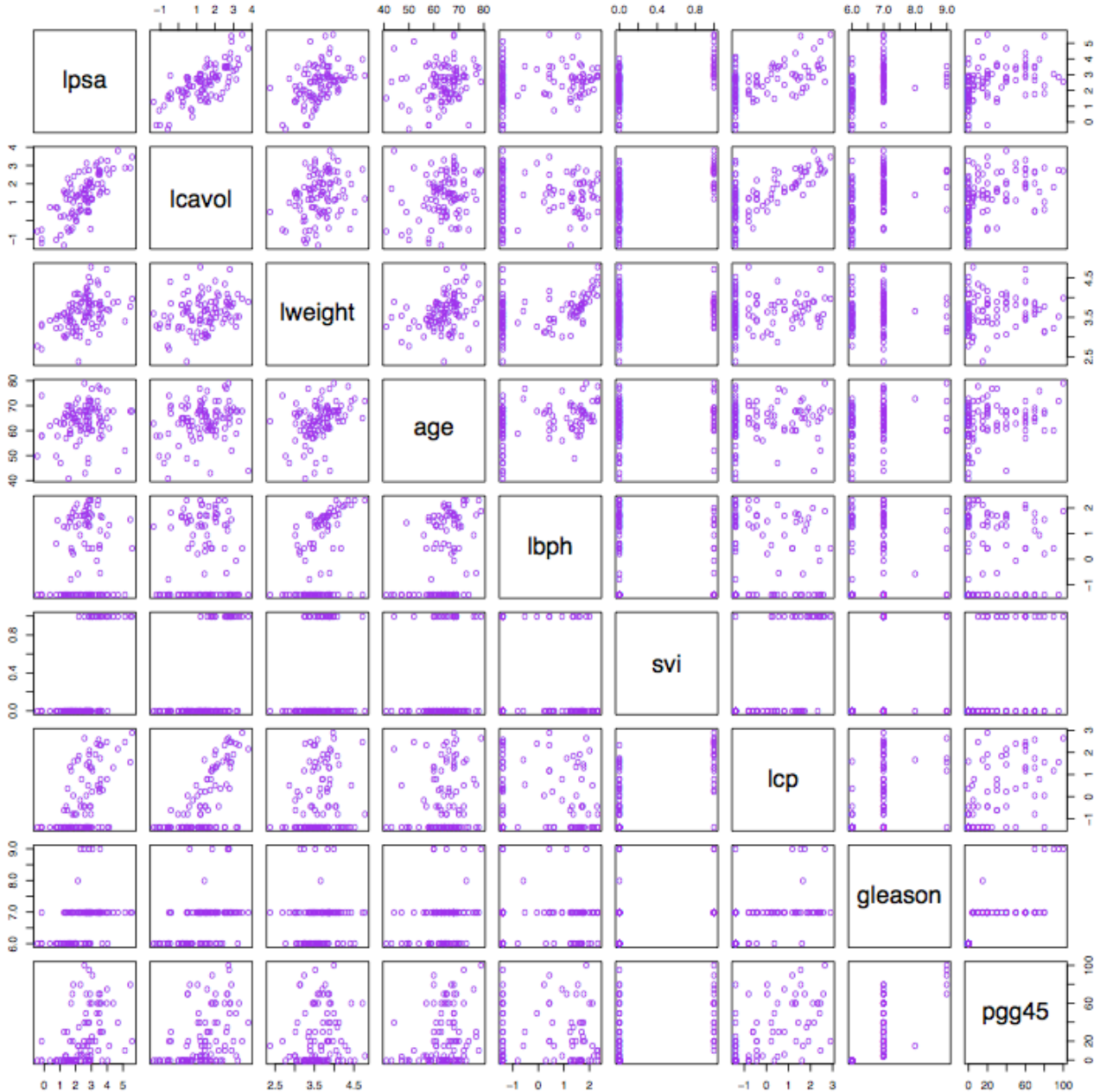
Data from the following study on prostate cancer

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 $Y = (Y_1, \ldots, Y_n)$ is assumed to be a $n-$dimensional random variable.

