Stats 102B - Week 5, Lecture 2

Miles Chen, PhD

Department of Statistics

Week 5 Wednesday

UCLA

Section 1

Maximum Likelihood Estimate of the Variance

Likelihood function of Linear Regression model with Gaussian Noise

$$L(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} P(t_n|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} N(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

We assume all observations of t_n are independent. We assume ϵ is normally distributed $N(0, \sigma^2)$, so the likelihood of our data is:

$$L(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

Maximizing the Likelihood function w.r.t $\hat{\mathbf{w}}$

When it's time to maximize the likelihood function w.r.t $\hat{\mathbf{w}}$, we maximize the log-likelihood, and we see that maximizing the log-likelihood function will be equivalent to minimizing the SSE.

$$\text{Likelihood} = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

$$\text{log-likelihood} = \sum_{n=1}^{N} \left(\frac{-1}{2}\log(2\pi) - \log(\sigma) + \frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

$$\text{log-likelihood} = \underbrace{\frac{-N}{2}\log(2\pi) - N\log(\sigma)}_{\text{constants}} + \underbrace{\frac{-1}{2\sigma^2}}_{\text{negative constant}} \underbrace{\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2}_{\text{SSE}}$$

Solution

The value of $\mathbf{w}$ that maximizes the log-likelihood is

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

which is the exact same solution that minimizes the SSE.

Minimizing the squared loss is equivalent to maximizing the likelihood if the errors are assumed to come from a Gaussian distribution.

Maximizing the Likelihood function w.r.t σ

We can maximize the likelihood function with respect to the standard deviation of the gaussian noise, σ. We plug in the MLE estimate of $\hat{\mathbf{w}}$ and treat it as a fixed value (we saw that is has no dependence on σ).

$$\log \mathsf{L} = \frac{-N}{2}\log(2\pi) - N\log(\sigma) + \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$$

$$\frac{\partial \log \mathsf{L}}{\partial \sigma} = 0 - \frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$$

$$0 = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$$

$$\frac{N}{\sigma} = \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$$

The MLE estimate of our variance turns out to be the mean squared error of our model.

The log-likelihood at the maximum:

We can plug in the estimate of $\widehat{\sigma^2}$ into the log-likelihood of our model to see the log-likelihood at the maximum.

$$
\begin{aligned}
\text{log-likelihood} &= \frac{-N}{2}\log(2\pi) - N\log(\sigma) + \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2 \\
&= \frac{-N}{2}\log(2\pi) - \frac{N}{2}\log(\widehat{\sigma^2}) + \frac{-1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\
&= \frac{-N}{2}\log(2\pi) - \frac{N}{2}\log(\widehat{\sigma^2}) + \frac{-N}{2} \\
&= \frac{-N}{2}(1 + \log(2\pi)) - \frac{N}{2}\log(\widehat{\sigma^2})
\end{aligned}
$$

Maximum Likelihood favors complex models

$$\text{log-likelihood} = \frac{-N}{2}(1 + \log(2\pi)) - \frac{N}{2}\log(\widehat{\sigma^2})$$

Because there's a negative constant in front of $\widehat{\sigma^2}$, the log-likelihood will continue to increase if we decrease $\widehat{\sigma^2}$.

σ^2 is the variance of the 'error' in the model. The error is the left-over variation after our model makes a prediction. One way to decrease the amount of error is to make a prediction function $f(\mathbf{x}; \mathbf{w})$ as complicated as possible so that the remaining error is as small as possible.

Increasing the complexity of a model can maximize the likelihood (and reduce the error), but intuitively, we know that a highly complex model may not always be the best.

Section 2

Fisher Information

Review: Derivatives of univariate functions

The first derivative of a function tells us the rate of change of the function. It is the slope of the line tangent to the function.

The second derivative of a function tells us how quickly slope changes - it tells us about the curvature of the function. A large absolute value of the second derivative indicates that the slope rapidly changes - indicating a sharp 'bend' in the curve.

The sign of the second derivative tells us about the concavity of the function.

If we know the first derivative is equal to 0 at some point x, and the second derivative is ...

- positive ($f''(x) > 0$) then the function is concave up at x, and x is a local minimum
- negative ($f''(x) < 0$) then the funciton is concave down at x, and x is a local max
- zero ($f''(x) = 0$) then we do not know the concavity of the function at x. x could be an inflection point.

Derivatives of multivariate functions: The Gradient

We can generalize some of the intuition of derivatives to higher dimensional functions. Let f be a function of a vector that produces a scalar ($f : \mathbb{R}^n \to \mathbb{R}$).

