Ridge and LASSO regression

2018.04.02

杜岳華

Outline

- Polynomial regression
- Power of model
- Overfitting
- Ridge regression
- LASSO regression
- LASSO for feature selection

Recall

$Y = f(X)$

Linear regression: $f(X) = aX + b$
Loss function (error): $\sum_i (\hat{y_i} - y_i)^2$ $= \sum_i (f(x_i) - y_i)^2$

Polynomial regression

$Y = f(X)$

$ = w_0 + w_1X + w_2X^2 + \dots + w_kX^k$

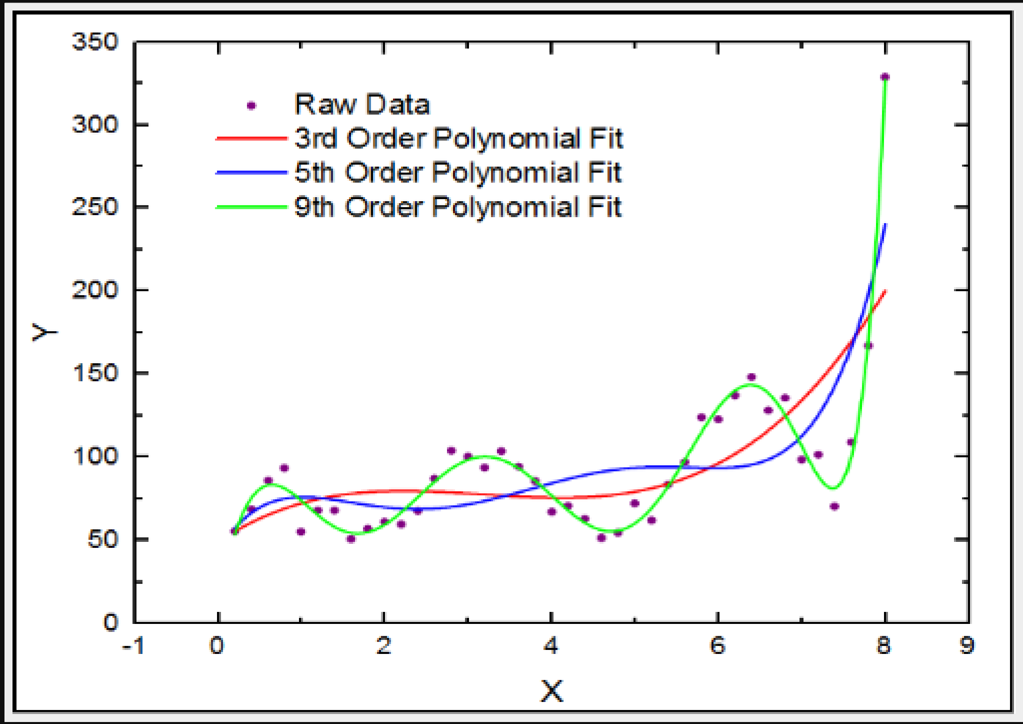
Loss function: $\mathcal{L} = \sum_i (f(x_i) - y_i)^2$
p.s. X is random variable，x_i is data

Power of model

1st order: $Y = w_0 + w_1X$

2nd order: $Y = w_0 + w_1X + w_2X^2$
3rd order: $Y = w_0 + w_1X + w_2X^2 + w_3X^3$
Power: 1st < 2nd < 3rd < ... < k-th

