MATH 559: BAYESIAN THEORY AND METHODS
TRADITIONAL MODEL SELECTION: AIC AND BIC

Model Selection and Parameter Estimation : Suppose we wish to choose one from a collection of models described by densities $f_1, f_2, \ldots, f_K$ with parameters $\theta_1, \theta_2, \ldots, \theta_K$. Let the true, data generating model be denoted f_0. We consider the KL divergence between f_0 and f_k:

$$KL(f_0, f_k) = \int \log\left(\frac{f_0(y)}{f_k(y;\theta_k)}\right) f_0(y)\, dy$$

and aim to choose the model using the criterion

$$\widehat{k} = \arg\min_k KL(f_0, f_k).$$

- We have that
$$KL(f_0, f_k) = \int \log f_0(y) f_0(y)\, dy - \int \log f_k(y;\theta_k) f_0(y)\, dy$$
so we then may choose k by
$$\widehat{k} = \arg\max_k \int \log f_k(y;\theta_k) f_0(y)\, dy.$$

- In reality, θ_k are typically unknown, so we consider estimating them using maximum likelihood procedures. We consider
$$KL(f_0, \widehat{f}_k) = \int \log\left(\frac{f_0(y)}{f_k(y;\widehat{\theta}_k)}\right) f_0(y)\, dy$$
where $\widehat{\theta}_k$ is obtained by maximizing the likelihood under model k, that is maximizing
$$\ell_{kn}(\theta) = \sum_{i=1}^{n} \log f_k(y_i;\theta)$$
with respect to θ for data $y_1, \ldots, y_n$.

- We will use the notation

$$\ell_k(y;\theta) \equiv \log f_k(y;\theta) \qquad \text{Single datum log-likelihood}$$

$$\ell_{kn}(\theta) \equiv \sum_{i=1}^{n} \log f_k(y_i;\theta) \qquad \text{Full data log-likelihood}$$

Asymptotic results for estimation under misspecification

- *Maximum likelihood as minimum KL:* The 'true' parameter value for model k, θ_{0k}, is defined as

$$\theta_{0k} = \arg\min_{\theta} KL(f_0, f_k) = \arg\min_{\theta} \int \log\left(\frac{f_0(y)}{f_k(y;\theta)}\right) f_0(y)\, dy.$$

For the minimization, under standard regularity conditions

$$\begin{aligned}\frac{\partial}{\partial\theta}\left\{\int \log\left(\frac{f_0(y)}{f_k(y;\theta)}\right) f_0(y)\, dy\right\} &= -\frac{\partial}{\partial\theta}\left\{\int \log f_k(y;\theta) f_0(y)\, dy\right\} \\ &= -\int\left\{\frac{\partial}{\partial\theta}\log f_k(y;\theta)\right\} f_0(y)\, dy = -\mathbb{E}_{f_0}\left[\dot{\ell}_k(Y;\theta)\right]\end{aligned}$$

and so θ_{0k} solves

$$\mathbb{E}_{f_0}\left[\dot{\ell}_k(Y;\theta_{0k})\right] = \mathbf{0}.$$

Under identifiability, we assume that is that there is a single θ_{0k} which solves this equation.

- *Asymptotics recap:* The sample-based equivalent calculation dictates that for the estimate $\widehat{\theta}_k$

$$\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_k(y_i;\widehat{\theta}_k) = \mathbf{0}$$

which coincides with ML estimation. Under identifiability assumptions, as by the Strong Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_k(Y_i;\theta_{0k}) \xrightarrow{a.s.} \mathbf{0},$$

we must have that $\widehat{\theta}_k \longrightarrow \theta_{0k}$ with probability 1 as $n \longrightarrow \infty$. If

$$\Psi_k(y;\theta_k) = \left.\frac{\partial^2 \log f_k(y;\theta)}{\partial\theta\partial\theta^\top}\right|_{\theta=\theta_k} \equiv \ddot{\ell}_k(y;\theta_k)$$

let

$$\mathcal{J}_{0k}(\theta_k) = \mathbb{E}_{f_0}\left[-\Psi_k(Y;\theta_k)\right].$$

