Model comparison: Deviance-based approaches

Patrick Breheny

February 19

Model comparison

- Thus far, we have looked at residuals in a fairly exploratory fashion to motivate the need for more flexible models
- Our next two lectures will focus on the issue of model comparison using more objective/systematic approaches
- Today's topic is the development of an AIC-like criterion for evaluating a series of models in terms of their predictive ability
- Thursday's lecture will focus on model comparison from a different perspective, that of considering, given two models M_1 and M_2, the probabilities $\Pr(M_1)$ and $\Pr(M_2)$ of each model being the correct one

Deviance

- A useful measure of how well the model fits the data, whether or frequentist and Bayesian, is the *deviance*:

$$D(\theta) = -2 \log p(\mathbf{y}|\theta)$$

- Remarks:
  - High values of $D(\theta)$ indicate low values of the log-likelihood and that the data deviates substantially from the model's assumptions
  - For normally distributed data with σ^2 treated as known, $D(\theta) = \mathrm{RSS}$
  - $D(\theta)$ is a function of θ and thus has a posterior distribution like any other quantity; it is calculated automatically by both BUGS and JAGS provided that `DIC=TRUE` (which it is, by default)

In-sample vs. out-of-sample prediction

- As you presumably know from earlier modeling courses, however, deviance measures only the *in-sample* accuracy of the model, and complex models will always fit the observed data better than simple models
- That does not mean, however, that complex models are always better – their estimates can be highly variable (frequentist viewpoint)/their posterior distributions can highly diffuse (Bayesian viewpoint)
- The true test of a model is *out-of-sample* prediction – how well the model can predict future observations – and simple models often outperform complex models in this regard

External validation and cross-validation

- In principle, one could use a portion of the data to fit a model, then go out and collect more data to evaluate the predictive ability of the model
- In reality, however, data is usually quite precious and we would like to use all of it to fit the model
- Another approach is *cross-validation*, in which one fits a model leaving some of the data out, predicts the left-out observations, and repeats the process so that each observation gets a turn being left out and predicted

AIC

- This can easily become computationally intensive, especially with complicated models
- For this reason, there has been considerable interest in both Bayesian and frequentist circles to working out numerical approximations to this quantity
- The most well-known frequentist approximation to this quantity is the *Akaike Information Criterion*:

$$AIC = D(\hat{\theta}) + 2p,$$

where p is the degrees of freedom in the model

Degrees of freedom

- In typical settings in which AIC is used, the degrees of freedom is the same as the number of parameters in the model (hence the p notation)
- Directly applying AIC to Bayesian settings, however, is somewhat problematic when informative priors are employed, as the prior restricts the freedom of the parameters
- For example, in our linear regression examples, the models with reference priors and skeptical priors had the same numbers of parameters, but the models were effective not equally "complex" and their posterior distributions were not equally diffuse

Quadratic approximation to the deviance

- Consider a quadratic approximation to the deviance about the posterior mean:

$$D(\boldsymbol{\theta}) \approx D(\bar{\boldsymbol{\theta}}) + D'(\bar{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T D''(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})$$

- Now, with $\mathbf{y}$ (and thus, $\bar{\boldsymbol{\theta}}$) fixed and $\boldsymbol{\theta}$ random, we have:

$$\mathrm{E}D(\boldsymbol{\theta}) \approx D(\bar{\boldsymbol{\theta}}) + \mathrm{tr}\left(\mathbf{V}\boldsymbol{\mathcal{I}}\right),$$

where $\mathbf{V}$ is the posterior variance of $\boldsymbol{\theta}$, $\boldsymbol{\mathcal{I}}$ is the Fisher information evaluated at the posterior mean, and $\mathrm{tr}\left(\mathbf{V}\boldsymbol{\mathcal{I}}\right)$ takes the place of p in the analogous sampling distribution approximation

p_D

- This approximation suggests a measure of the *effective number of parameters* in a model (Spiegelhalter *et al.*, 2002):

$$\begin{aligned} p_D &\equiv \text{tr}\left(\mathbf{V}\boldsymbol{\mathcal{I}}\right) \\ &\approx \bar{D} - D(\bar{\boldsymbol{\theta}}) \end{aligned}$$

