S&DS 242/542: Theory of Statistics

Lecture 19: Parameter estimation in misspecified models

"All models are wrong, but some are useful."

— George Box, 1976

Model misspecification

In our examples thus far, we typically started with the assumption $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. Here $f(x \mid \theta)$ plays two different roles:

- It describes the model we choose to *fit* to the data
- It describes the *true* distribution of the data

In reality these can be different, for various reasons:

- The true distribution of the data is unknown, and any model we fit is only an approximation to the truth
- The data is well-described by a complex model, but we choose to fit a simpler model that has more interpretable parameters
- We fit a model that is only intended for black-box prediction, rather than accurately describing the data-generating process

Q: How may we interpret the fitted parameter $\hat{\theta}$ and model $f(x \mid \hat{\theta})$ if $f(x \mid \theta)$ does not truly describe the distribution of the data?

Interpreting estimates in misspecified models

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^2$ are covariate-response pairs, which are IID with some unknown distribution $g(x, y)$.

We fit a linear regression model $Y \sim \beta_0 + \beta_1 X$ with parameters $\beta = (\beta_0, \beta_1)$ using least squares:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

Even if the true relation between X and Y is not linear, we may fit this model and interpret $(\hat{\beta}_0, \hat{\beta}_1)$ as representing the "best linear predictor" for Y based on X. For large n, we expect $(\hat{\beta}_0, \hat{\beta}_1)$ to be close to the minimizers of the population squared error

$$\underset{\beta_0, \beta_1}{\arg\min}\, \mathbb{E}_{(X,Y)\sim g}[(\beta_0 + \beta_1 X - Y)^2]$$

Interpreting estimates in misspecified models

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} g(x)$, for an unknown distribution $g(x)$.

We fit a model $f(x \mid \theta)$ with parameters $\theta \in \mathbb{R}^k$ using the method of moments. By definition, the estimated parameters $\hat{\theta}$ are those for which

$$\mathbb{E}_{\hat{\theta}}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \ldots, \quad \mathbb{E}_{\hat{\theta}}[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

This may be understood as estimating the distribution $f(x \mid \theta)$ within our model that matches $g(x)$ in its first k moments.

For large n, we expect $\hat{\theta}$ to be close to the value of θ such that

$$\mathbb{E}_{\theta}[X] = \mathbb{E}_{X\sim g}[X], \quad \ldots, \quad \mathbb{E}_{\theta}[X^k] = \mathbb{E}_{X\sim g}[X^k]$$

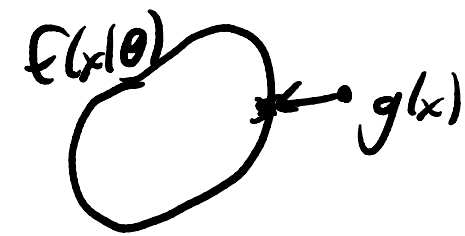
Interpreting estimates in misspecified models

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} g(x)$, for an unknown distribution $g(x)$. We fit a model $f(x \mid \theta)$ over parameters $\theta \in \Omega$ using maximum likelihood. What is the interpretation of $\hat{\theta}$?

Theorem

Let $D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta))$ be the Kullback-Leibler (KL) divergence from $f(x \mid \theta)$ to $g(x)$, and suppose that

$$\theta \mapsto D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta))$$



has a unique minimizer $\theta^* \in \Omega$. Then, under regularity conditions for $f(x \mid \theta)$, the MLE $\hat{\theta}$ converges to θ^* in probability as $n \to \infty$.

For large n, the MLE estimates the "KL projection" of the true distribution $g(x)$ onto our model $\{f(x \mid \theta) : \theta \in \Omega\}$.

Kullback-Leibler divergence

Kullback-Leibler divergence

For two discrete distributions with PMFs f and g on a sample space $\mathcal{X}$, the **Kullback-Leibler (KL) divergence** from f to g is

$$D_{\mathrm{KL}}(g\|f) = \sum_{x\in\mathcal{X}} g(x)\log\frac{g(x)}{f(x)}$$

For two continuous distributions with PDFs f and g on the real line, the KL divergence from f to g is similarly

$$D_{\mathrm{KL}}(g\|f) = \int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{f(x)}\,dx$$

In both cases, this may be written as an expectation

$$D_{\mathrm{KL}}(g\|f) = \mathbb{E}_g\left[\log\frac{g(X)}{f(X)}\right] = \mathbb{E}_g[\log g(X)] - \mathbb{E}_g[\log f(X)]$$

where $\mathbb{E}_g$ indicates expectation with respect to $X \sim g(x)$.