The remaining variables $X_1, \ldots, X_p$ 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

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

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

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

Then we can rewrite the model in matrix notation as

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

Interpretation of coefficients for continuous inputs

Consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$. Typically $E[\epsilon_i] = 0$, hence

$$E[Y_i | X_{1i} = x] = \beta_0 + \beta_1 x$$

Note that $E[Y_i | X_{1i} = 0] = \beta_0$. Hence β_0 is the expected value of the variable Y for $X_1 = 0$.

Consider $X_{1i} = x$ and $X_{1i} = x + 1$. Note that

$$E[Y_i | X_{1i} = x + 1] - E[Y_i | X_{1i} = x] = \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x = \beta_1$$

Hence β_1 is the expected change in Y for a unit change in X.

Now consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$. Do the interpretation of β_0 and β_1 change?

Example: Prostate Cancer Regression Coefficients

Test your understanding 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

Qualitative inputs

Suppose that we have a categorial variable G as input, taking the values, $\{A, B, C\}$.

We can regress the response variable Y into G using dummy variables. This is done by the following procedure:

- Choosing a reference category -say- A.
- For the remaining categories, B and C, create indicator (dummy) variables.
- For example in the case of category B form the variable X_Bi taking the value 1 if the individual i is in category B and 0 otherwise.
- Use the regression model

$$Y_i = \beta_0 + \beta_B X_{Bi} + \beta_C X_{Ci} + \epsilon_i$$

The model above is also known as ANOVA.

Qualitative inputs

In the model above note that $E[Y_i|X_{Bi} = 0, X_{Ci} = 0]$. This is equal to β_0 that correspond to the mean of Y in A.

If we take $E[Y_i|X_{Bi} = 1, X_{Ci} = 0] - E[Y_i|X_{Bi} = 0, X_{Ci} = 0]$ we will get

$$\beta_0 + \beta_B - \beta_0 = \beta_B$$

Hence β_B is the mean difference of Y between B and A.

The parameter β_C is interpreted in a similar manner.

Selecting a different reference category (e.g. C) will give an equivalent model with slightly different interpretation.

Linear Basis Functions

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

$$\begin{aligned} Y &= h(X)\beta + \epsilon \text{ or else} \\ Y_i &= \beta_0 + \beta_1 h_1(X)_i + \cdots + \beta_m h_m(X)_i + \epsilon_i \end{aligned}$$