Use of over-powered model

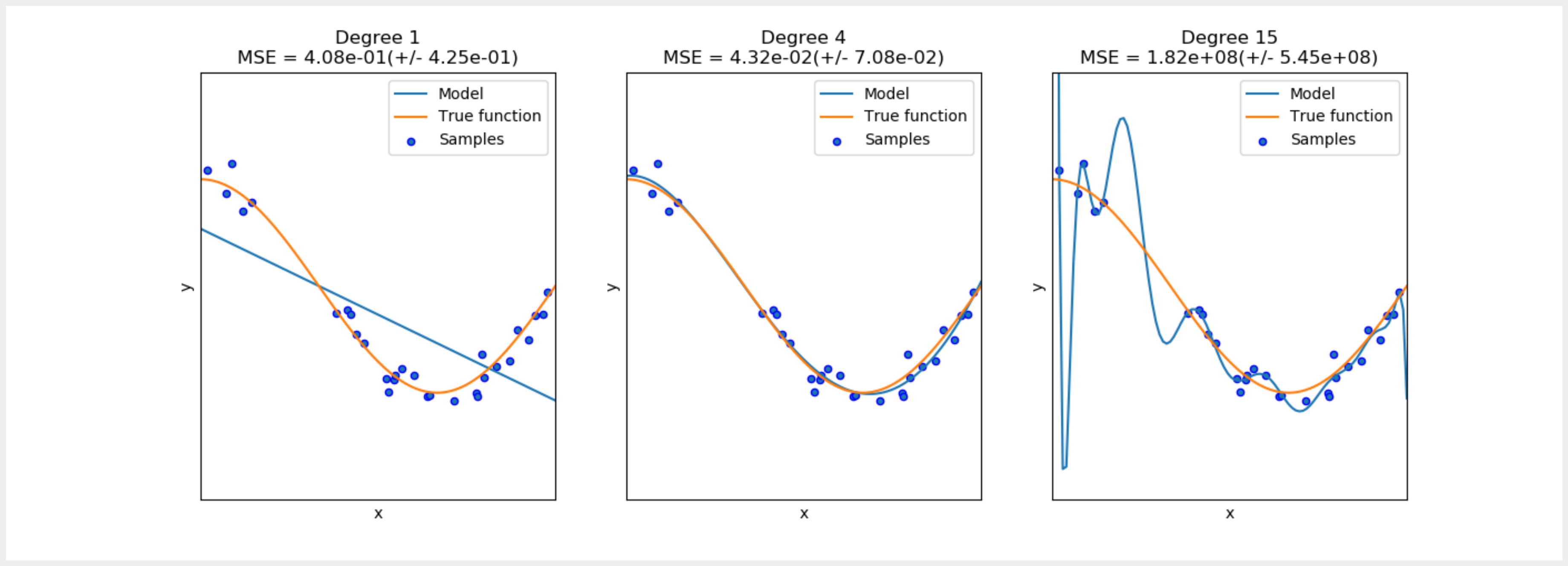


Use of over-powered model

那我們是不是拿全宇宙最強大的模型就可以 fit 任何的資料了？

事情沒有你想像的單純！

Overfitting



那怎麼辦？

Regularization

Add penalty!

