Model checking and improvement (ch 6)

- Sensitivity analysis
- Sanity check
- Posterior predictive checking
- Model comparison
  - predictive performance
  - DIC
  - Bayes factor

Slide 1

Sensitivity analysis

- How much different assumptions in the model and prior affect the inference?
  - test alternative models and priors
  - alternative models may be changed to a hierarchical model including continuum between the models
    - eg. hierarchical model instead of separate and pooled models
    - eg. t-distribution instead of fixed long-tailed and normal distribution
  - robust models are useful for testing sensitivity to observations
    - eg. t-distribution instead of normal
- Compare results of essential inference
  - eg. extreme quantiles more sensitive than mean or median
  - eg. extrapolation more sensitive than interpolation

Slide 2

Sanity check

- In practice often there is some knowledge, which was not formally included in the model and prior
  - if modeling results are in conflict with other knowledge, formal assumptions need to be reformulated
  - eg, if in bioassay posterior would indicate possibility that tested chemical would reduce the chance of death

Slide 3

External validation

- Compare model predictions to new observations
  - best approach
  - commonly used approach in science generally
  - if possible, predict something which has not been measured before
    - cf eg. predictions of travel of light bending due to sun made by Einstein
  - most obvious weaknesses of the model can be detected without external validation

Slide 4

Partial validation

- Simplest approximation of the external validation
  - part of the observations is left out when computing the posterior
  - compare posterior predictions to left out observations
  - pros/cons
    + simple
    + relatively robust
    - only part of the observations used to update posterior

Slide 5

Posterior predictive checking

- Internal validation
- Is the model internally consistent?
  - samples from the posterior predictive distribution should resemble the original data
  - get samples from the posterior predictive distribution and compare them to the data
    · systematic deviations indicate faults in the model
  - emulates external validation using the observed data also as new observations
    · using the data twice is a problem
    · can reveal some problems anyway
- Gelman *et al.* posterior predictive checking is pragmatic approach, not formal

Slide 6

Posterior predictive checking - example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma^2)$
  - prior $(\mu, \log \sigma) \propto 1$
- Esim9_2.m

Slide 7

Posterior predictive checking

- Data y
- Parameters θ
- Replicated data y^{rep}
  - assume, the observed data is generated by a process, which is well described by the model M having parameters θ
  - replicated data could be observed, if test were repeated
  - replace "true" generating process with the model and parameters θ

$$p(y^{\mathrm{rep}}|y, M) = \int p(y^{\mathrm{rep}}|\theta, M)p(\theta|y, M)d\theta$$

- Test quantity or discrepancy measure $T(y, \theta)$
  - summary statistic which used to compare data and predictive samples

Slide 8

Posterior predictive checking

- Posterior predictive p-value

$$\begin{aligned} p &= \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y) \\ &= \int\int I_{T(y^{\text{rep}},\theta)\geq T(y,\theta)} p(y^{\text{rep}}|\theta)p(\theta|y)dy^{\text{rep}}d\theta \end{aligned}$$

  where I is indicator function

Slide 9

  - is samples $(y^{\text{rep}\,l}, \theta^l)$ from the posterior predictive distribution, this can be estimated by the proportion of samples for which

$$T(y^{\text{rep}\,l}, \theta^l) \geq T(y, \theta^l), \quad l = 1, \ldots, L$$

- Posterior predictive p-value (ppp-value) estimates whether discrepancy between the model and the data could be chance given model assumptions

Posterior predictive checking - example

- Independence in binomial tests
  - model $y \sim \text{N}(\mu, \sigma^2)$
  - prior $(\mu, \log\sigma) \propto 1$
- Observations in order: 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
- $T =$ number of switches in the sequence

Slide 10

  - observed $T(y) = 3$
  - if observations were independent, which would be the distribution of the number of switches if experiment would be repeated?
- Esim9_3.m

Posterior predictive checking

- Selecting test quantity
  - testing properties which directly correspond to parameters in the model should not be tested, because parameters have been fitted to the data
  - test properties, which not directly parametrized
  - different test may have large differences
- Esim9_4.m

Slide 11

Posterior predictive checking

- Omnibus tested
  - χ^2-discrepancy

$$T(y,\theta) = \sum_i \frac{(y_i - E(Y_i|\theta))^2}{\mathrm{var}(y_i|\theta)}$$

