Linear Regression

$$P(y|x,\theta) = \mathcal{N}(y|w^T x, \sigma^2)$$

multivariate linear regression

Multiple data points

$$y_i = w^T x_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

$$y_i = \sum_a w_a x_{ia} + \varepsilon_i$$

Sometimes we have nonlinear relation expressed through a basis expansion

$$P(y|x,\theta) = \mathcal{N}(y|w^T \phi(x), \sigma^2)$$

- For example $\phi(x) = [1 \; x \; \cdots \; x^d]$

The structure of the problems ~~is~~ are the same

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1D} \\ x_{21} & \cdots & x_{2D} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{ND} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^d \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

If ε_i's are independent,

$$\log \prod_i p(y_i | w) = \log \frac{e^{-\sum (y_i - w^T x_i)^2 / 2\sigma^2}}{(\sqrt{2\pi\sigma^2})^n}$$

$$= -\frac{1}{2\sigma^2} \sum_i (y_i - w^T x_i)^2 - \frac{N}{2} \log 2\pi\sigma^2$$

Max over $w \Rightarrow \min \sum_i (y_i - w^T x_i)^2$

Least Square fit.

$$RSS(w) = \sum_i (y_i - w^T x_i)^2$$

Residual sum of squares

$$RSS(w) = \|\varepsilon\|^2$$

In general we will write

$$y = Xw + \varepsilon$$

$$\begin{pmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{ND} \end{pmatrix} \text{ or } \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ \vdots & & & & \\ 1 & x_N & x_N^2 & \cdots & x_N^d \end{pmatrix} \text{ or } \cdots \begin{pmatrix} ?? \end{pmatrix}$$

X is called the design matrix.

$$P(y|w) \propto e^{-(y-Xw)^T(y-Xw)/2\sigma^2}$$

So, the object to minimized is negative log likelihood (NLL) upto some linear transformation.

$$NLL(w) = \frac{1}{2}(y-Xw)^T(y-wx)$$

$$= \frac{1}{2} w^T X^T X w - w^T X^T y + \frac{1}{2} y^2$$

Now take w derivative and set to zero

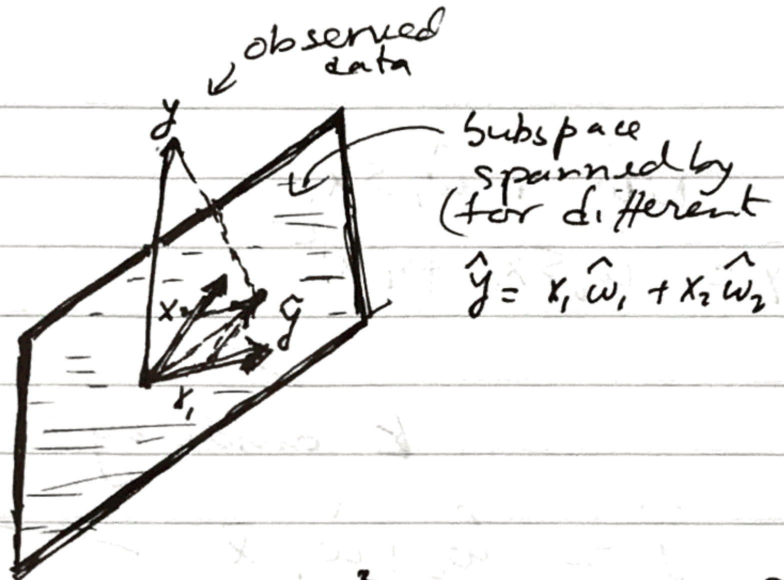
$$X^T X w = X^T Y$$

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T Y$$

ordinary least square

$$y = X_1 w_1 + X_2 w_2 + \cdots + \varepsilon$$

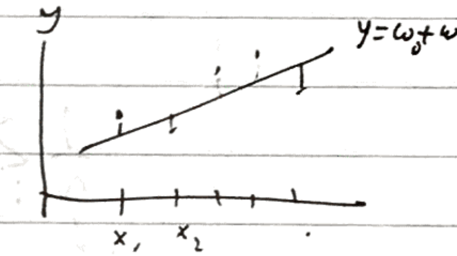
columns of the design matrix



$$\hat{y} = x_1 \hat{\omega}_1 + x_2 \hat{\omega}_2$$

$(y - \hat{y})^2$ is the RSS

Let us take the matrix formulation out for a test drive on the familiar road of linear regression