The input to the function is a vector: $\mathbf{w} = (w_1, w_2, ...w_n)$

The gradient of the function with respect to $\mathbf{w}$ at $\mathbf{w}$ is itself a vector:

$$\nabla f(\mathbf{w}) = \frac{\partial f}{\partial \mathbf{w}} = \left(\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \cdots, \frac{\partial f}{\partial w_n} \right)$$

We can think of this as the multivariate generalization of a derivative.

Second derivatives of multivariate functions: The Hessian

If all of the second partial derivatives of f are continous over the domain, then the Hessian (pronounced Hesh'en) matrix $\mathbf{H}$ of f is an $n \times n$ matrix:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_n} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_n \partial w_1} & \frac{\partial^2 f}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_n^2} \end{bmatrix}$$

This is a matrix of second derivatives, and can be thought of as the multivariate generalization of the second derivative.

The Hessian

Another way to think of the Hessian is to form the columns of the matrix by taking the 'gradient' of each component of the gradient.

$$\nabla f(\mathbf{w}) = \frac{\partial f}{\partial \mathbf{w}} = \left(\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \cdots, \frac{\partial f}{\partial w_n} \right)$$

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_n} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_n \partial w_1} & \frac{\partial^2 f}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_n^2} \end{bmatrix}$$

Positive Definite / Negative Definite Matrices

For a univariate function, the sign of the second derivative tells us about the concavity of the function.

For a multivariate function, the "definiteness" of a matrix can tell us about the curvature of the function.

If the gradient of a function is equal to 0 at some point $\mathbf{x}$, and the Hessian matrix ...

- is positive-definite (all eigenvalues of the matrix are > 0) then the function is "concave up" at $\mathbf{x}$, and $\mathbf{x}$ is a local minimum
- is negative-definite (all eigenvalues of the matrix are < 0) then the function is "concave down" at $\mathbf{x}$, and $\mathbf{x}$ is a local maximum
- has both positive and negative eigenvalues then the function has a saddle point at $\mathbf{x}$. It is concave up in one direction and concave down in another direction.
- has eigenvalues equal to 0, then we need more information to determine concavity.

More properties of the Hessian Matrix

See: https://en.wikipedia.org/wiki/Hessian_matrix

The Hessian Matrix is symmetric (usually). Schwarz's theorem: If second derivatives are continuous, the order of differentiation does not matter.
(https://en.wikipedia.org/wiki/Symmetry_of_second_derivatives#Schwarz's_theorem)

The Hessian Matrix is equivalent to the Jacobian matrix of the gradient of the function.

The Hessian as a measure of curvature

The Hessian is also useful because it can inform us about the curvature of the function.

When we apply it to the likelihood or log-likelihood function, we can get an idea of the curvature of the function, which tells us how **confident we can be in our parameter estimates**.

Example:

In the following example, we will demonstrate the the second derivative of a univariate probability function tells us how confident we can be in our parameter estimates.

After this example, we will try to generalize it to the multivariate case.

Univariate likelihood

Let's imagine (once again) that we have a pouch with some red and blue marbles in it. The proportion that is red θ is a fixed, but unknown value.

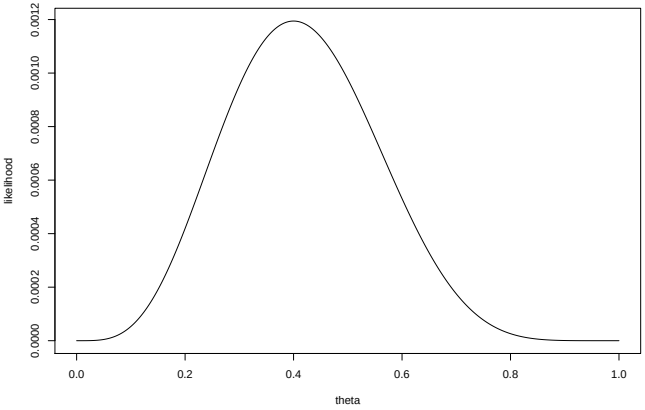
Let's say we draw 10 marbles with replacement. We observe red 4 times. Our likelihood function will be:

$$L(\theta|\text{data: 4 red out of 10 draw}) = \theta^4(1-\theta)^6$$

When we graph this, we get:

```
theta <- seq(0,1,by = 0.001)
likelihood = theta^4 * (1-theta)^6
plot(theta, likelihood, type = 'l')
```

10 draws, 4 Red



10 draws, 4 Red

The likelihood function shows us that 0.4 is the value of theta with the maximum likelihood. This makes intuitive sense.

Also note that the likelihood function also seems to indicate that 0.5 or 0.3 have high likelihoods as well.

While 0.4 is the most likely value, other values still have high likelihoods.