  - Deviance

$$T(y,\theta) = -2\log p(y|\theta) = -2\sum_i \log p(y_i|\theta)$$

    where -2 due to historical reasons and properties of normal distribution
- Neither of these indicate any problem in the speed of light experiment since there are only two problematic observations
  - more sensitive choice is $\max_i(-2\log p(y_i|\theta))$
- Esim9_5.m

Slide 12

Posterior predictive checking

- Interpretation of the posterior predictive p-values
  - p-value almost 0 or 1 indicate that model does not model well this property of the data
  - if p-value almost 0 or 1, exact value not interesting
  - goal is not to evaluate whether the model is "true", but estimate discrepancies between the model and data, and estimate whether the observed discrepancies might be arise due to a chance

Slide 13

Posterior predictive checking

- Multiple comparisons
  - if multiple test quantities are used, the probability that one of the test quantities produce p-value near 0 or 1 increase
  - there are multiple hypothesis testing correction terms, but
    - Gelman *et al:* goal is not to formally accept or reject the model, but understand its limitations in realistic applications
    - since dependencies between test quantities are unknown, correction term is difficult to estimate...

Slide 14

Posterior predictive checking

- Very small or big p-value
  - how model can be improved?
  - eg. in light of speed, t-model or mixture model, with separate component for deviating observations
- Moderate p-value
  - how model could be checked in some other way?
  - eg. in SAT example pooled model can not be rejected based on the data, but hierarchical model changes the essential inference

Slide 15

Other conflict measurements*

- In addition of checking predictive distributions, in hierarchical models, it possible to analyse distributions in other levels, too
  - eg. O'Hagan (2003). HSSS model criticism, in Green et al, eds, *Highly Structured Stochastic Systems*, pp. 423-444. Oxford University Press.

Slide 16

Model improvement

- Model construction often iterative
  1. Make a model
  2. Model checking and sensitivity analysis
     - if defects go to step 3
     - if no defect go to step 4
  3. Improve model and go to step 2

Slide 17

  4. Make predictions and decisions

Iterative model improvement

- Original prior information I
- C denote new information obtained in model checking
- Improved model $M^{'}$ and posterior conditioned on I and C

$$p(\theta|y, M^{'}, C, I)$$

Slide 18

but since C based on model checking (ie given y), y has been used twice

- In theory iterative model construction is not right in Bayesian framework
- In practice often no harm done, although i may difficult to estimate how much double use of data has effect
  - eg. if improved model has much larger marginal likelihood, there would be no difference if we had started with both models and would integrate over them

Marginal likelihood

Slide 19

- Bayes' rule

$$p(\theta|y, M) = \frac{p(y|\theta, M)p(\theta|M)}{p(y|M)}$$

where

$$p(y|M) = \int p(y|\theta, M)p(\theta|M)d\theta$$

  - $p(y|M)$ is normalization term
  - $p(y|M)$ also called marginal likelihood, from which the parameters have been integrated out

Posterior probability of the model

Slide 20

- What if alternative models M_1 and M_2?
  - and background information I, which includes the prior probabilities for the models
- We could compute posterior probabilities

$$p(M_j|y, I) = \frac{p(y|M_j, I)p(M_j|I)}{p(y|I)}$$

where $p(y|M_j, I)$ is the marginal likelihood of the model M_j (evidence)

- If we compare the posterior probabilities of two models, normalization term $p(y|I)$ cancels out (I included implicitly)

$$\frac{p(M_2|y)}{p(M_1|y)} = \frac{p(y|M_2)}{p(y|M_1)}\frac{p(M_2)}{p(M_1)}$$

Posterior probabilities and Bayes factor

- Proportion of the posterior probabilities

$$\frac{p(M_2|y)}{p(M_1|y)} = \frac{p(y|M_2)}{p(y|M_1)}\frac{p(M_2)}{p(M_1)}$$

- $p(M_2)/p(M_1)$ from the prior and $p(y|M_2)/p(y|M_1)$ from the data through likelihoods

Slide 21

- If prior probabilities assumed $p(M_1) \approx p(M_2)$ then we have term, called Bayes factor

$$\frac{p(y|M_2)}{p(y|M_1)} = \mathrm{BF}(M_2; M_1)$$

- Bayes factor is the ratio of marginal likelihoods (evidences)

Posterior probabilities of models

- Often when Bayes-factor examined, the posterior probabilities of the models also mentioned
- Nice idea that we could have the probability of a model or an hypothesis, but...