$$Y_i = \omega_0 + \omega_1 x_i + \varepsilon_i$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & \cdots & x_1 \\ \vdots & & \vdots \\ 1 & \cdots & x_N \end{pmatrix}}_{X} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$X^T X = \begin{pmatrix} N & \sum x_i \\ \sum_i x_i & \sum x_i^2 \end{pmatrix} \qquad X^T Y = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}$$

$$X^T X \hat{\omega} = X^T Y \Rightarrow \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$\sum y_i = N\hat{w}_0 + \hat{w}_1 \sum x_i$$

$$\sum x_i y_i = \hat{w}_0 \sum x_i + \hat{w}_1 \sum x_i^2$$

divide by N

$$\bar{y} = \hat{w}_0 + \hat{w}_1 \bar{x}$$

The line passes through $(\bar{x}, \bar{y})$

multiply by $\bar{x}$ and subtract

divide by N

$$\frac{1}{N} \sum x_i y_i = \hat{w}_0 \bar{x} + \hat{w}_1 \frac{1}{N} \sum_i x_i^2$$

$$\frac{1}{N} \sum_i x_i y_i - \bar{x}\bar{y} =$$ [redacted] $$\hat{w}_1 \left(\frac{1}{N} \sum_i x_i^2 - \bar{x}^2 \right)$$

Empirical covariance [redacted]

[redacted] empirical variance of x

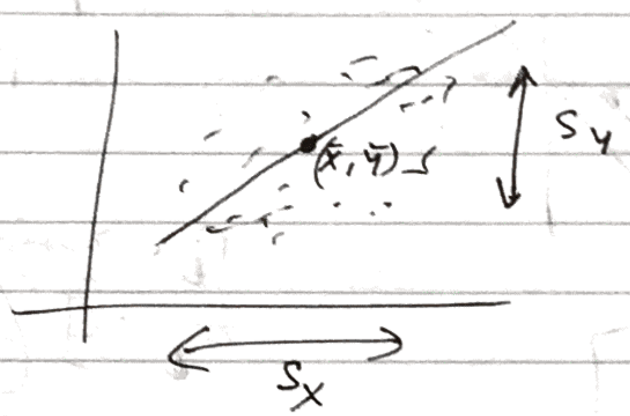
$$\left[\frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \right] = \hat{w}_1 \left[\frac{1}{N} \sum_i (x_i - \bar{x})^2 \right]$$

Correlation coefficient

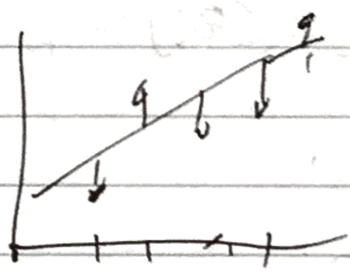
$$r_{x,y} = \frac{\frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y})}{\underbrace{\sqrt{\frac{1}{N} \sum_i (x_i - \bar{x})^2}}_{S_x} \underbrace{\sqrt{\frac{1}{N} \sum_i (y_i - \bar{y})^2}}_{S_y}}$$