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

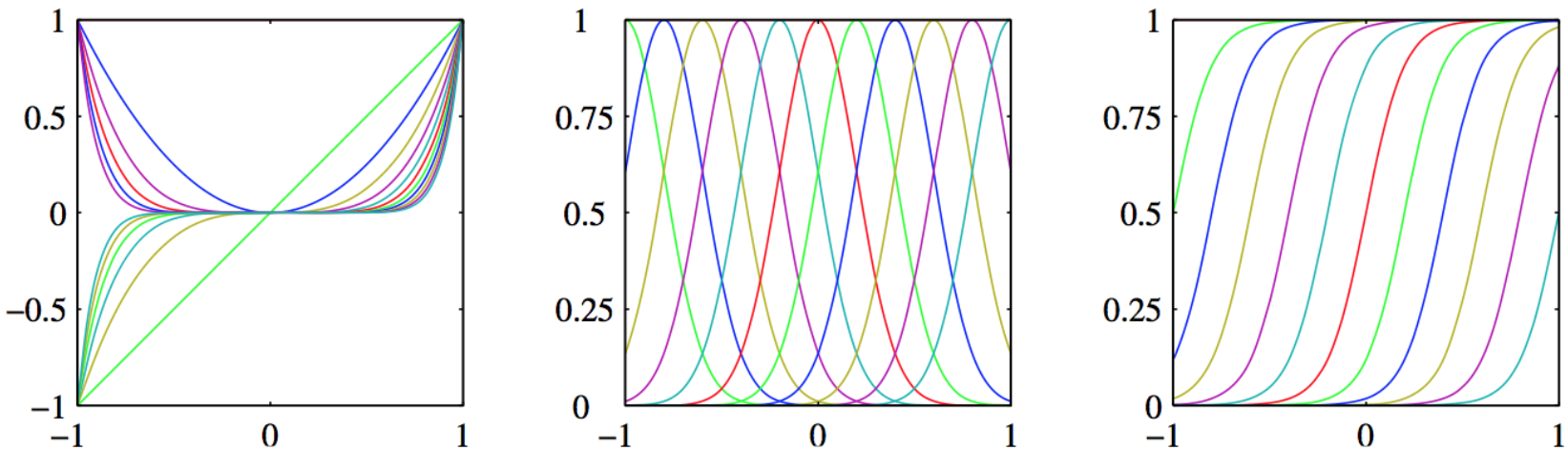


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

MLE of Linear Regression

We assume that the errors ϵ are distributed as $N(0_n,\ \sigma^2 I_n)$ where 0_n is the vector of n zeros and I_n is the identity matrix of dimension n.

Hence $y \sim N(X\beta,\ \sigma^2 I_n)$. The likelihood function can be written as

$$\begin{aligned} f(y|X,\beta,\sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)^2\right\}, \end{aligned}$$

where X_i denotes the i-th row of the matrix X.

Maximising the likelihood wrt β is the same as minimising the blue terms wrt β. Hence MLE is the same as the least squares estimator.

MLE of Linear Regression

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

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

The variance of the MLE is given by

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

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

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

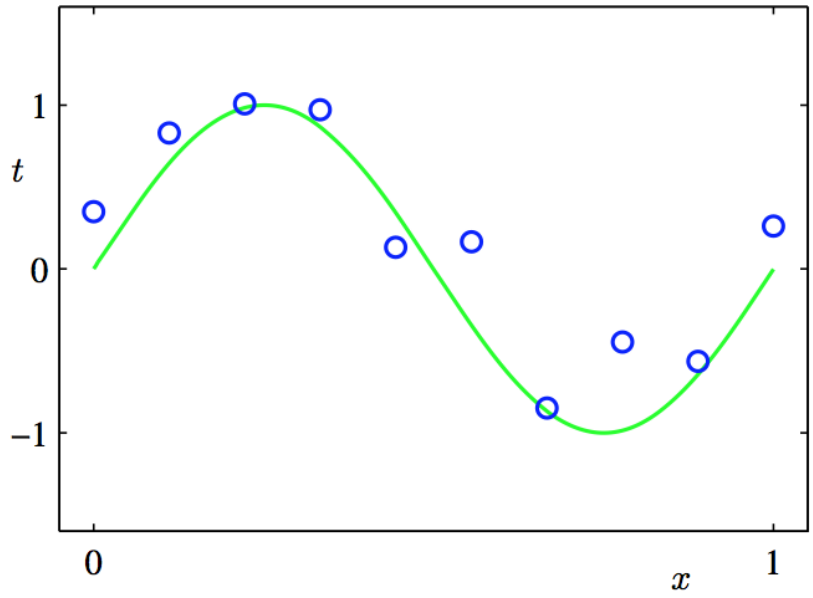
Example: Prostate Cancer Regression Coefficients

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

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.