The quantity $\widehat{J}_k(\theta_k)$ is the sample-based version

$$\widehat{J}_k(\widehat{\theta}_k) = -\frac{1}{n}\sum_{i=1}^{n} \Psi_k(Y_i;\widehat{\theta}_k)$$

and note that $\widehat{J}_k(\widehat{\theta}_k) \longrightarrow \mathcal{J}_k(\theta_{0k})$ with probability 1 as $n \longrightarrow \infty$. Recall also the definition

$$\mathcal{I}_{0k}(\theta_k) = \mathbb{E}_{f_0}\left[\ell_k(Y;\theta_k)\ell_k(Y;\theta_k)^\top\right].$$

Finally, we have the sample-based quantity

$$\widehat{I}_k(\widehat{\theta}_k) = \frac{1}{n}\sum_{i=1}^{n} \ell_k(Y_i;\widehat{\theta}_k)\ell_k(Y_i;\widehat{\theta}_k)^\top.$$

- *Statistical properties:* Using a second order Taylor expansion around $\widehat{\theta}_k$ of the first derivative of the log-likelihood at $\theta = \theta_{0k}$, we have

$$\begin{aligned}\dot{\ell}_{kn}(\theta_{0k}) &= \dot{\ell}_{kn}(\widehat{\theta}_k) + \ddot{\ell}_{kn}(\widehat{\theta}_k)(\theta_{0k} - \widehat{\theta}_k) + \frac{1}{2}(\theta_{0k} - \widehat{\theta}_k)^\top \dddot{\ell}_{kn}(\theta^*)(\theta_{0k} - \widehat{\theta}_k) \\ &= \ddot{\ell}_{kn}(\widehat{\theta}_k)(\theta_{0k} - \widehat{\theta}_k) + R_n(\theta^*)\end{aligned}$$

as $\dot{\ell}_{kn}(\widehat{\theta}_k) = \mathbf{0}$, $||\widehat{\theta}_k - \theta^*|| < ||\widehat{\theta}_k - \theta_{0k}||$, and where the remainder term is $R_n(\theta^*)$. By definition

$$-\ddot{\ell}_{kn}(\widehat{\theta}_k) = n\widehat{J}_k(\widehat{\theta}_k).$$

We have that $R_n(\theta^*) = \mathrm{o}_p(n)$, that is,

$$\frac{R_n(\theta^*)}{n} \xrightarrow{p} 0,$$

and

$$\widehat{J}_k(\widehat{\theta}_k) \xrightarrow{a.s.} \mathcal{J}_{0k}(\theta_{0k})$$

as $n \longrightarrow \infty$. Rewriting the expansion above, we have

$$\sqrt{n}(\widehat{\theta}_k - \theta_{0k}) = \left\{\widehat{J}_k(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\dot{\ell}_{kn}(\theta_{0k})\right\} + \left\{\widehat{J}_k(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{R_n(\theta^*)}{n}\right\}.$$

Thus, by the Central Limit Theorem

$$\sqrt{n}(\widehat{\theta}_k - \theta_{0k}) \xrightarrow{d} Normal_d(\mathbf{0}, \Sigma_k)$$

say, where Σ_k denotes the limiting variance-covariance matrix when sampling is under f_0, and

$$\frac{1}{\sqrt{n}}\dot{\ell}_{kn}(\theta_{0k}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_k(Y_i;\theta_{0k}) \xrightarrow{d} Normal_d(\mathbf{0}, \mathcal{I}_{0k}(\theta_{0k})).$$

Finally, by Slutsky's Theorem

$$\left\{\widehat{J}_k(\widehat{\theta}_k)\right\}^{-1}\left\{\frac{R_n(\theta^*)}{n}\right\} \xrightarrow{p} \mathbf{0}.$$

Therefore, by equating the asymptotic variances of the above quantities, we must have

$$\mathcal{I}_{0k}(\theta_{0k}) = \mathcal{J}_{0k}(\theta_{0k})\Sigma_k\mathcal{J}_{0k}(\theta_{0k})$$

yielding that

$$\Sigma_k = \{\mathcal{J}_{0k}(\theta_{0k})\}^{-1}\mathcal{I}_{0k}(\theta_{0k})\{\mathcal{J}_{0k}(\theta_{0k})\}^{-1}$$