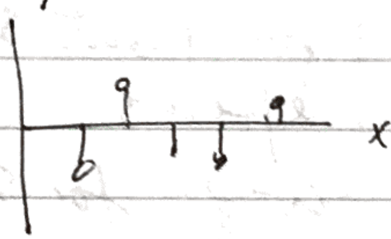
$$r_{x,y} \, s_x s_y = \hat{\omega}_1 s_x^2$$

$$\hat{\omega}_1 = \frac{r_{xy} \, s_y}{s_x}$$

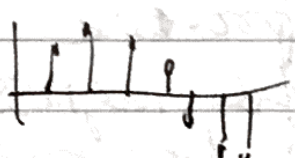


It is a good practice to check what the residuals

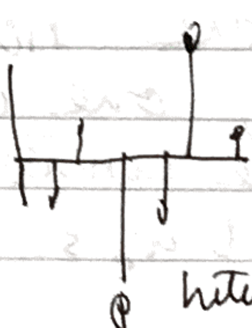




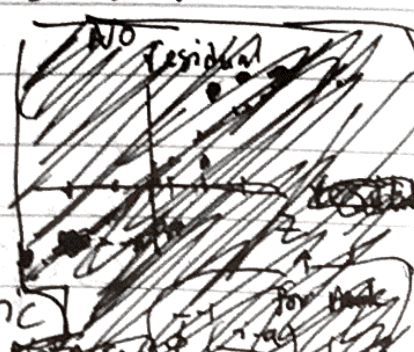
and check whether the residuals look like independent, homoskedastic and normal



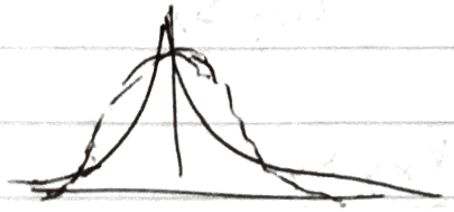
Not uncorrelated

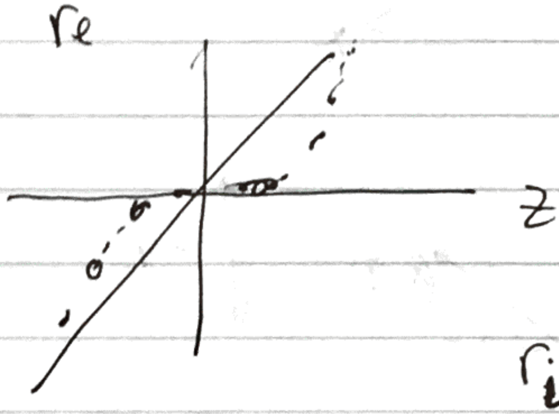


heteroskedastic



Checking normality
Imagine the residual distr^n





normal probability plot

r_i residual

$$z_i = \Phi^{-1}\left(\frac{i-a}{N+1-2a}\right)$$

$\boxed{\text{Matlab function normplot}}$

approx $\frac{i}{N}$ $\quad \Phi^{-1}\left(\frac{i}{N}\right)$ gives the z value with $P(z \leq z_i) = \frac{i}{N}$

a corrects for finite size effects

$a = \frac{3}{8}$ for small N $(N \leq 10)$

$a = .5$ for large N $(N > 10)$

When the normal error model is right, one could provide confidence limits for the regression coefficients w_a's as well for predicted values of y. One could also do hypothesis tests.

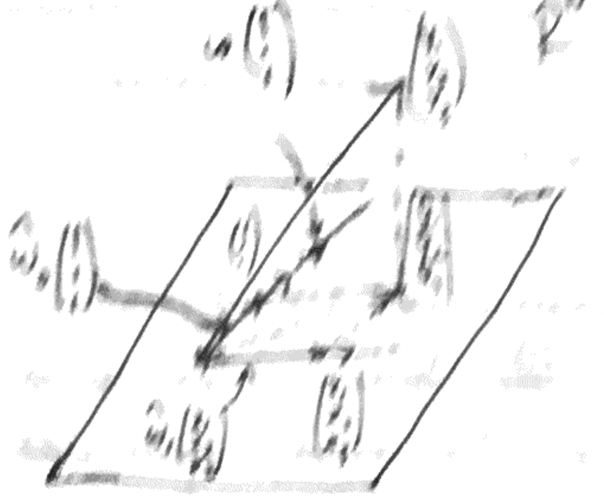
We will discuss an example that it related to the issue of complexity, but is often posed as a test for explanatory relevance of feature.

