# Model Selection

General Strategies:

- Penalize
- Cross validation
- Model posterior
- Model Averaging

~~Pbs~~ Issue:

- Models with more DOF appear to fit training data better (but predict test data worse)
- Can't select on fitting LLH alo

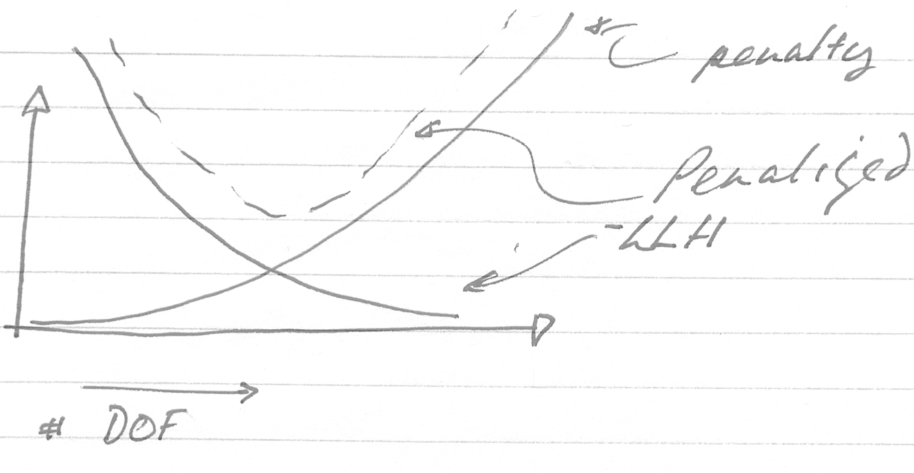
## Penalization methods

- wish to select from $M_1 \ldots M_K$
- compare $L_1(\hat{\theta}) + P(M_1)$

$$\vdots$$

$$L_K(\hat{\theta}) + P(M_K)$$

Typically




Popular penalties:

AIC

$$-2\log(L(\hat{\theta}|y)) + 2K$$

# of parameters in model

$AIC_c$

# of sam

$$-2\log(L(\hat{\theta}|y)) + 2K\left(\frac{n}{n-K-1}\right)$$

~~MDL~~

BIC ≡ MDL

$$-2\log(L(\hat{\theta}|y)) + K\log(n)$$

## Which one to use?

→ never AIC, unless $n \gg K$

→ otherwise, opinion is divided

## mechanics

- fit each of $M_1 \ldots M_K$
- evaluate favored criterion
- smallest value wins.

## Cross validation

- AIC, $AIC_c$ are attempts to estimate

$$KL(f, g(\theta)) = \int f(x) \log\left(\frac{f(x)}{g(x|\hat{\theta})}\right) dx$$

- But we might do this directly.

Assume we fit model to $x_1 \cdots x_n$,

now evaluate $\frac{-1}{N-n} \sum_{i=n+1}^{N} \log p(x_i|\hat{\theta})$

$$\approx - \int f \log p(x|\hat{\theta})\, dx .$$

$$= KL(f, p(x|\hat{\theta})) + H_f$$

· we don't know this, but it is shared for all models

∴ take model with smallest held-out averaged LLH

→ <u>waste of data</u>:

<u>soln</u>: Average over multiple UAR splits of $x_i$

→ <u>expensive computationally</u>:

- <u>buy a faster comp</u>

Model posteriors and Bayesian m.s.

typical Bayesian model, we will have

$M_1 \cdots M_K \leftarrow$ Models

each will have $\theta \leftarrow$ hyperparameters which parametrize the priors of the parameters. E.g. GP, we had length scale, ~~ratio of~~ model var, ~~to~~ noise

and parameters $w \leftarrow$ eg. pars of linear regression.

We will select a model by

$$p(M_i \mid y, X) = \frac{p(y \mid X, M_i)\, p(M_i)}{p(y \mid X)}$$

Now $p(y|X, M_i)$

$$= \int p(y|X, M_i, \theta)\, p(\theta|H_i)\, d\theta$$

$\boxed{\text{hyper priors}}$

~~and~~

~~$p(\theta$~~

<u>Marginal likelihood, or evidence</u>

$$p(y|X, M_i, \theta) = \int p(y|X, M_i, \theta, w)\, p(w|\theta, H_i)\, dw$$

<u>Problem</u>: · all these integrals

<u>Strategies</u>:

· get lucky, and have analy

· sampled ests

· fix $\theta = \hat{\theta}$ and work with

$p(\cancel{y} M_i | y, X, \hat{\theta})$

## Model Averaging

e.g. • $$p(y_*|y, X, x_*) = \int p(y_*|y, X, x_*, w)\, p(w|y, X)\, d$$

• In ~~f~~ some strong sense, best thing to do

$$p(y_*|y, X, x_*) = \sum_i p(y_* | y, X, x_*, M_i)\, p(M_i | y, X).$$

BUT

- Can be violently impractical
- Can obscure comprehension

PRO's

- occasional, important practical examples where not averaging ~~obs~~ leads to false sense of secu