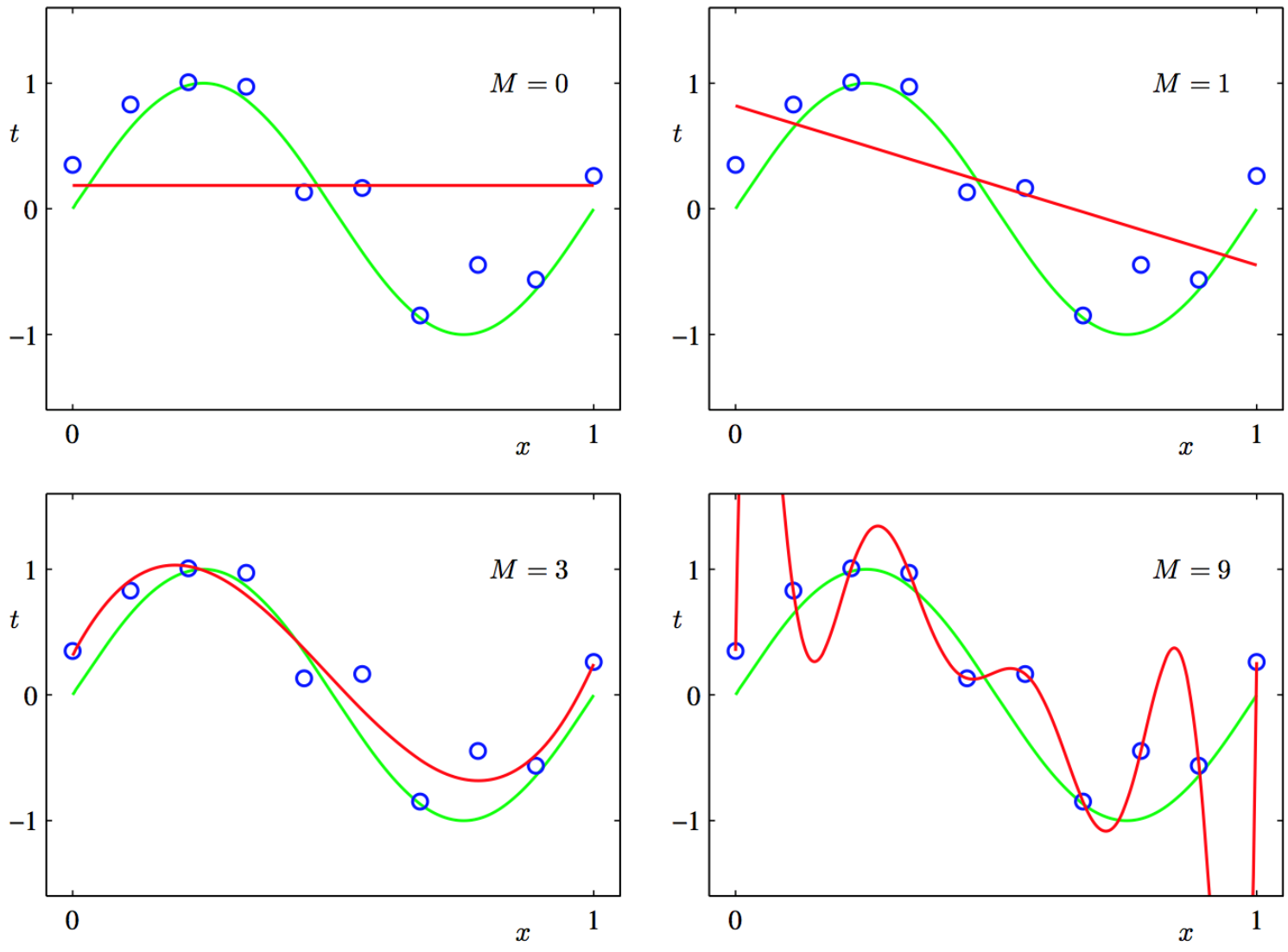


We fit a linear regression model with polynomial basis functions

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

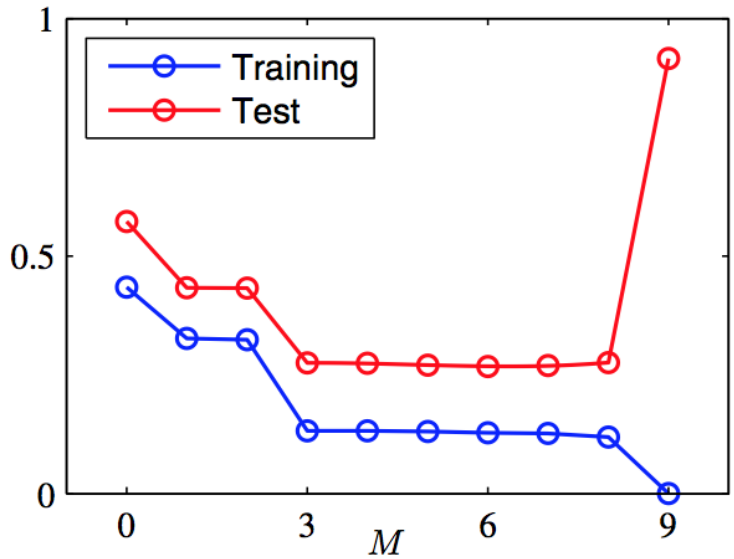
Example: Fit of different polynomials