Loss function: $\mathcal{L} = \sum_i (f(x_i) - y_i)^2 + {\Large \Omega}$

Regularization

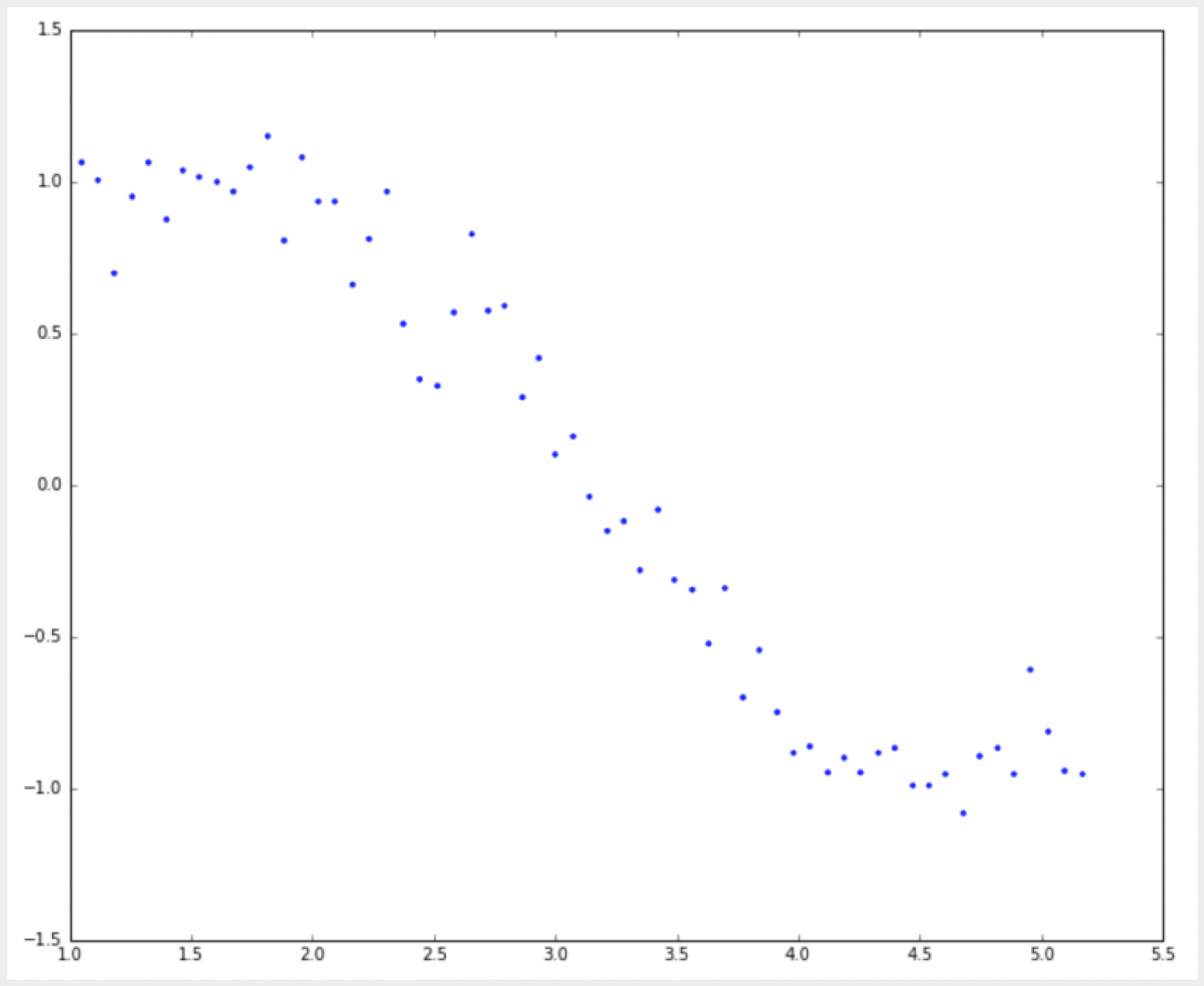
$\mathcal{L} = \sum_i (w_0 + w_1x_i + \dots + w_kx_i^k - y_i)^2 + {\Large \Omega}$