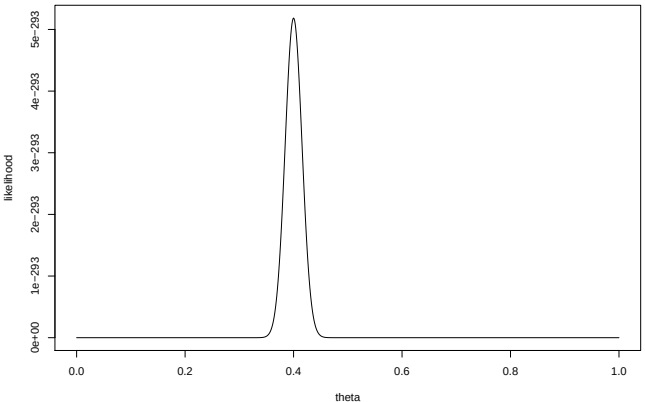
Same thing, more data

Let's say we draw 1000 marbles with replacement. We observe red 400 times. Our likelihood function will be:

$$L(\theta|\text{data: 400 red out of 1000 draw}) = \theta^{400}(1-\theta)^{600}$$

```
theta <- seq(0,1,by = 0.001)
likelihood = theta^400 * (1-theta)^600
plot(theta, likelihood, type = 'l')
```

1000 draws, 400 red



1000 draws, 400 red

Like the earlier graph, this graph also has a maximum likelihood at $\theta = 0.4$.

In contrast, this graph is much more pointy at 0.4.

This pointyness tells us that only values that are close to 0.4, like 0.39 or 0.41 still have high likelihood. There is almost 0 likelihood that θ is 0.3 or 0.5.

There is much less variance in what θ can be.

Relating it back to the second derivative

Let's look at the second derivative of the log-likelihood of these functions:

$$\log L_{\text{10 draws}} = \log(\theta^4(1-\theta)^6) = 4\log\theta + 6\log(1-\theta)$$
$$\frac{\partial \log L}{\partial \theta} = \frac{4}{\theta} + \frac{6}{1-\theta}$$
$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{-4}{\theta^2} + \frac{-6}{(1-\theta)^2}$$

$$\log L_{\text{1000 draws}} = \log(\theta^{400}(1-\theta)^{600}) = 400\log\theta + 600\log(1-\theta)$$
$$\frac{\partial \log L}{\partial \theta} = \frac{400}{\theta} + \frac{600}{1-\theta}$$
$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{-400}{\theta^2} + \frac{-600}{(1-\theta)^2}$$

The second derivative of the log-likelihood

When we evaluate the second derivative of the log-likelihood at the point $\theta = 0.4$, we get:

$$\frac{\partial^2 \log L_{\text{10 draws}}}{\partial \theta^2} = \frac{-4}{0.4^2} + \frac{-6}{(1-0.4)^2} = -41.667$$

$$\frac{\partial^2 \log L_{\text{1000 draws}}}{\partial \theta^2} = \frac{-400}{0.4^2} + \frac{-600}{(1-0.4)^2} = -4166.7$$

The second derivative measures the curvature of the function. This tells us what we already know: the likelihood function of the sample of 1000 has much more curvature (is more pointy).

Curvature as a measure of information

The fact that the likelihood function has much more curvature (the slope changes rapidly) in the second graph means that the data we have provides more information about the location of θ. This makes sense, with 1000 draws we have 100 times more information about what theta should be than we did after 10 draws.

The relationship between the curvature of the likelihood function and the information we have about the parameter is tied together with the Fisher Information.

Fisher Information

Fisher Information is a way for us to measure the information a random variable can provide about some unknown parameter. Formally, it is the variance of the score. (The score is the derivative/gradient of the log-likelihood.)

Let $f(X;\theta)$ be a probability density (or mass) function of X given some parameter θ.

If the log-likelihood is twice differentiable with respect to θ, the Fisher information is also equal to the **negative of the expected value of the second derivative of the log-likelihood**. That is:

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right], \text{can also be written as } \mathcal{I}(\theta) = -\mathbb{E}_{f(X;\theta)}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

The expected value is taken with respect to the random variable whose distribution is defined by $f(X;\theta)$.

Fisher Information for our Bernouli experiment

Let X be a random Bernouli variable. For a single Bernouli observation, the Fisher information is:

$$
\begin{aligned}
\mathcal{I}(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log(\theta^X(1-\theta)^{1-X})\right] \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log(\theta^X) + \frac{\partial^2}{\partial\theta^2}\log(1-\theta)^{1-X}\right] \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}X\log(\theta) + \frac{\partial^2}{\partial\theta^2}(1-X)\log(1-\theta)\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial\theta}\frac{X}{\theta} + \frac{\partial}{\partial\theta}\frac{-(1-X)}{(1-\theta)}\right] \\
&= -\mathbb{E}\left[\frac{-X}{\theta^2} + \frac{-(1-X)}{(1-\theta)^2}\right] \\
\mathcal{I}(\theta) &= \frac{\mathbb{E}(X)}{\theta^2} + \frac{\mathbb{E}(1-X)}{(1-\theta)^2}
\end{aligned}
$$