- To gain some insight into p_D, let's consider the linear regression case:

$$p_D = \text{tr}\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\Omega}_0 + \boldsymbol{\mathcal{I}})^{-1}\right],$$

or, roughly, the number of coefficients to be estimated times the fraction of the posterior precision that comes from the information

p_D: Another interpretation

- Another way of thinking about p_D is that it represents the difference between the posterior mean deviance and the deviance of the posterior mean
- If the posterior for $\boldsymbol{\theta}$ is relatively concentrated around $\bar{\boldsymbol{\theta}}$, $D(\boldsymbol{\theta})$ will typically be near $D(\boldsymbol{\theta})$ and p_D will be small
- If the posterior for $\boldsymbol{\theta}$ is diffuse, $D(\boldsymbol{\theta})$ might be very large for some values of $\boldsymbol{\theta}$, leading to a large $\bar{D}$ and a large p_D

Example: Regression

- For a more specific example, let's go back to our alcohol metabolism example and consider two models: the likelihoods are the same, but the priors are

$$M_R : \beta_j \sim \mathrm{N}(0, 0.0001^{-1})$$
$$M_S : \beta_j \sim \mathrm{N}\left(0, \{10\mathrm{Var}(\mathbf{x}_j)\}^{-1}\right) \qquad j \neq 0$$

- This produces p_D estimates of

$$M_R : \hat{p}_D = 5.0$$
$$M_S : \hat{p}_D = 3.2$$

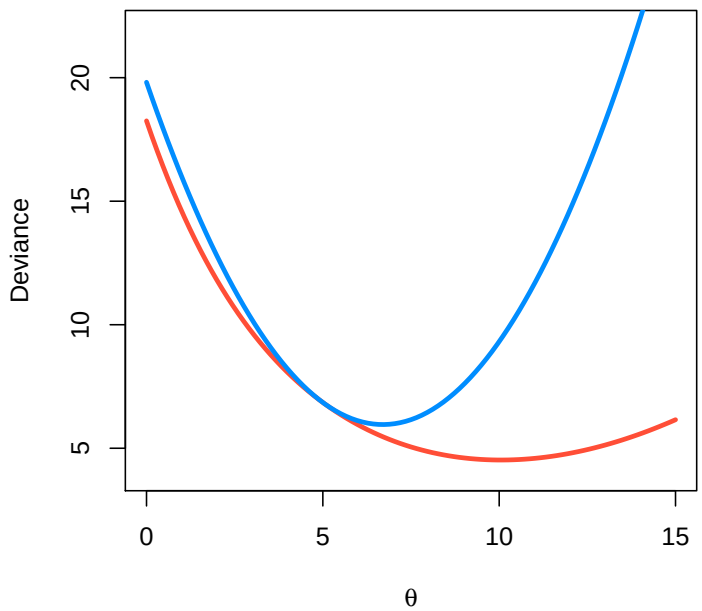
p_D: Caveat

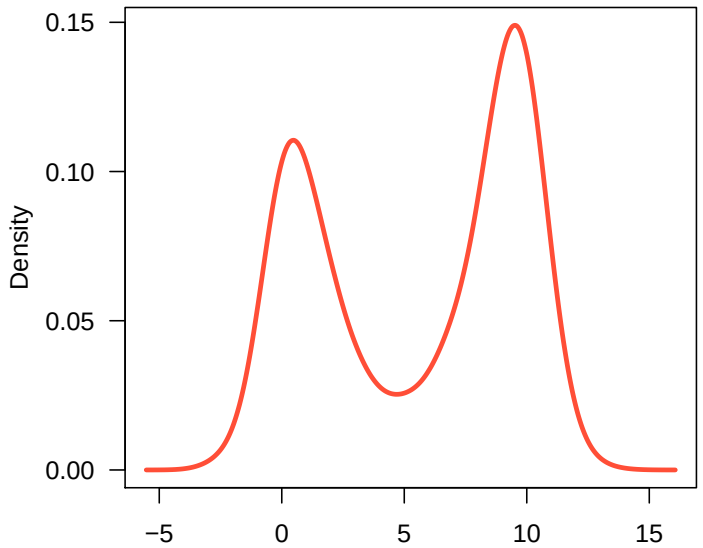
- It should be noted that this p_D estimate is only as good as the quadratic approximation to the deviance about the posterior mean
- As an example of a case where this approximation is poor, suppose we have the following model:

$$Y \sim t_4(\mu, 1)$$
$$\mu \sim t_4(0, 1)$$

and observe $y = 10$

p_D: Caveat (cont'd)





For this example, $\hat{p}_D = -1.1$

p_V

- An alternative estimate of the effective number of parameters in situations with weak priors is to note that when the prior is weak, $\mathbf{V} \approx \boldsymbol{\mathcal{I}}_{\bar{\boldsymbol{\theta}}}$, and thus

$$D(\boldsymbol{\theta}) \approx D(\bar{\boldsymbol{\theta}}) + \chi^2_p$$

- This approximation suggests $\mathrm{E}D(\boldsymbol{\theta}) \approx p$ as before and $\mathrm{Var}\{D(\boldsymbol{\theta})\} \approx 2p$ and thus the following estimator (Gelman *et al.*, 2004):

$$\hat{p}_V = \mathrm{Var}\{D(\boldsymbol{\theta})\}/2$$

p_V (cont'd)