Consider these nested models

$$y_i = w_0 + \varepsilon_i$$

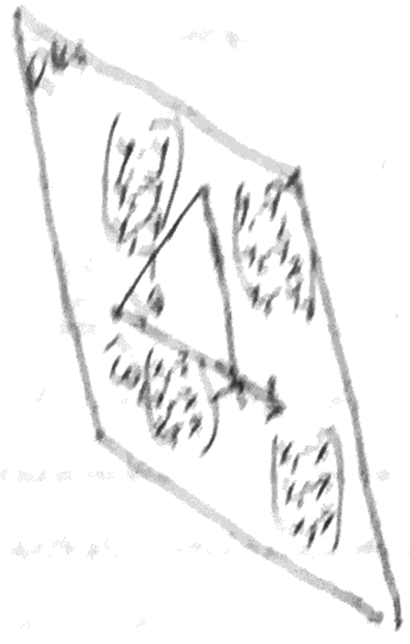
$$y_i = w_0 + w_1 x_i + \varepsilon_i \qquad w_1 = 0$$

If we do two parameter least square, we will usually find a nonzero $\hat{w}_1$. One could ask whether one should just stick to the simpler model with $w_1 = 0$.



How do I test anything about $\hat{\omega}_1$

Project to the subspace perpendicular to $\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$



$$\cot\theta = \frac{\hat{\omega}_1 \sqrt{\sum_i (x_i-\bar{x})^2}}{\sqrt{\sum_i (y_i-\hat{y}_i)^2}}$$

$$t = \frac{\hat{\omega}_1 \sqrt{\sum_i (x_i-\bar{x})^2}}{\sqrt{\frac{\sum_i (y_i-\hat{y}_i)^2}{N-2}}} = \frac{\hat{\omega}_1 \sqrt{\frac{\sum_i (x_i-\bar{x})^2}{N}}}{\sqrt{\frac{\sum_i (y_i-\hat{y}_i)^2}{N}}} \sqrt{N-2}$$

$$= \frac{\hat{\omega}_1 s_x}{s_{y.x}} \sqrt{N-2} \qquad \text{d.f. } N-2$$

Note that $\sum_i (y_i-\bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$$= \sum_i (y_i-\hat{y}_i)^2 + 2\sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) + \sum_i (\hat{y}_i-\bar{y})^2$$

$$\hat{y}_i - \bar{y} = \hat{\omega}_0 + \hat{\omega}_1 x_i - (\hat{\omega}_0 + \hat{\omega}_1 \bar{x})$$
$$= \hat{\omega}_1 (x_i - \bar{x})$$

$$\sum_i (y_i - \hat{y}_i)(x_i - \bar{x}) = 0$$

So Total variance [redacted] Unexplained var Explained variance

[redacted] $\sum_i (y_i-\bar{y})^2 = \sum_i (y_i-\hat{y}_i)^2 + \sum_i (\hat{y}_i-\bar{y})^2$

[redacted] RSS_0 [redacted] $= (RSS_1$ [redacted] $+ ESS$ $\quad \hat{\omega}_1^2 \sum_i (x_i-\bar{x})^2$

So $t^2 = \dfrac{\sum_i (\hat{y}_i-\bar{y})^2/1}{\sum_i (y_i-\hat{y}_i)/(N-2)}$ = F statistic with dfs 1, N-2

In general, if we went from a p_0 parameters to p_1 parameters and [redacted] RSS_0 [redacted] reduced to [redacted] RSS_1 [redacted], then

$F =$ [redacted] $\dfrac{(RSS_0 - RSS_1)/(p_1-p_0)}{RSS_1/(N-p_1-1)}$

would be distributed as $F_{p_1-p, N-p_1-1}$ if the smaller model is sufficient and the larger model is only fitting noise.

For polynomial fits, there is a natural structure: ~~[illegible]~~ degree of the polynomial. Many multivariate linear regression problems, it is not obvious what to fit first and what next.