Slide 22

Posterior probabilities of models

- Does $p(M_1|y)$ mean probability, that M_1 is true?
  - oops, we forgot I?
- Does $p(M_1|y, I)$ mean probability, that M_1 is true given I?
  - oops, forgot normalization $p(y|I) = \sum_{M_j \in \mathcal{M}} p(y|M_j, I)p(M_j|I)$?
  - how many models in a model space $\mathcal{M}$?

Slide 23

  - if $p(y|I)$ not computed, posterior probability unknown
    - ratios of posterior probabilities known

Posterior probabilities of models

- Compare to posterior distribution of the parameters
  - we could add parameter M, which gets values $j = 1, 2, \ldots$
  - how should parameters be handled in the Bayesian approach?
- Integrate over the unknowns!
  - if uncertainty about the model, integrate over M, ie. over alternative models
  - asymptotic frequency properties similarly

Slide 24

    - asymptotic consistency of posterior and counter-examples
- Asymptotic consistency of the Bayes factor follows from the asymptotic consistency of the posterior
  - posterior converges to one model, which posterior probability $\longrightarrow 1$
  - who has infinite amount of data?

Posterior probabilities and Bayes factor

- Posterior probabilities of the models can be used to integrate over the alternative models (*Bayesian model averaging (BMA)*)
  - BMA is no different to integrating over the parameters (although it has own name)
  - if possible expanding to continuous model family is preferred
  - eg. in SAT example possible to integrate over separate and pooled model, but hierarchical model includes them as special cases and continuum between them

Slide 25

Posterior probabilities and Bayes factor

- Posterior probabilities of the models have been used for model selection
  - corresponds to maximizing the marginal posterior probability
  - works well if only a few parameters or no parameters at all
  - more there are parameters worse the Bayes factor works

Slide 26

Posterior probabilities and Bayes factor

Slide 27

- Prior sensitivity of the Bayes factor is due to integration over the prior distribution

$$p(y|M_1) = \int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1$$

  - if $p(\theta_1|M_1)$ is not proper, BF is not defined
  - even if $p(\theta_1|M_1)$ is proper, BF still sensitive to prior
  - problem gets worse when the dimensionality of the θ increases
  - problem gets worse when the prior information is vague

- Prior sensitivity of the Bayes factor can be seen also from the chain rule

$$p(y|M_1) = p(y_1|M_1)p(y_2|y_1, M_1), \ldots, p(y_n|y_1, \ldots, y_{n-1}, M_1)$$

  where the first terms in product are sensitive to prior

- If lot of data compared to model complexity, Bayes factor may work ok
  - asymptotic property

Computing Bayes factor

Slide 28

- Computing evidence often difficult

$$p(y|M_1) = \int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1$$

- Some alternatives
  - analytic solution only for closed form posteriors
  - normal approximation
  - variational methods
  - expectation propagation
  - many Monte Carlo -methods

- Estimating evidence with MCMC much more difficult than obtaining samples from the posteriors distribution
  - currently most used approaches are trans-dimensional MCMC (eg. RJMCMC) or importance sampling

Why estimating Bayes factor is difficult with MCMC

- Computing the evidence often difficult

$$p(y|M_1) = \int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1$$

- Direct Monte Carlo-approximation would be

Slide 29

$$p(y|M_1) \approx \frac{1}{L}\sum_{l=1}^{L} p(y|\theta_1^{(l)}, M_1)$$

where $\theta_1^{(l)}$ from the prior distribution

- often vague prior
- if even moderate amount of data, likelihood is concentrated on much smaller region

$\longrightarrow$ very small proportion of samples hist the interesting region

Bayes factor

- Despite its shortcomings Bayes factor is much used in model selection
- Usage is however decreasing
  - problems more obvious with complex models
  - predictive approaches work better