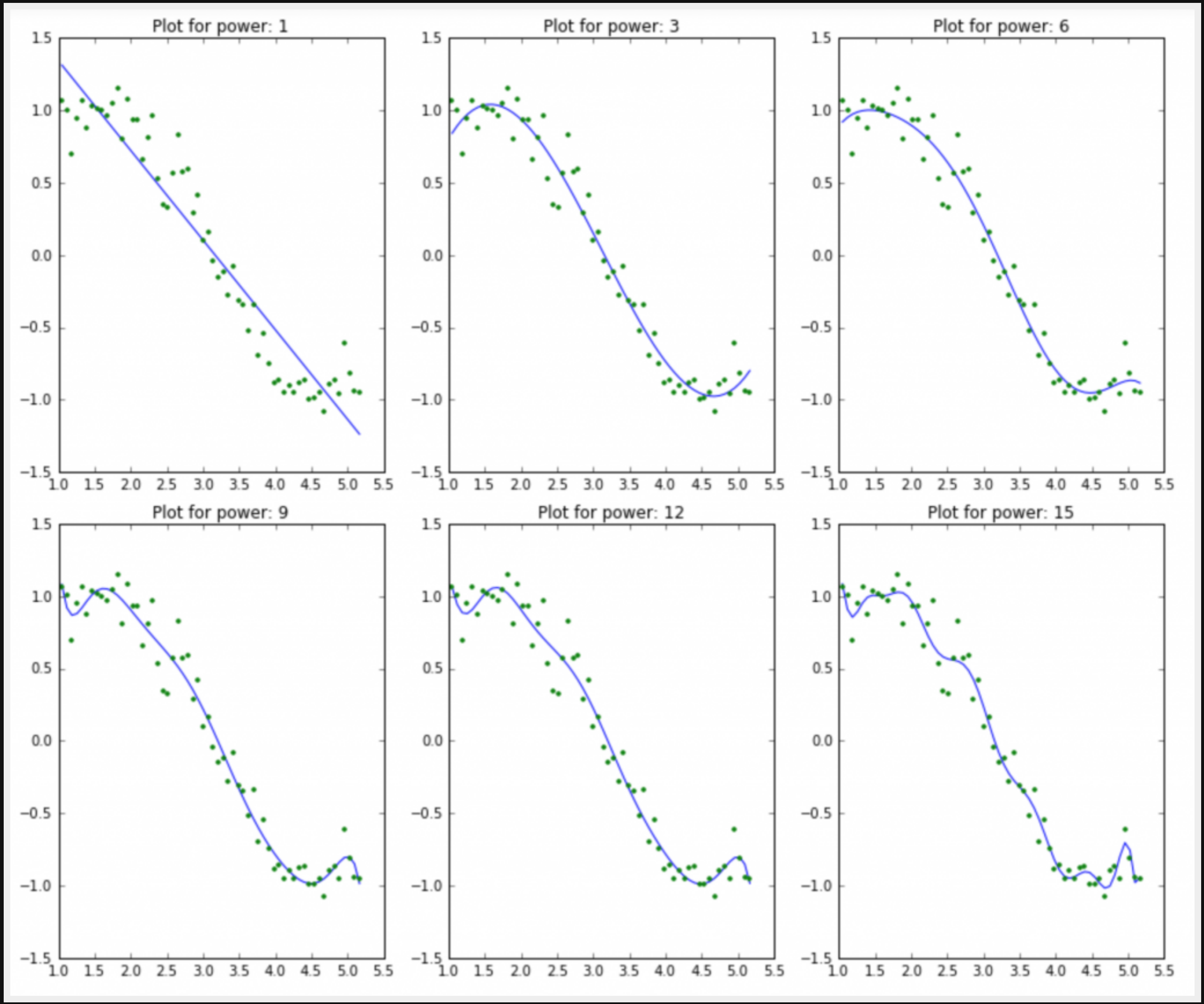
$\mathcal{L} = \sum_i (\sum_j w_jx_i^j - y_i)^2 + {\Large \lambda\sum_j (w_j)^2}$