- p_V has certain advantages, such as the fact that it cannot be negative and can always be calculated (at no extra difficulty or cost), even when other approaches cannot (we will see an example of this later)
- However, it is worth noting that it is not a meaningful estimator for the effective number of parameters in cases with informative priors
- For example, in our regression example from earlier,

$$M_R : \hat{p}_V = 5.9$$
$$M_S : \hat{p}_V = 23.5$$

Extracting p_D from BUGS/JAGS

- In both BUGS and JAGS, an estimate of the effective number of parameters is provided by `print(fit)` if using `R2OpenBUGS` or `R2jags`; alternatively, one can go to Inference $\rightarrow$ DIC in the OpenBUGS GUI
- However, it is worth being aware of the fact that p_D is computed by BUGS but not by JAGS; `print(fit)` returns p_V as the degrees of freedom estimate in `R2jags`

p_{opt}

- JAGS has its own approach for calculating degrees of freedom, based on expected values of Kullback-Leibler divergences between multiple MCMC chains (Plummer, 2008)
- The details are beyond the scope of the course, but the concept is the same:

$$\mathrm{E}D(\boldsymbol{\theta}) = \bar{D} + p_{opt},$$

where p_{opt} is a measure of how optimistic $\bar{D}$ is as a measure of the actual out-of-sample deviance $\mathrm{E}D(\boldsymbol{\theta})$

Obtaining p_{opt} in `R2jags`

- One can obtain p_{opt} in `R2jags` through the `dic.samples` function:

```
fit <- jags(...)
dev <- dic.samples(fit$model, n.iter, type="popt")
dev
```

- Note that, unlike the calculation of p_D in BUGS, this (a) requires `n.chains` to be at least 2, and (b) requires a separate MCMC calculation
- For the regression example:

$$M_R : \hat{p}_{opt} = 12.66$$
$$M_S : \hat{p}_{opt} = 10.29$$

DIC

- The measure of fit, $\bar{D}$, may be combined with the measure of model complexity, p_D, to produce the *Deviance Information Criterion*:

$$\text{DIC} = \bar{D} + p_D$$

- Note that

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D;$$

  thus, in cases with weak prior information, where $\bar{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ and $p_D \approx p$, $\text{DIC} \approx \text{AIC}$
- One may define similar criteria for the other model complexity measures:

$$\text{DIC}_V = \bar{D} + p_V$$
$$\text{DIC}_{opt} = \bar{D} + p_{opt}$$

Interpretation of DIC

- It seems fairly clear that the absolute scale of DIC is fairly meaningless, as it depends on factors such as normalizing constants, but is there a meaningful relative scale?
- In other words, suppose $\mathrm{DIC}_1 = 100$ and $\mathrm{DIC}_2 = 105$; is that a meaningful difference?
- For nested models, a rough rule of thumb is that AIC differences less than 2 are insignificant, while AIC differences larger than 10 essentially rule out the model with the larger AIC
- This would then seem to be a reasonable rule of thumb for DIC with nested models as well, but no meaningful rules of thumb have been proposed for non-nested models, or for DIC_{opt} (indeed, it may be that a meaningful rule isn't possible)

Regression example

	$\bar{D}$	p_D	p_{opt}	DIC	DIC_{opt}
Reference	111.4	5.0	12.3	116.4	123.6
Skeptical	122.8	3.2	10.8	125.9	133.7
Interactions	106.7	11.2	35.5	121.8	142.3
Intrx., > 0	110.8	6.2	22.1	117.0	132.9
0-intercept	105.3	4.9	15.1	110.2	120.4

Hills data

	$\bar{D}$	p_D	p_{opt}	DIC	DIC_{opt}
Normal	288.5	4.0	12.4	292.4	300.9
t_5	260.9	4.2	13.4	265.1	274.3
$\nu \propto \nu^{-1}$	249.1	5.5	14.3	254.8	263.4

Roots data

- For the roots data, the usual DIC approach cannot be used, since $\log p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is not clearly defined for mixture distributions
- The DIC_{opt} approach does work in general for mixture distributions, although it fails here, calculating a p_{opt} contribution of `NaN` for all the observations with `Roots = 0`
- The p_V approach is straightforward, however:

	$\bar{D}$	p_V	DIC_V
Poisson	1574	2.0	1576
ZIP	1068	31.4	1099