Slide 30

Bayes factor and BMA

- Algorithms designed for computing Bayes factor, can be used to integrate over the model space
- In BMA prior sensitivity is usually lesser problem
- parameters depend on the model structure, and thus uncertainty on parameters and model structure is a prior dependent

  $\longrightarrow$ if uncertainty about model structure, there is need to think more about priors for parameters

Slide 31

Utility of the model

- When model is checked, we might ask how good the model is?
  - before model used in practice
  - can estimate expected utility, ie. how much benefit obtained by using the model for predictions
  - expected predictive performance

Slide 32

Bayesian decision making

- Expected utility $E[U(x)|d] = \int U(x)p(x|d)dx$

Slide 33

Expected utility of the model

- Application specific utility functions very useful
  - eg. money, life years, etc.
- If general goodness of the predictive distribution is interesting, or scientific inference, or application specific utility not yet known, then useful utility function is predictive log-density
  - ie. log-density of the predictive distribution

Slide 34

$$\log p(\tilde{y}|y, M)$$

this has also information theoretic justification

Expected utility of a model

- Often predictive distribution replaced with a *plug-in* estimate

$$\log p(\tilde{y}|\hat{\theta}(y), M),$$

  where $\hat{\theta}(y)$ eg. posterior mean
  - makes computations easier, but not Bayesian

Slide 35

- Sometime log-density multiplied by -2, and then called deviance

$$D(y, \theta) = -2 \log p(y|\theta, M)$$

Expected utility of a model and external validation

- True utility can be find out by using the model
  - external validation
- Expected utility of the model
  - estimates the result from the external validation

Slide 36

Expected utility of a model and y^{rep}

- Expected deviance (eq 6.11 in book)

$$D_{\text{avg}}^{\text{pred}}(y) = \mathrm{E}[D(y^{\text{rep}}, \hat{\theta}(y))],$$

  where expectation over the distribution of y^{rep}
  - assume that distribution of y^{rep} is the "true" data generating distribution

Slide 37

- In external validation y^{rep} replaced with future observations
- Several ways to approximate the expectation

Estimates of the expected utility of the model

- Data predictive
  - y^{rep} same as y, ie y used twice
  - corresponds to "training error" in machine learning
- Partial predictive
  - divide data in two parts
  - y^{rep} is the part not used in posterior computations
  - corresponds to "test error" in machine learning

Slide 38

- Cross-validation predictive *
  - improvement to partial predictive
- DIC
  - estimate correction term to data predictive
- Other predictive approaches *

Effective number of parameters

- In data estimate y^{rep} same as y
  - additionally if using plug-in deviance

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

  - data used twice and estimate over-optimistic
  - optimism is due to fitting parameters to the observed data

Slide 39

  - amount of data fitting can be estimated by estimating the effective number of parameters

Effective number of parameters

- Effective number of parameters is due to
  - total number of parameters (there is also models with infinite number of parameters)
  - prior effect
  - dependency between parameters
  - number of observations $(p_{\text{eff}} \leq n)$

Slide 40

  - amount of uncertainty and complexity of the underlying phenomenon
  - ie how much the parameters have been fitted to the observed data

Information criterion*

- Akaike (1973) derived how to estimate the "expected utility" in combination of frequentist and information theoretic framework
- Akaike summarized the criterion

$$*\mathrm{IC} = \mathrm{fit} + \mathrm{complexity}$$

  - which unfortunately has lead to situation, where the expected utility estimate is forgotten

Slide 41

- In AIC model complexity is the total number of parameters p
  - uses maximum likelihood and assumes $n \longrightarrow \infty$
  - assumes, that θ_0 is in the parameter space

Deviance Information Criterion (DIC)

- Proposal for Bayesian information criterion
- DIC uses deviance

$$D(y, \theta|M) = -2 \log p(y|\theta, M)$$

- If posterior distribution of $p(\theta|y)$ is close to normal ,
  then distribution of $(D - D_{\min})$ is close to χ^2_ν-distribution

Slide 42

- For some models it can be shown that when $n \longrightarrow \infty$, then $\nu \longrightarrow p$,
  where p is the total number of parameters
- For some models it can be shown that when $n \longrightarrow \infty$, then $\nu \longrightarrow p_{\mathrm{eff}}$,
  where p_{eff} is the effective number of parameters