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



Training and Test Error - Overfitting

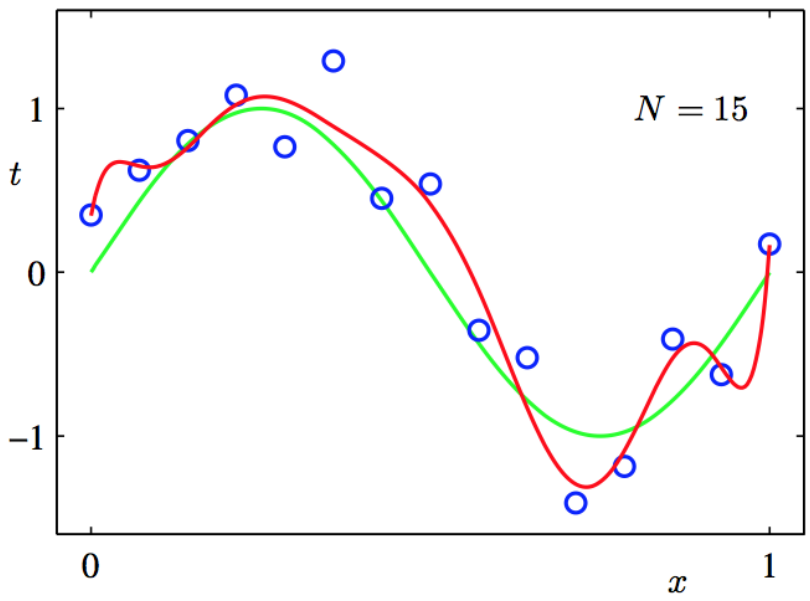
In addition to the training data set of 10 observations we simulate 100 more points in the same way and assess the training and test error.