Ridge regression

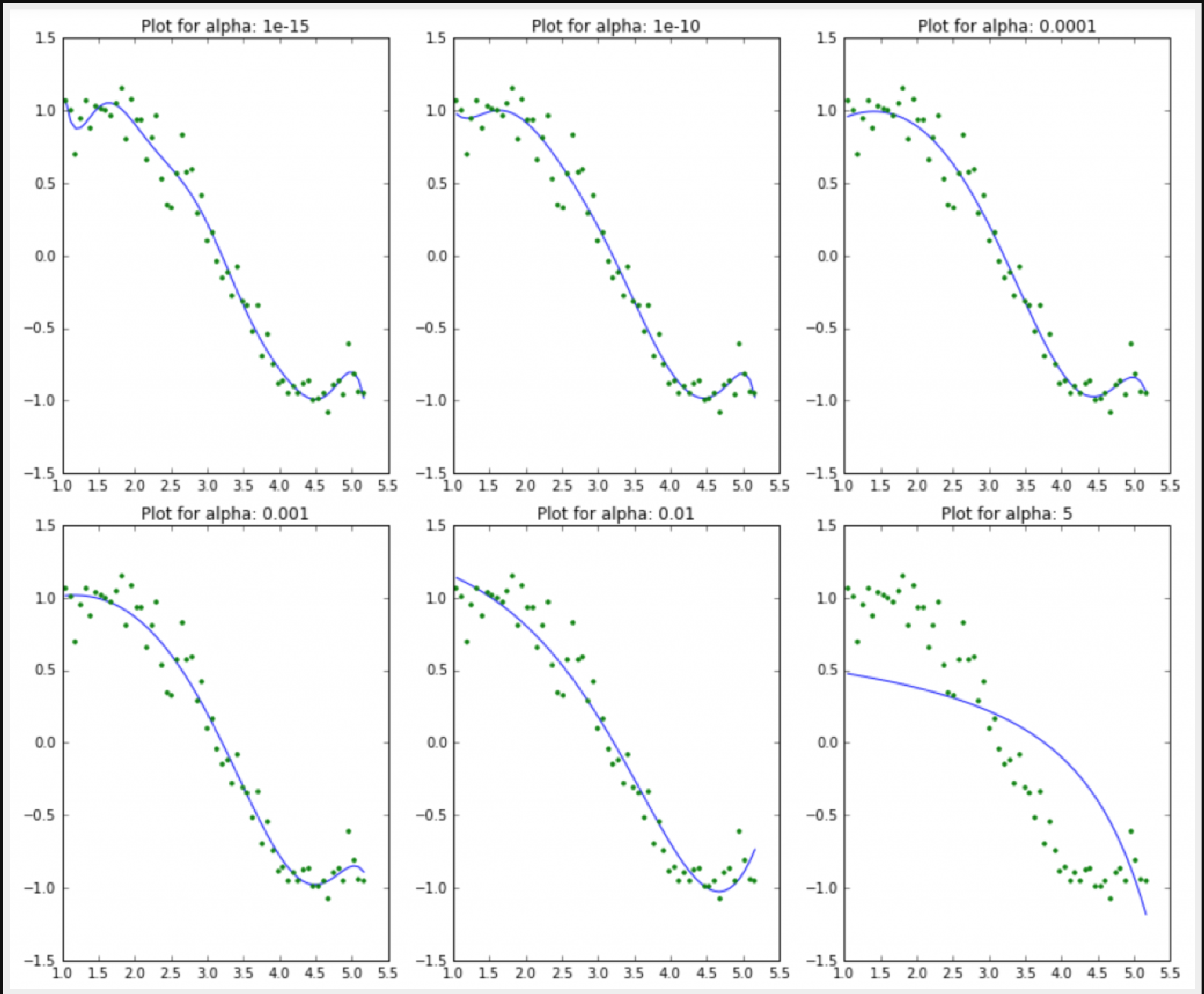
$\mathcal{L} = \sum_i (\sum_j w_jx_i^j - y_i)^2 + \lambda\sum_k (w_k)^2$

$\mathcal{l}^2$ norm: $\sum_k (w_k)^2$
Purpose: reduce the complexity of model


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5



Plot for power: 1
Plot for power: 3
Plot for power: 6
Plot for power: 9
Plot for power: 12
Plot for power: 15
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