Many different methods has been proposed how to choose a subset of variables to fit. In general they run into the trouble that there are many such subsets.

Stepwise Forward method chooses the variable that explains the most variance significantly. It then considers adding one more etc et. They all have essentially some multitesting issues

Alternative → Shrinkage based methods.

Ridge regression

Prior $p(w) = \prod_j \mathcal{N}(w_j | 0, \tau^2)$

Now the MAP estimation of w

$$\underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + w^T x_i, \sigma^2)$$

$$+ \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$$

Equivalent to minimizing

$$J(w) = \sum_{i=1}^{N} (y_i - (w_0 + w^T x_i))^2 + \lambda \|w\|_2^2$$

where $\lambda = \frac{\sigma^2}{\tau^2}$ and $\|w\|_2^2 = \sum_{j=1}^{D} w_j^2$

$= w^T w$

The parameter w_0 is there to handle $\bar{y}$ and is not regularized

Ridge regression or penalized least squares

The intuition is that w_i would be small if it does not improve the square error significantly

w_0 derivative set to zero

$$\sum_i \left[y_i - (w_0 + w^T x_i) \right] = 0$$

$$w_0 = \bar{y} - w^T \bar{x}$$

Let us center both y_i and x_i

$y_i - \bar{y} \rightarrow y_i$ $\qquad$ $x_i - \bar{x} \rightarrow x_i$

$$\sum_i (y_i - w^T x_i)^2 + \lambda w^T w$$

$$= (y - Xw)^2 + w^T w$$

w deriv

$$2X^T(X\hat{w} - y) + 2\lambda \hat{w} = 0$$

$$(\lambda I_D + X^T X)\hat{w} = X^T y$$

$$\hat{w}_{ridge} = (\lambda I_D + X^T X)^{-1} X^T y$$

Relationship to PCA, SVD etc.

A quick introduction to singular value decompositio

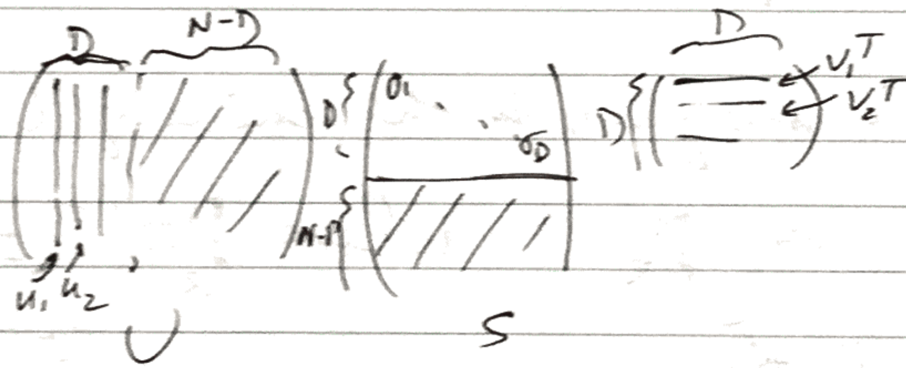
$$\underset{N\times D}{X} = \underset{N\times N}{U}\ \underset{N\times D}{S}\ \underset{D\times D}{V^T}$$

U, V unitary (orthogonal for real X) $\Rightarrow U^T U = UU^T = I_N$

$V^T V = VV^T = I$

$$S = \begin{array}{c} \\ D \\ N-D \end{array} \begin{array}{c} \overset{D}{\longleftrightarrow} \\ \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_D \\ \hline & 0 & \end{pmatrix} \end{array}$$

(We assume $N > D$)



$$X = \sum_{i=1}^{D} \sigma_i u_i v_i^T$$

You can think of it this way

The Gramians X^TX and XX^T are symmetric matrices. They can be diagonalized $X^TX = \tilde{V} \Lambda_D \tilde{V}^T$ and $XX^T = \tilde{U} \tilde{\Lambda}_N \tilde{U}^T$