KL divergence between normal distributions

Let $f = \mathcal{N}(\mu_0, \sigma^2)$ and $g = \mathcal{N}(\mu_1, \sigma^2)$ with common variance σ^2.
To compute $D_{\mathrm{KL}}(g \| f)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}, \quad g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

$$\Rightarrow \log \frac{g(x)}{f(x)} = -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2}$$

$$= \left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu_1}{\sigma^2}x - \frac{\mu_1^2}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^2}x^2 - \frac{\mu_0}{\sigma^2}x + \frac{\mu_0^2}{2\sigma^2}\right)$$

$$= \frac{\mu_1 - \mu_0}{\sigma^2}x - \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

$$D_{KL}(g \| f) = \mathbb{E}_g\left[\frac{\mu_1 - \mu_0}{\sigma^2}X - \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}\right]$$

KL divergence between normal distributions

$$= \frac{\mu_1 - \mu_0}{\sigma^2} \underbrace{\mathbb{E}_g[X]}_{=\mu_1} - \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

$$= \frac{2\mu_1(\mu_1 - \mu_0)}{2\sigma^2} - \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

$$= \frac{\mu_1^2 - 2\mu_0\mu_1 + \mu_0^2}{2\sigma^2} = \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$$

So $D_{KL}(g \| f) = \frac{(\mu_1 - \mu_0)^2}{2\sigma^2}$

KL divergence between Bernoulli distributions

Let $f = \text{Bernoulli}(p)$, $g = \text{Bernoulli}(q)$. To compute $D_{\mathrm{KL}}(g\|f)$:

$$\log \frac{g(x)}{f(x)} = \begin{cases} \log \frac{q}{p} & \text{if } x=1 \\ \log \frac{1-q}{1-p} & \text{if } x=0 \end{cases}$$

$$\Rightarrow D_{KL}(g\|f) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right]$$

$$= q \cdot \log \frac{q}{p} + (1-q) \cdot \log \frac{1-q}{1-p}$$

Note that this expression is *not* symmetric in (p, q), so $D_{\mathrm{KL}}(g\|f) \neq D_{\mathrm{KL}}(f\|g)$.

KL divergence between Bernoulli distributions

For $f = \text{Bernoulli}(p)$ and $g = \text{Bernoulli}(q)$, we have

$$D_{\mathrm{KL}}(g\|f) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$$

For p close to q, this may be approximated by Taylor expansion:

$$\log p \approx \log q + (p-q)\cdot\frac{1}{q} + \frac{(p-q)^2}{2}\cdot\left(-\frac{1}{q^2}\right)$$

$$\Rightarrow \log q - \log p \approx -\frac{p-q}{q} + \frac{(p-q)^2}{2q^2}$$

$$\log(1-p) \approx \log(1-q) + (p-q)\cdot\left(-\frac{1}{1-q}\right) + \frac{(p-q)^2}{2}\cdot\left(-\frac{1}{(1-q)^2}\right)$$

$$\Rightarrow \log(1-q) - \log(1-p) \approx \frac{p-q}{1-q} + \frac{(p-q)^2}{2(1-q)^2}$$

KL divergence between Bernoulli distributions

$$D_{KL}(q\|p) = q(\log q - \log p) + (1-q)(\log(1-q) - \log(1-p))$$

$$\approx \frac{(p-q)^2}{2q} + \frac{(p-q)^2}{2(1-q)}$$

$$= \frac{(p-q)^2}{2q(1-q)} \approx \frac{(p-q)^2}{2p(1-p)}$$

Si $q \approx p$

$$D_{KL}(q\|p) \approx \frac{(p-q)^2}{2q(1-q)} \approx \frac{(p-q)^2}{2p(1-p)}$$

KL divergence between Binomial distributions

Let $f = \text{Binomial}(n, p)$ and $g = \text{Binomial}(n, q)$. Then

$$\begin{aligned}\log \frac{g(x)}{f(x)} &= \log \left(\binom{n}{x} q^x (1-q)^{n-x} \Big/ \binom{n}{x} p^x (1-p)^{n-x} \right) \\ &= x \log \frac{q}{p} + (n-x) \log \frac{1-q}{1-p}\end{aligned}$$

Applying $\mathbb{E}_g[X] = nq$,

$$\begin{aligned}D_{\mathrm{KL}}(g \| f) &= \mathbb{E}_g[X] \log \frac{q}{p} + (n - \mathbb{E}_g[X]) \log \frac{1-q}{1-p} \\ &= n \left(q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \right) \\ &= n \times D_{\mathrm{KL}}(\text{Bernoulli}(q) \| \text{ Bernoulli}(p))\end{aligned}$$

For p close to q, this is $D_{\mathrm{KL}}(g \| f) \approx n \times \frac{(p-q)^2}{2q(1-q)}$.