- Recall that under correct specification, we have the identity

$$\mathcal{I}_{0k}(\theta_{0k}) = \mathcal{J}_{0k}(\theta_{0k})$$

so we have the simplification

$$\Sigma_k = \{\mathcal{I}_{0k}(\theta_{0k})\}^{-1} = \{\mathcal{J}_{0k}(\theta_{0k})\}^{-1}.$$

Model Selection using Minimum KL

- *The Expected KL Divergence:* Suppose that $\mathbf{X}$ and $\mathbf{Y}$ represent data sets of size n drawn independently from f_0, and consider the quantity

$$\int \log f_k(\mathbf{x}; \widehat{\theta}_k(\mathbf{Y})) f_0(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{X|Y}[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]$$

We will write X and Y in the subscript of the expectations for clarity.

In the conditional expectation, we consider the expectation with respect to the distribution of X. That is, we

  - collect data $\mathbf{Y}$, use the data to construct the estimator $\widehat{\theta}_k(\mathbf{Y})$, then
  - compute the expectation of the function $\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{Y}))$ over the distribution of $\mathbf{X}$ for fixed $\mathbf{Y}$; this is a random quantity, a function of $\mathbf{Y}$,
  - consider the expectation of this result over the distribution of $\mathbf{Y}$,
  - choose k by

$$\widehat{k} = \arg\max_k \mathbb{E}_Y[\mathbb{E}_{X|Y}[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]]$$

We consider first the expansion of the inner integral around the true value for model k, θ_{0k}. For an arbitrary data set $\mathbf{y}$, under regularity conditions

$$\begin{aligned}\mathbb{E}_X[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{y}))|\mathbf{Y} = \mathbf{y}] = \; & \mathbb{E}_X[\log f_k(\mathbf{X}; \theta_{0k})] + n\mathbb{E}_X\left[\dot{\ell}_k(X; \theta_{0k})\right]^\top (\widehat{\theta}_k(\mathbf{y}) - \theta_{0k}) \\ & + \frac{n}{2}(\widehat{\theta}_k(\mathbf{y}) - \theta_{0k})^\top \mathbb{E}_X\left[\ddot{\ell}_k(X; \theta_{0k})\right](\widehat{\theta}_k(\mathbf{y}) - \theta_{0k}) + \mathrm{o}_p(n)\end{aligned}$$

where X denotes a single random variable. By definition $\mathbb{E}_X\left[\dot{\ell}_k(X; \theta_{0k})\right] = \mathbf{0}$, and

$$\mathbb{E}_X\left[\ddot{\ell}_k(X; \theta_{0k})\right] = -\mathcal{J}_{0k}(\theta_{0k})$$

so therefore

$$\mathbb{E}_X[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{y}))|\mathbf{Y} = \mathbf{y}] = \mathbb{E}_X[\log f_k(\mathbf{X}; \theta_{0k})] - \frac{n}{2}(\widehat{\theta}_k(\mathbf{y}) - \theta_{0k})^\top \mathcal{J}_k(\theta_{0k})(\widehat{\theta}_k(\mathbf{y}) - \theta_{0k}) + \mathrm{o}_p(n)$$

We then must compute

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]].$$

The term $\mathbb{E}_X[\log f_k(\mathbf{X}; \theta_{0k})]$ does not depend on $\mathbf{Y}$. For the quadratic term, under sampling data $\mathbf{Y}$ independently from f_0, we may rewrite

$$n(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k})^\top \mathcal{J}_{0k}(\theta_{0k})(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k}) = \{\sqrt{n}(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k})\}^\top \{\mathcal{J}_{0k}(\theta_{0k})\} \{\sqrt{n}(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k})\}$$

where from the previous result

$$\sqrt{n}(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k}) \xrightarrow{d} Normal_d(0, \Sigma_k).$$

Therefore, by standard results for quadratic forms

$$\mathbb{E}_Y\left[n(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k})^\top \mathcal{J}_{0k}(\theta_{0k})(\widehat{\theta}_k(\mathbf{Y}) - \theta_{0k})\right] \xrightarrow{a.s.} \text{Trace}\left(\mathcal{J}_{0k}(\theta_{0k})\Sigma_k\right)$$

and so we have that

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[\log f_k(\mathbf{X}; \widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]] = \mathbb{E}_X[\log f_k(\mathbf{X}; \theta_{0k})] - \frac{1}{2}\text{Trace}\left(\mathcal{J}_{0k}(\theta_{0k})\Sigma_k\right) + \mathrm{o}_p(n) \qquad (*)$$