Since $Tr(X^TX)^n = Tr(XX^T)^n$

$$\sum_{i=1}^{D} \lambda_i^n = \sum_{j=1}^{N} \tilde{\lambda}_j^n$$

Only way to satisfy this $\{\tilde{\lambda}_j\}_{j=1}^{N}$

$$= \{\lambda_1, \dots, \lambda_D, \underbrace{0, 0, \dots, 0}_{N-D}\}$$

$\lambda_i \geq 0$ since X^TX is positive semidefinite. Choose $\sigma_i = \sqrt{\lambda_i}$

Only ambiguity in the choice of U, V

is $V \Rightarrow \tilde{V} P_R$ where $P_R = \begin{pmatrix} \pm 1 & & \\ & \pm 1 & \\ & & \ddots \end{pmatrix}$

$U \Rightarrow \tilde{U} P_L$ $\quad P_L = \begin{pmatrix} \pm 1 & & \\ & \pm 1 & \\ & & \ddots \end{pmatrix}$

Consider the matrix $\tilde{V}$

and its rows. $\begin{pmatrix} \tilde{v}_1^T \\ \tilde{v}_2^T \end{pmatrix}$

and the matrix $\tilde{U}$ and its columns $(\tilde{u}_1, \tilde{u}_2 \dots)$

Note that $(X^TX)\tilde{v}_j = \lambda_j \tilde{v}_j$

Then $XX^T(X\tilde{v}_j) = \lambda_j X\tilde{v}_j$

But $XX^T\tilde{u}_j = \lambda_j \tilde{u}_j$

So $X\tilde{v}_j = \sigma_j \tilde{u}_j \Rightarrow \tilde{v}_j^T X^T X \tilde{v}_j = \gamma_j^2$

But then $\gamma_j^2 = \lambda_j \Rightarrow \gamma_j = \pm\sigma_j$

Using the ± 1 in P_L, P_R we could define u_j, v_j s.t.

$$X v_j = \sigma_j u_j \quad \text{etc..}$$

$$X = U S V^T$$

Now, for centered y, X in ridge regression

$$\hat{w}_{ridge} = (\lambda I_D + X^TX)^{-1} X^T y$$

using SVD representation of X:

$$\hat{w}_{ridge} = V(\lambda I_D + S^TS)^{-1} \underbrace{V^TV}_{I_D} S U^T y = V(S^2 + \lambda I)^{-1} S U^T y$$

Ridge regression predictions are

$$\hat{y} = X \hat{w}_{ridge} = USV^T V (S^2 + \lambda I)^{-1} S U^T y$$

$$= \sum_{j=1}^{D} u_j \tilde{S}_{jj} u_j^T y$$

where $\tilde{S}_{jj} = \frac{\sigma_j^2}{\sigma_j^2 + \lambda}$

So $$\hat{y} = \sum_{j=1}^{D} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y$$

The corresponding least square answer is

$$\hat{y} = \sum_{j=1}^{D} u_j u_j^T y$$

For directions u_j where $\sigma_j^2 \ll \lambda$ u_j would not have much of an effect on prediction in ridge regression as opposed to pure least square.

One might consider $\sum_i \frac{\sigma_i^2}{\sigma_i^2+\lambda}$ to be an effective degree of freedom here.

Why is this a good thing?

Well log likelihood $\sim -\frac{1}{2\sigma^2}(Xw-Y)^T(Xw-Y)$

$\Rightarrow$ Fisher Information $= \frac{1}{\sigma^2} X^T X$

$\Rightarrow$ $\mathrm{Cov}(\hat{w}) \sim \sigma^2 (X^T X)^{-1}$

That mean $\hat{w}$ is most ambiguous in direction which are given by eigenvectors of $X^T X$ corresponding to lowest eigenvalues $\lambda_i = \sigma_i^2$. The factor $\frac{\sigma_i^2}{\sigma_i^2+\lambda}$ suppresses these directions, preventing effect of noise from influencing $\hat{w}$ and subsequently the prediction.

