A Bayesian view of model complexity

Angelika van der Linde
University of Bremen, Germany

1. Intuitions
2. Measures of dependence
   between observables and parameters
3. Occurrence in predictive model comparison
4. Evaluation
5. Discussion

1. Intuitions

1.1 What is a statistical model ?

for a Bayesian: likelihood *and* prior

$$\mathcal{M} = (\{p(y|\theta)|\theta \in \Theta\},\, p(\theta))$$

- note: $\mathcal{M}$ focused on θ
- special case: $p(\theta)$ not informative

1.2 What is model complexity (mc) ?

(i) $\theta \longrightarrow y$

explanatory power of θ for y
potential of fitting y with θ

- mc = no of parameters

$$mc = p$$

gen. in lin. regression (Wahba, 1990):

$$Y = \mu + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad \widehat{\mu} = Sy$$

- $$mc = tr(S)$$

(ii) $y \longrightarrow \theta$

discriminatory power of y for θ

sensitivity of $\widehat{\theta}$ to y

how hard to learn θ from y

estimation variance

in linear regression:

$$tr(cov(SY)) = \sigma^2 tr(S^2) \underset{S\, orth.\, proj.}{=} \sigma^2 tr(S)$$

- $$mc = tr(S) = \sum_i s_{ii}$$

  s_{ii} sensitivity of estimate $\widehat{\mu}_i$ to y_i

- $$gdf(\theta) = \sum_i cov_{Y|\theta}(\widehat{\mu}_i(Y), Y_i)$$

  (Ye, 1998)

(iii) information criteria for model comparison

$$IC = fit + complexity$$

$$BIC = -2 \log p(y|\widehat{\theta}_{ML}(y)) + 2p \log n$$

$$AIC = -2 \log p(y|\widehat{\theta}_{ML}(y)) + 2p$$

$$DIC = -2 \log p(y|E(\vartheta|y)) + 2p_D$$

$$
\begin{aligned}
p_D &= -2E_{\theta|y}(\log p(y|\theta)) - 2\log p(y|E(\theta|y)) \\
&= \overline{D} - D(\overline{\theta})
\end{aligned}
$$

and many others (GIC, TIC, BPIC, WAIC)

- mc by comparison rather elusive

conclusion: issues in model complexity

- duality $y \leftrightarrows \theta$
- but: $y \longrightarrow \theta$ or $y \longrightarrow \widehat{\theta}$?
- idea:

  $\boxed{\text{mc} = \text{measure of dependence between } y \text{ and } \theta}$

  Bay: θ random, no problem
  freq: θ fix, but $\widehat{\theta}(Y)$ random

2. Measures of dependence

"mutual information"

$$I(Y, \Theta) = E_{\theta y}[\log \frac{p(\theta, y)}{p(\theta)p(y)}]$$

symmetric

$$J(Y, \Theta) = I(Y, \Theta) + E_{\theta} E_{y}[\log \frac{p(\theta)p(y)}{p(\theta, y)}]$$

properties

- dual in y and θ
- *measure of variability of* $p(y|\theta)$ w.r.t. $p(y)$

$$J(Y, \Theta) = E_\theta J(p(y|\theta), p(y))$$

- in exponential families

$$J(Y, \Theta) = tr(cov_Y(E(\theta|Y), t(Y))) = tr(cov_\Theta(\theta, E(t(Y)|\theta)))$$

- prior and posterior version:
  prior with Y and Θ
  posterior with future replication $\widetilde{Y}$ and Θ_{post}
- coincidence with $tr(S)$ in regression ?
  prior: NO
  posterior: YES !

posterior J in detail

$$J(\widetilde{Y}, \Theta_{post}) = E_{\theta|y} J(p(\widetilde{y}|\theta), p(\widetilde{y}|y))$$

$$J(\widetilde{Y}, \Theta_{post}) = \text{variability of } p(\widetilde{y}|\theta) \text{ having learnt about } \theta$$

$$J(\widetilde{Y}, \Theta_{post}) \underset{\text{expofams}}{=} E_{\theta|y} J(p(\widetilde{y}|\theta), p(\widetilde{y}|\overline{\theta}))$$

3. Occurrence in predictive model comparison

3.1 Targets

assessments of models according to

- data fit: '*prior prediction*' of Y by Θ
- fit of future obs: '*posterior prediction*' of $\widetilde{Y}$ by Θ_{post}
- parameters averaged w.r.t. Θ or Θ_{post}
- represented by $\widehat{\theta}(y)$

predictive success measured by e.g.