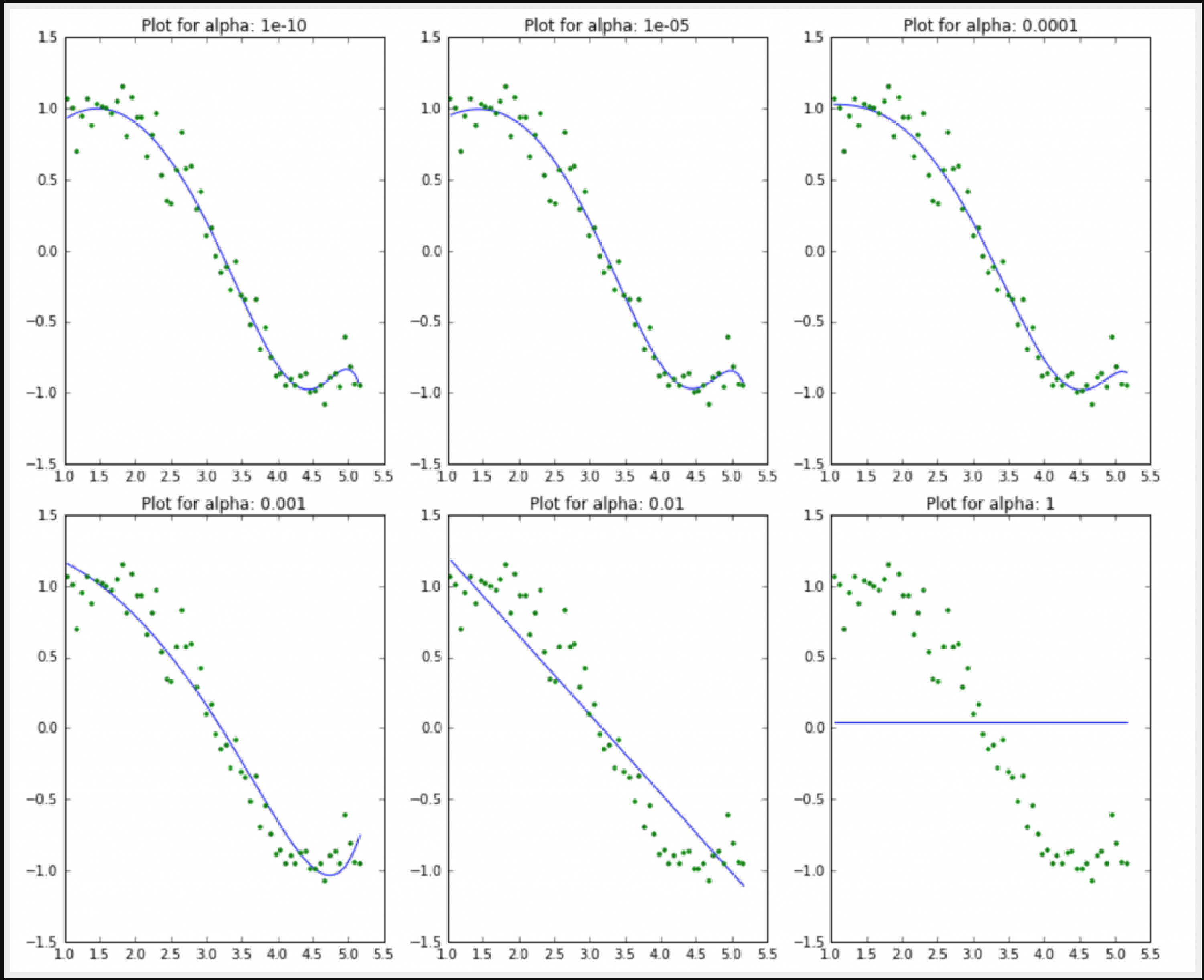

Plot for alpha: 1e-15
Plot for alpha: 1e-10
Plot for alpha: 0.0001
Plot for alpha: 0.001
Plot for alpha: 0.01
Plot for alpha: 5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5

LASSO regression

Least Absolute Shrinkage and Selection Operator

$\mathcal{L} = \sum_i (\sum_j w_jx_i^j - y_i)^2 + \lambda\sum_k |w_k|$

$\mathcal{l}^1$ norm: $\sum_k |w_k|$
Purpose: reduce the effect of uncorrelated terms


Plot for alpha: 1e-10
Plot for alpha: 1e-05
Plot for alpha: 0.0001
Plot for alpha: 0.001
Plot for alpha: 0.01
Plot for alpha: 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5

LASSO on multiple regression

$Y = f(X_1, X_2, \dots, X_k)$

$ = w_0 + w_1X_1 + w_2X_2 + \dots + w_kX_k$

$ = 0.001 + 7.5X_1 + 0.35X_2 + \dots + 7.2X_k$
$ = \ \ + 7.5X_1 + \ \ + \dots + 7.2X_k$

LASSO for feature selection

Y	X_1	X_2	X_3
$\dots$	$\dots$	$\dots$	$\dots$
$\checkmark$	$\checkmark$	$\times$	$\times$

補充

$\mathcal{l}^2 norm: \sum_k (w_k)^2 = \lVert \mathbf{w} \rVert_2$

$\mathcal{l}^1 norm: \sum_k |w_k| = \lVert \mathbf{w} \rVert_1$

$\mathcal{l}^p norm: \sum_k |w_k|^p = \lVert \mathbf{w} \rVert_p$

Q&A