The first term on the right hand side cannot be computed, as θ_{0k} is not known. Thus we repeat the same operation, but instead expand

$$\mathbb{E}_X[\log f_k(\mathbf{X};\theta_{0k})]$$

about $\widehat{\theta}_k(\mathbf{X})$. We have

$$\log f_k(\mathbf{x};\theta_{0k}) = \ell_{kn}(\widehat{\theta}_k(\mathbf{x}))+\dot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{x}))^\top(\theta_{0k}-\widehat{\theta}_k(\mathbf{x}))+\frac{1}{2}(\widehat{\theta}_k(\mathbf{x})-\theta_{0k})^\top\ddot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{x}))(\widehat{\theta}_k(\mathbf{x})-\theta_{0k})+\mathrm{o}(n)$$

Again $\dot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{x})) = \mathbf{0}$, and writing

$$-(\widehat{\theta}_k(\mathbf{x}) - \theta_{0k})^\top\ddot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{x}))(\widehat{\theta}_k(\mathbf{x}) - \theta_{0k})$$

as

$$\{\sqrt{n}(\widehat{\theta}_k(\mathbf{x}) - \theta_{0k})\}^\top\left\{-\frac{1}{n}\ddot{\ell}_k(\widehat{\theta}_k(\mathbf{x}))\right\}\{\sqrt{n}(\widehat{\theta}_k(\mathbf{x}) - \theta_{0k})\}$$

we have by the results discussed above that the expectation over $\mathbf{X}$ of this quadratic form converges to

$$\text{Trace}\,(\mathcal{J}_{0k}(\theta_{0k})\Sigma_k)$$

as by standard theory

$$\sqrt{n}(\widehat{\theta}_k(\mathbf{X}) - \theta_k) \xrightarrow{d} Normal_d(0,\Sigma_k).$$

Thus

$$\mathbb{E}_X[\log f_k(\mathbf{X};\theta_{0k})] = \mathbb{E}_X[\log f_k(\mathbf{X};\widehat{\theta}_k(\mathbf{X}))] - \frac{1}{2}\text{Trace}\,(\mathcal{J}_{0k}(\theta_{0k})\Sigma_k) + \mathrm{o}_p(n).$$

Therefore, by (*) we now have

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[\log f_k(\mathbf{X};\widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]] = \mathbb{E}_X[\log f_k(\mathbf{X};\widehat{\theta}_k(\mathbf{X}))] - \text{Trace}\,(\mathcal{J}_{0k}(\theta_{0k})\Sigma_k) + \mathrm{o}_p(n)$$

By the earlier identity

$$\mathcal{J}_{0k}(\theta_{0k})\Sigma_k = \mathcal{I}_{0k}(\theta_{0k})\,\{\mathcal{J}_{0k}(\theta_{0k})\}^{-1}$$

Therefore, $\mathbb{E}_Y[\mathbb{E}_{X|Y}[\log f_k(\mathbf{X};\widehat{\theta}_k(\mathbf{Y}))|\mathbf{Y}]]$ can be consistently estimated by

$$\log f_k(\mathbf{y};\widehat{\theta}_k(\mathbf{y})) - \text{Trace}\left(\widehat{I}_k(\widehat{\theta}_k(\mathbf{y}))\left\{\ \widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))\right\}^{-1}\right)$$

If the approximating model f_k is close in KL terms to the true model f_0, we would expect that

$$\text{Trace}\left(\mathcal{J}_{0k}(\theta_{0k})\,\{\mathcal{I}_{0k}(\theta_{0k})\}^{-1}\right) \doteq \dim(\theta_k) = d_k$$

say. This yields the criterion: choose k to maximize

$$\log f_k(\mathbf{y};\widehat{\theta}_k(\mathbf{y})) - d_k$$

or equivalently to minimize

$$-2\log f_k(\mathbf{y};\widehat{\theta}_k(\mathbf{y})) + 2d_k$$

This is **Akaike's Information Criterion** (AIC).