	$prior$	$post$
ave	$\log E_\Theta[p(y\|\theta)] = \log p(y)$	$E_{\Theta_{post}}[\log p(\widetilde{y}\|\theta)]$
rep	NA	$\log p(\widetilde{y}\|\widehat{\theta}(y))$

should be large,
prior: for data y
posterior: on average over $\widetilde{Y}$ (and Y, if rep ?)
Bay and freq versions !

examples of predictive targets

- model evidence: $\log E_{\Theta}[p(y|\theta)] = \log p(y)$, related to BIC and Bayes factors

  prior predictive, average

- AIC: $E_{y|\theta_0} E_{\widetilde{y}|\theta_0}[-2 \log p(\widetilde{y}|\widehat{\theta}_{ML}(y))]$

  posterior predictive, representative

- DIC: $E_{\widetilde{y}|y}[-2 \log p(\widetilde{y}|\overline{\theta})]$

  posterior predictive, representative

- DIC+, BPIC: $E_{\widetilde{y}|y} E_{\Theta_{post}}[-2 \log p(\widetilde{y}|\theta)]$

  posterior predictive, average

3.2 Key decompositions

$$-\log p(y) = -\log p(y|\theta) + \log \frac{p(y,\theta)}{p(y)p(\theta)} \tag{1}$$

$$-\log p(y|\theta^+) = -\log p(y|\theta) + \log \frac{p(y|\theta)}{p(y|\theta^+)} \tag{2}$$

taking expectations yields

$$E_{Y,\Theta}[\log \frac{p(y,\theta)}{p(y)p(\theta)}] = I(Y,\Theta) = E_\Theta I(p(y|\theta), p(y)) \tag{3}$$

$$E_{Y,\Theta}[\log \frac{p(y|\theta)}{p(y|\theta^+)}] = E_\Theta I(p(y|\theta), p(y|\theta^+)) \tag{4}$$

more building blocks

- symmetry assumption (approximation)

$$\boldsymbol{J}(p(y|\theta), p(y|\theta^+)) \approx \boldsymbol{2I}(p(y|\theta), p(y|\theta^+))$$

  achievable under second order Taylor approx

- link between 'ave' and 'rep' in expofams

$$J(Y, \Theta) = E_\Theta J(p(y|\theta), p(y|E(\theta)))$$

conclusions

- $IC = fit + complexity$
  is based on decompositions
  major variants

$$E_Y[-\log p(y)] = H(Y) = H(Y|\Theta) + I(Y,\Theta)$$

$$E_Y E_\Theta[-\log p(Y|\Theta)] = H(Y|\Theta) + J(Y,\Theta)$$

- a cloud of targets yields a cloud of measures of mc
  but all can be identified as variants of KL-divergences

- problem

  decompositions hold

    only for *model specific* expectations w.r.t. $Y, \widetilde{Y}$

    not for *true/common* expectations w.r.t. $Y, \widetilde{Y}$!!!

  invocation of a '*good model assumption*'

3.3 summary of crucial issues

- average targets are Bayesian
  representative targets are frequentist
- Is it meaningful to invoke a good model assumption already in the definition of a predictive target (eg DIC) ?
- Is there a definite trade-off between fit and complexity in model comparison ?
  YES! not in any particular IC,
  but according to the decompositions (laws of prob)

- Is is meaningful to think of models
  as defined by a focus
  and of mc depending on focus and data ?

- how to assess *uniformity* of performance, i.e.
  accuracy of approximations and estimates
  used in the derivation of an IC
  across models ???

  decompositions may provide benchmarks

4. Evaluations of posterior mc

- Gaussian case $Y, \widetilde{Y} \sim N(\theta, \Sigma)$, $\theta \sim N(0, K)$

$$
\begin{aligned}
& J(\widetilde{Y}, \Theta_{post}) \\
= \; & tr(I_p + \Sigma K^{-1})^{-1} \\
= \; & p(\frac{n\tau^2}{\sigma^2 + n\tau^2}) \quad \rightarrow p \\
\Sigma = \frac{s^2}{n} I_n & \qquad\qquad \tau^2 \rightarrow \infty \\
K = \tau^2 I_n & \qquad\qquad n \rightarrow \infty
\end{aligned}
$$

- exponential families

$$J(\widetilde{Y}, \Theta_{post}) = tr(cov_{\Theta}(\theta, E(t(Y)|\theta))$$

  p_D of DIC estimates $J(\widetilde{Y}, \Theta_{post})$
  in exponential families

- general, by simulation (Plummer, 2002)

$$J(\widetilde{Y}, \Theta_{post}) = E_{\Theta^{(1)}} E_{\Theta^{(2)}}[I(p(y|\theta^{(1)}), p(y|\theta^{(2)}))]$$

5. Discussion

- Bayesian view
  - allows to think of mc as measure of dependence
  - emphasizes average targets
  thus leading more naturally to decomp. of entropy
  than Taylor expansions for representative targets

- ICs only one option to estimate targets
  alternatives:
  - with independence assumption: CV
  - without true/common distribution of $Y, \widetilde{Y}$:
  simulation

- singular models without regularity assumptions
  under investigation in machine learning community
  (Watanabe, WAIC, CV)
- information theory
  drives statistical thinking