Applied to the familiar territory of single variable linear regression,

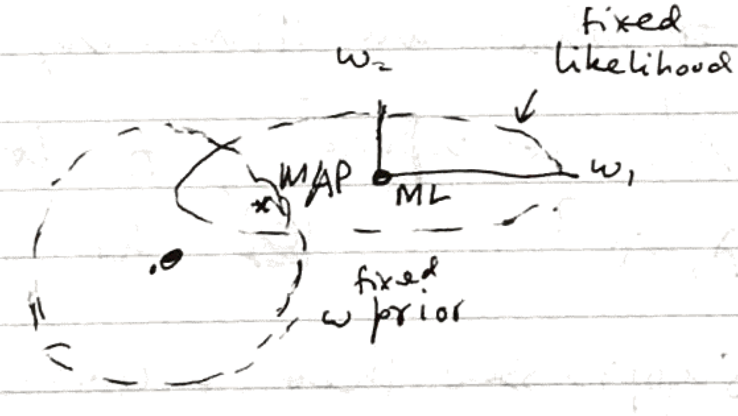
$$\underset{w_1}{\text{argmin}} \left[\sum_i \left\{ y_i - \bar{y} - w_1 (x_i - \bar{x}) \right\}^2 + \lambda w_1^2 \right]$$

$$\left(\hat{w}_1\right)_{\text{ridge}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 + \lambda} = \frac{r_{xy} s_x s_y}{s_x^2 + \lambda / N}$$

$$= \left(\hat{w}_1\right)_{\text{OLS}} \frac{N s_x^2}{N s_x^2 + \lambda}$$

$\hat{w}_1$ is 'shrunk'

Two variables

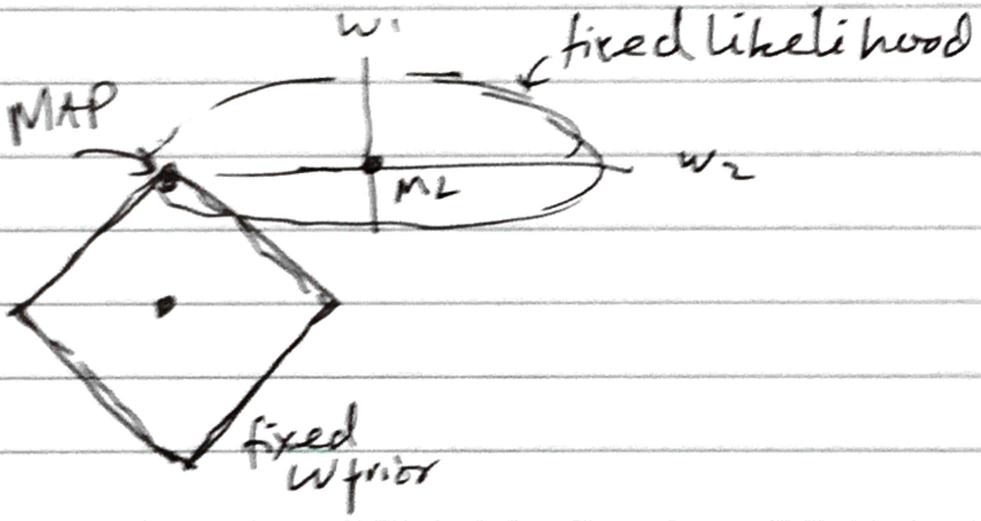


$\hat{w}_1$, which is more error-prone, is shrunk more than, $\hat{w}_2$.

If we really want some w_i's set to zero, one could use the laplace prior

l_1 norm

$$P(w) \propto e^{-\lambda \|w\|_1} = e^{-\lambda \sum_{j=1}^{D} |w_j|}$$



The l_1 penalty could lead to sparse w's: many $w_j = 0$.
$\Rightarrow$ LASSO (least absolute shrinkage and selection operator).

Both ridge regression & LASSO are convex optimization problems