Note: The approximation

$$\text{Trace}\left(\widehat{I}_k(\theta_k(\mathbf{y}))\left\{\widehat{J}_k(\theta_k(\mathbf{y}))\right\}^{-1}\right) \doteq d_k$$

is potentially poor, but in fact this empirical trace can be computed and used in place of d_k.

- *The Bayesian Information Criterion:* The Bayesian Information Criterion (BIC) uses an approximation to the marginal likelihood function to justify model selection. For data $\mathbf{y}$, we have the posterior distribution for model k as

$$\pi_n(\theta) = \frac{\mathcal{L}_{kn}(\theta)\pi_0(\theta)}{\int \mathcal{L}_{kn}(t)\pi_0(t)\ dt}.$$

where $\mathcal{L}_{kn}(\theta)$ is the likelihood function. The denominator is

$$f_k(\mathbf{y}) = \int \mathcal{L}_{kn}(t)\pi_0(t)\ dt$$

that is, the *marginal likelihood*. We have that

$$\ell_{kn}(\theta) = \log \mathcal{L}_{kn}(\theta) = \sum_{i=1}^{n} \log f_k(y_i;\theta)$$

is the log-likelihood for model k. By a Taylor expansion of $\ell_{kn}(\theta)$ around $\widehat{\theta}_k(\mathbf{y})$, we have

$$\ell_{kn}(\theta) = \ell_{kn}(\widehat{\theta}_k(\mathbf{y})) + (\theta - \widehat{\theta}_k(\mathbf{y}))^\top \dot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{y})) + \frac{1}{2}(\theta - \widehat{\theta}_k)^\top \ddot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{y}))(\theta - \widehat{\theta}_k(\mathbf{y})) + \mathrm{o}(n).$$

We have that $\dot{\ell}_{kn}(\widehat{\theta}_k(\mathbf{y})) = \mathbf{0}$, and by the usual quadratic approximation approach we have that

$$\mathcal{L}_{kn}(\theta) \doteq \mathcal{L}_{kn}(\widehat{\theta}_k) \exp\left\{-\frac{n}{2}(\theta - \widehat{\theta}_k(\mathbf{y}))^\top \widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))(\theta - \widehat{\theta}_k(\mathbf{y}))\right\}.$$

Suppose that prior $\pi_0(\theta) = c$ in the neighbourhood of $\widehat{\theta}_k(\mathbf{y})$. Then, for large n

$$\begin{aligned} f_k(\mathbf{y}) &= \int \mathcal{L}_{kn}(t)\pi_0(t)\ dt \\ &\doteq \int c\mathcal{L}_{kn}(\widehat{\theta}(\mathbf{y})) \exp\left\{-\frac{n}{2}(t - \widehat{\theta}_k(\mathbf{y}))^\top \widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))(t - \widehat{\theta}_k(\mathbf{y}))\right\}\ dt \\ &= c\mathcal{L}_k(\widehat{\theta}(\mathbf{y}))(2\pi)^{d_k/2}|n\widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))|^{-1/2} \end{aligned}$$

where $d_k = \dim(\theta_k)$, as the integrand is proportional to a multivariate Normal pdf. But

$$|n\widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))|^{-1/2} = n^{-d_k/2}|\widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))|^{-1/2}.$$

Therefore the marginal likelihood approximation is

$$f_k(\mathbf{y}) \doteq \mathcal{L}_{kn}(\widehat{\theta}(\mathbf{y}))(2\pi)^{d_k/2} n^{-d_k/2} |\widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))|^{-1/2}$$

or on the $-2\log$ scale we have

$$-2\log f_k(\mathbf{y}) \doteq -2\ell_{kn}(\widehat{\theta}(\mathbf{y})) + d_k \log n + \text{constant}$$

where

$$\text{constant} = -d_k \log(2\pi) + \log|\widehat{J}_k(\widehat{\theta}_k(\mathbf{y}))| - 2\log c.$$

is $\mathrm{o}(1)$ and is hence negligible for large n. Therefore, the BIC is defined as

$$\text{BIC} = -2\ell_{kn}(\widehat{\theta}_k(\mathbf{y})) + d_k \log n = -2\sum_{i=1}^{n} \log f_k(y_i;\widehat{\theta}_k(\mathbf{y})) + d_k \log n.$$