Properties of KL divergence

$D_{\mathrm{KL}}(g\|f)$ is an information-theoretic measure of discrepancy between g and f. In general, $D_{\mathrm{KL}}(f\|g) \neq D_{\mathrm{KL}}(g\|f)$.

It satisfies the following basic properties:

- If $f = g$, then $D_{\mathrm{KL}}(g\|f) = 0$ because $\log \frac{g(x)}{f(x)} = 0$ for all x.
- $D_{\mathrm{KL}}(g\|f) \geq 0$ for any f and g. This follows from applying Jensen's inequality to the convex function $-\log x$:

$$D_{KL}(g\|f) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right]$$

$$= \mathbb{E}_g\left[-\log \frac{f(x)}{g(x)}\right]$$

$$\geq -\log \mathbb{E}_g\left[\frac{f(x)}{g(x)}\right] = -\log \underbrace{\int_{-\infty}^{\infty} \frac{f(x)}{g(x)} \cdot g(x)\, dx}_{=1}$$

$$= 0$$

KL divergence within a parametric model

Consider a parametric model $f(x \mid \theta)$ and two parameters $\theta_0, \theta_1 \in \mathbb{R}$. Then

$$D_{\mathrm{KL}}(f(x \mid \theta_0) \,\|\, f(x \mid \theta_1)) = \underbrace{\mathbb{E}_{\theta_0}[\log f(X \mid \theta_0)]}_{=L(\theta_0)} - \underbrace{\mathbb{E}_{\theta_0}[\log f(X \mid \theta_1)]}_{=L(\theta_1)}$$

Fixing θ_0, let $L(\theta) = \mathbb{E}_{\theta_0}[\log f(X \mid \theta)]$ be the population log-likelihood function when the true parameter is θ_0. If θ_0 and θ_1 are close to each other, then we may apply the Taylor expansion

$$L(\theta_1) \approx L(\theta_0) + (\theta_1 - \theta_0)L'(\theta_0) + \frac{1}{2}(\theta_1 - \theta_0)^2 L''(\theta_0)$$

Recall that

- $L'(\theta_0) = 0$ because θ_0 maximizes $L(\theta)$
- $L''(\theta_0) = -I(\theta_0)$, the Fisher information at θ_0

KL divergence within a parametric model

Thus, for parameters θ_0 and θ_1 close to each other,

$$D_{\mathrm{KL}}(f(x \mid \theta_0) \,\|\, f(x \mid \theta_1)) \approx \frac{I(\theta_0)}{2}(\theta_1 - \theta_0)^2$$

The KL divergence is approximately the squared difference between the parameter values, scaled by one-half times the Fisher information.

This gives another interpretation of the Fisher information:
It relates the KL divergence — an information-theoretic measure of the discrepancy between $f(x \mid \theta_0)$ and $f(x \mid \theta_1)$ — to the squared difference of the model parameters.

The MLE in misspecified models

Interpretation of the MLE

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} g(x)$, for an unknown distribution $g(x)$. We fit a model $f(x \mid \theta)$ over parameters $\theta \in \Omega$ using maximum likelihood.

Theorem

Let $D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta))$ be the Kullback-Leibler (KL) divergence from $f(x \mid \theta)$ to $g(x)$, and suppose that

$$\theta \mapsto D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta))$$

has a unique minimizer $\theta^* \in \Omega$. Then, under regularity conditions for $f(x \mid \theta)$, the MLE $\hat{\theta}$ converges to θ^* in probability as $n \to \infty$.

Proof sketch

By definition, the MLE $\hat{\theta}$ maximizes

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta)$$

By the Law of Large Numbers, as $n \to \infty$, for each $\theta \in \Omega$ this converges in probability to the expected log-likelihood function

$$L(\theta) = \mathbb{E}_g[\log f(X \mid \theta)]$$

Here, $\mathbb{E}_g$ is now the expectation with respect to the true distribution $X \sim g(x)$.

In Lecture 14, we assumed that $g(x) = f(x \mid \theta^*)$ for some true parameter $\theta^* \in \Omega$. Then $L(\theta)$ is maximized at θ^*, and this explained consistency of the MLE.