Fisher Information for our Bernouli experiment

The expected value of a Bernouli random variable X is equal to θ.

$$\begin{aligned}
\mathcal{I}(\theta) &= \frac{\mathbb{E}(X)}{\theta^2} + \frac{\mathbb{E}(1-X)}{(1-\theta)^2} \\
&= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\
&= \frac{1}{\theta} + \frac{1}{(1-\theta)} \\
&= \frac{(1-\theta)+\theta}{\theta(1-\theta)} \\
\mathcal{I}(\theta) &= \frac{1}{\theta(1-\theta)}
\end{aligned}$$

Fisher Information for our Bernouli experiment

Fisher Information is additive, so with n observations, we have:

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}$$

Which aligns with what we observed.

Back to the multi-variate case

The previous example demonstrated that the second derivative of the likelihood function told us how confident we could be in our parameter estimates.

We generalize this to a multivariate setting.

The second derivative in a multivariate setting is the Hessian matrix.

Back to the multi-variate case

The Fisher Information matrix for a multivariate function tells us how much information our data provides about the unknown parameters θ.

(If the log-likelihood is twice differentiable), the Fisher information is the negative expected value of the matrix of second derivatives (the Hessian) of the log-likelihood.

$$[\mathcal{I}(\theta)]_{i,j} = -\mathbb{E}_{f(X;\theta)}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(X;\theta)\right]$$

Fisher Information for Linear Regression

For linear regression, the unknown parameters are the coefficients in $\mathbf{w}$.

The outcome variable is $\mathbf{t}$. We assume the errors are normally distributed. The probability function of each observation t_n is the normal PDF with mean $\mathbf{w}^T\mathbf{x}_n$ and variance σ^2

$$p(t_n|\mathbf{w}, \mathbf{X}, \sigma^2) = N(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

In the likelihood function, $\mathbf{x}_n$ and σ^2 are fixed values.

The Fisher information is the negative expected value of the Hessian of the log-likelihood. The expected value is taken with respect to the outcome variable $\mathbf{t}$ whose distribution is defined by the probability function $p(t_n|\mathbf{w}, \mathbf{X}, \sigma^2)$

$$\mathcal{I}(\mathbf{w}) = -\mathbb{E}_{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)}\left[\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}^T} \log \prod_n N(\mathbf{w}^T\mathbf{x}_n, \sigma^2)\right]$$

Log-likelihood of Linear regression

We expand the log-likelihood function of linear regression. If our matrix $\mathbf{X}$ has k columns, the vector $\mathbf{w}$ is $k \times 1$.

$$\text{Likelihood} = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

$$\text{log-likelihood} = \sum_{n=1}^{N} \left(\frac{-1}{2}\log(2\pi) - \log(\sigma) + \frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

$$\text{log-likelihood} = \frac{-N}{2}\log(2\pi) - N\log(\sigma) + \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

Linear Regression: First derivative of the log-likelihood

$$\log L = \frac{-N}{2}\log(2\pi) - N\log(\sigma) + \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = 0 - 0 + \frac{-1}{\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{w}^T\mathbf{x}_n)$$

$$\underbrace{\frac{\partial \log L}{\partial \mathbf{w}}}_{k\times 1} = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{t} - \mathbf{X}^T\mathbf{X}\mathbf{w})$$ see lecture 5-1 for the steps

Linear Regression: Second derivative of the log-likelihood

$$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^T} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{t} - \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} \mathbf{w} \right)$$

$$\underbrace{\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^T}}_{k \times k} = -\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}$$

The Hessian Matrix of the log-likelihood

This is the Hessian of the log-likelihood (matrix equivalent of the second-derivative)

$$\mathbf{H}(\log L) = \frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^T} = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}$$

The Fisher Information is equal to the negative expected value of the above.

$$\mathcal{I}(\mathbf{w}) = -\mathbb{E}\left[-\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right]$$

With respect to the outcome variable t_n and $t_n \sim N(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$, $\mathbf{X}$ and σ are constants. Thus,

$$\mathcal{I}(\mathbf{w}) = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}$$

Interpretation

For linear regression:

- The model of our data is $t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$. The $\mathbf{w}$ represents the "true" set of parameter weights in the population.
- σ^2 is the amount of variance we have in the noise parameters ϵ
- If there is lots of noise, σ^2 is large; the Fisher Information is small; we have less confidence in our estimates of $\mathbf{w}$
- If there is little of noise, σ^2 is small; the Fisher Information is large; we have more confidence in our estimates of $\mathbf{w}$