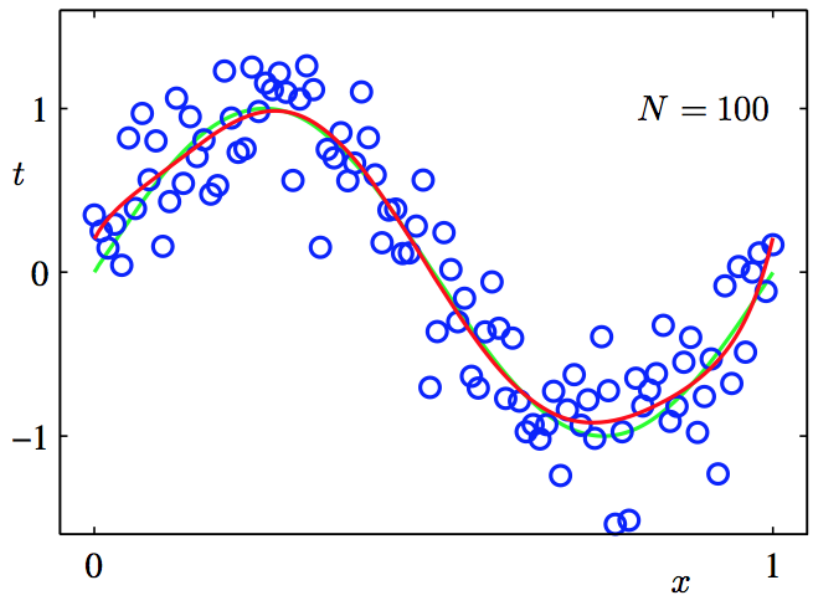


The test error decreases until $M = 6$ and increase afterwards. The training error keeps decreasing and drops to 0 for $M = 9$. MLE leads to overfit.

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

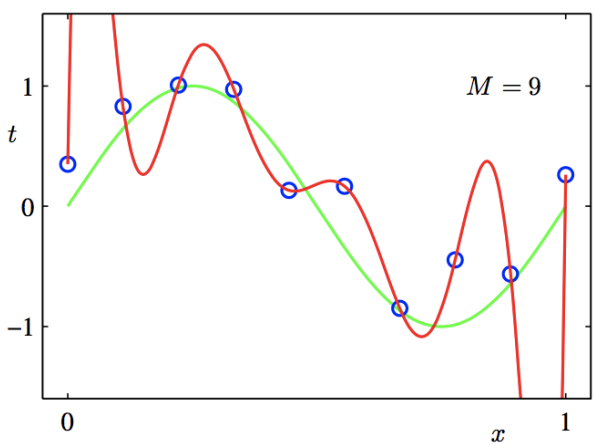
$$\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i + \cdots + \beta_M x_i^M)^2 + \lambda \sum_{i=1}^{M} \beta_i^2.$$

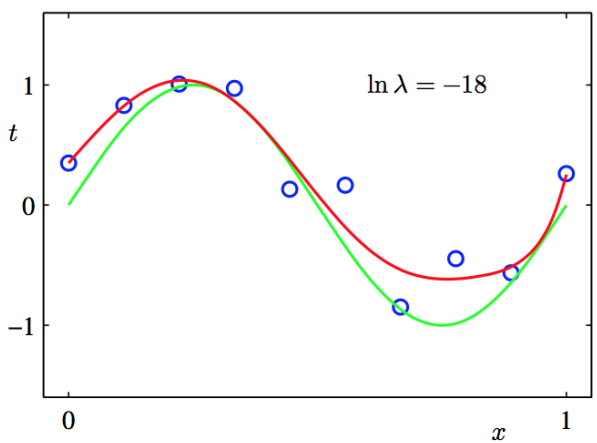
In the general case of a linear regression model $y = X\beta + \epsilon$ the above is minimised at the point

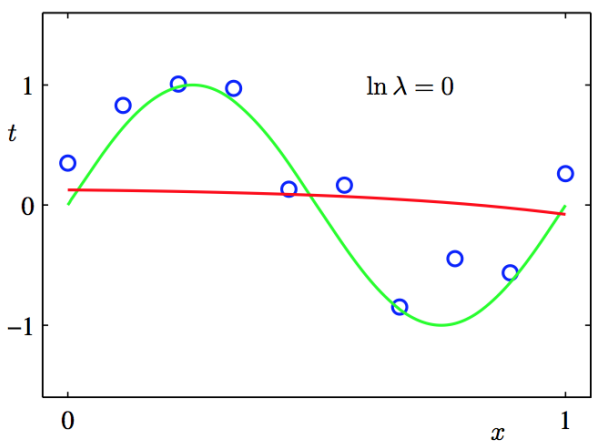
$$\hat{\beta}^{\lambda} = (X^T X + \lambda^2 I_p)^{-1} X^T y$$

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

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 σ^2

We will first assume that σ^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 β 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 with $g = 1/\lambda$.