Proof sketch

If $g(x)$ does not belong to our model $\{f(x \mid \theta) : \theta \in \Omega\}$, let us write

$$\begin{aligned} L(\theta) &= \mathbb{E}_g[\log f(X \mid \theta)] \\ &= \mathbb{E}_g[\log g(X)] - \mathbb{E}_g\left[\log \frac{g(X)}{f(X \mid \theta)}\right] \\ &= \mathbb{E}_g[\log g(X)] - D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta)) \end{aligned}$$

The first term $\mathbb{E}_g[\log g(X)]$ does not depend on θ. Thus the maximizer of $L(\theta)$ is the minimizer of $D_{\mathrm{KL}}(g(x) \,\|\, f(x \mid \theta))$, which is assumed in the theorem to be a unique value $\theta^* \in \Omega$. So we expect, as $n \to \infty$,

$$\hat{\theta} \to \theta^*$$

Example: MLE in a misspecified Poisson model

Suppose we fit the model Poisson(λ) to observations $X_1, \ldots, X_n$ whose true distribution is $g(x)$.

To compute $D_{KL}(g(x) \| f(x|\lambda))$ when

$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ is the PMF of Poisson(λ):

$$D_{KL}(g(x) \| f(x|\lambda)) = \mathbb{E}_g\left[\log \frac{g(X)}{f(X|\lambda)}\right]$$

$$= \mathbb{E}_g[\log g(X)] - \mathbb{E}_g[-\lambda + X\log\lambda - \log(X!)]$$

$$= \mathbb{E}_g[\log g(X)] + \lambda - \mathbb{E}_g[X]\cdot\log\lambda + \mathbb{E}_g[\log(X!)]$$

Example: MLE in a misspecified Poisson model

Set $0 = \frac{d}{d\lambda} D_{KL}(g(x) \| f(x|\lambda))$

$= 1 - \frac{\mathbb{E}_g[X]}{\lambda}$

Minimum of $D_{KL}(g(x) \| f(x|\lambda))$ is $\lambda^* = \mathbb{E}_g[X]$

So the MLE $\hat{\lambda}$ satisfies

$$\hat{\lambda} \to \lambda^* = \mathbb{E}_g[X]$$

in probability as $n \to \infty$. This coincides with the direct conclusion using the Law of Large Numbers and the explicit form $\hat{\lambda} = \bar{X}$.

Example: Asymptotic variance of the MLE

Continuing this example, since $\hat{\lambda} = \bar{X}$, by the CLT we have

$$\sqrt{n}\,(\hat{\lambda} - \lambda^*) \to \mathcal{N}(0, \mathrm{Var}_g[X])$$

in distribution as $n \to \infty$. So $\mathrm{Var}[\hat{\lambda}] \approx \frac{\mathrm{Var}_g[X]}{n}$ for large n.

If we estimate this variance based on the Fisher information in our assumed Poisson(λ) model, we would obtain an incorrect estimate:

$$I(\lambda) = \mathrm{Var}_\lambda\left[\frac{\partial}{\partial\lambda}\log\frac{e^{-\lambda}\lambda^X}{X!}\right] = \mathrm{Var}_\lambda\left[\frac{X}{\lambda} - 1\right] = \frac{1}{\lambda}$$

Thus, since $\hat{\lambda} \approx \lambda^* = \mathbb{E}_g[X]$ for large n, the usual plug-in estimate for $\mathrm{Var}[\hat{\lambda}]$ would be

$$\frac{1}{nI(\hat{\lambda})} = \frac{\hat{\lambda}}{n} \approx \frac{\lambda^*}{n} = \frac{\mathbb{E}_g[X]}{n} \neq \frac{\mathrm{Var}_g[X]}{n}$$

Uncertainty quantification in misspecified models

- It is often meaningful to fit a parametric model $f(x \mid \theta)$ that is only an approximation to the true distribution of data.
- If maximum likelihood is used to fit the model, then $\hat{\theta}$ may be understood as estimating a parameter θ^* representing the KL projection of the data distribution onto the assumed model.
- However $\mathrm{Var}[\hat{\theta}]$ is, in general, **not** given by $\frac{1}{nI(\theta^*)}$.

  Thus the usual model-based estimate of $\mathrm{Var}[\hat{\theta}]$ given by $\frac{1}{nI(\hat{\theta})}$ may be incorrect in a misspecified model, and the confidence interval $\hat{\theta} \pm z^{(\alpha/2)} \sqrt{\frac{1}{nI(\hat{\theta})}}$ may have incorrect coverage for θ^*.