Notes on Bayesian Linear Regression

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

The prior N(μ_0, $\sigma^2\Omega_0$) shrinks the parameters to μ_0. This can be interpreted as prior information.

The amount of force is determined by Ω_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 Lassso

If we use the Laplace prior La$(0, 1/\gamma)$ for β 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

Unit information prior for Linear Regression

In the linear regression model, the Fisher information for β based on n observations is

$$I(\beta) = -E\left[-\frac{1}{\sigma^2}X^TX\right] = \frac{X^TX}{\sigma^2}$$

Unit information takes the average over n observations so the variance is set to $n\sigma^2(X^TX)^{-1}$. This implies setting $\Omega_0 = n(X^TX)^{-1}$, so $(X^TX)^{-1}$ instead of I_p and with $g = n$.

It is also possible to let g be an unknown parameter and estimated by the data. This is known as the Zelner's g prior.

Bayesian Linear Regression model with unknown σ^2

Assume y is N($X\beta, \sigma^2 I_n$). Assign N($\mu_0, \sigma^2\Omega_0$) as prior for β (given σ^2) and the IGamma(α_0,β_0).

The posterior for $\pi(\beta, \sigma^2|y, X)$ is then the product of the IGamma(α_n, β_n) and the N($\mu_n, \sigma^2\Omega_n^2$) where

$$\begin{aligned}
\mu_n &= (X^T X + \Omega_0^{-1})^{-1}(\Omega_0^{-1}\mu_0 + X^T y) \\
\Omega_n &= (X^T X + \Omega_0^{-1})^{-1}, \\
\alpha_n &= \alpha_0 + \frac{n}{2}, \\
\beta_n &= \beta_0 + \frac{1}{2}(y^T y + \mu_0^T \Omega_0^{-1}\mu_0 + \mu_n^T \Omega_n^{-1}\mu_n).
\end{aligned}$$

The marginal posterior for β, $\pi(\beta|y, X)$ is the multivariate t distribution with $2\alpha_n$ degrees of freedom, location μ_n and scale $\frac{\beta_n}{\alpha_n}\Omega_n$

Marginal Posterior of β and Predictive distribution

To obtain credible intervals for β we could use the t distribution. Alternatively, Monte Carlo can be used; more general choice.

We draw N samples from $\pi(\beta|y)$ and use Monte Carlo to calculate credible intervals, density plots etc.

Samples from $\pi(\beta|y)$ can be drawn in the following way:

1. Generating samples σ_i^2 from IGamma(α_n, β_n), $i = 1, \ldots, N$,
2. Draw β_i sample based on each σ_i^2 from N$(\mu_n, \sigma_i^2 \Omega_n)$,

For the predictive distribution for a new observation $y*$ based on covariates X_* we can use the additional step of drawing y_{*i} from $N(X_* \beta_i, \sigma_i^2)$ for each (β_i, σ_i^2).

Bayesian Model Selection

To compare models we will need to compute the model evidence for each model.

We can then use the model with the highest marginal likelihood. The use of unit information prior is the default option to guard against the Jeffreys-Lindley paradox.

Computing the model evidence is generally a very difficult task but here we can use the following trick. We can write

$$\begin{aligned} \pi(\beta, \sigma^2|y, X) &= \frac{\pi(y|\beta, \sigma^2, X)\pi(\beta, \sigma^2)}{\pi(y|X)}, \text{ or else} \\ \pi(y|X) &= \frac{\pi(y|\beta, \sigma^2, X)\pi(\beta, \sigma^2)}{\pi(\beta, \sigma^2|y, X)}, \text{ for all } \beta, \sigma^2. \end{aligned}$$

The expression above contains known Normal and Inverse Gamma pdfs so we can just evaluate for -say- the posterior mean of β, σ^2.

Reading

Gelman et al:

Chapter 14