Deviance Information Criterion (DIC)

- DIC approximates the expected utility
- Effective number of parameters measures the optimism due to double use of data
- *fit*-part is (plug-in) deviance $D(y, \mathrm{E}_\theta[\theta])$
- *complexity*-part is $2p_{\mathrm{eff}}$

Slide 43

- DIC

$$\mathrm{DIC} = D(y, \mathrm{E}_\theta[\theta]) + 2p_{\mathrm{eff}}$$

Deviance Information Criterion (DIC)

- Properties of χ^2_ν-distribution (book p. 575)
  $E[\theta] = \nu$ and $\mathrm{Var}[\theta] = 2\nu$
- Effective number of parameters in two ways
  - mean of transferred χ^2_ν-distribution
    $p^{(1)}_{\mathrm{eff}} = \mathrm{E}_\theta[D(y, \theta)] - D(y, \mathrm{E}_\theta[\theta])$
  - variance of transferred χ^2_ν-distribution
    $p^{(2)}_{\mathrm{eff}} = \frac{1}{2}\mathrm{Var}[D(y, \theta)|y] = \frac{1}{2}\frac{1}{L-1}\sum_{l=1}^{L}\left(D(y, \theta^l) - \mathrm{E}_\theta[D(y, \theta)]\right)^2$

Slide 44

- These estimates have slightly different properties
  - $p^{(1)}_{\mathrm{eff}}$ depends on parametrization (due to the term $D(y, \mathrm{E}_\theta[\theta])$)
  - in $p^{(2)}_{\mathrm{eff}}$ estimation of variance can be sensitive, since χ^2_ν-distribution only asymptotically, and true distribution might have long tail
  - some other differences, too, since χ^2_ν only asymptotically
  - non consensus, which one is better

Deviance Information Criterion (DIC)

- DIC can also be presented in a form

$$\mathrm{DIC} = \mathrm{E}_\theta[D(y,\theta)] + p_{\mathrm{eff}}$$

  which useful, specially if $p_{\mathrm{eff}}^{(2)}$ used, since then DIC independent of the parametrization

Slide 45

- On the other hand, form in the book is obtained by replacing p_{eff} with $p_{\mathrm{eff}}^{(1)} = \mathrm{E}_\theta[D(y,\theta)] - D(y, \mathrm{E}_\theta[\theta])$

$$\mathrm{DIC} = 2\,\mathrm{E}_\theta[D(y,\theta)] - D(y, \mathrm{E}_\theta[\theta])$$

Deviance Information Criterion (DIC)

- DIC is quick and easy to compute given posterior samples
  - popular due to simplicity of computation
  - available eg. in WinBUGS-software
  - useful approximation, as long as limitations remembered

Slide 46

Deviance Information Criterion (DIC)

- Problems in DIC
  - use of point-estimate in predictive distribution underestimated the uncertainty
  - asymptotic assumptions often do not hold
    - posterior may be far from normal
    - observations may be dependent
  - result can be sensitive to parametrization, and thus requires case-by-case analysis

Slide 47

  - estimating the uncertainty in the expected utility estimate is not straight forward
  - can estimate p_{eff} to be negative
  - can be bad estimate in regression for even small extrapolation
  - justification not completely Bayesian

Deviance Information Criterion (DIC)

- How large difference in DIC is significant?
- Can be approximate crudely (cf. Bayes factor)

$$\begin{aligned} \mathrm{DF}(M_1, M_2) &= \exp((\mathrm{DIC}_2 - \mathrm{DIC}_1)/2) \\ p(\mathrm{DIC}_1 < \mathrm{DIC}_2) &\approx \mathrm{DF}/(1+\mathrm{DF}) \end{aligned}$$

  - ie. difference is significant if difference is larger than 6

Slide 48

- Spiegelhalter writes in DIC FAQ
  - difference larger than 10 is significant
  - difference smaller than 5 is not significant

Model selection

Slide 49

- True Bayesian way is to integrate over all uncertainties
  - no need for model selection
  - after some iteration in model building we can't find defects anymore and are happy with the model
- In practice we may need to reduce the model
  - nested models
  - need to make model easier to explain
  - need to educe measurement cost for covariates
  - need to reduce computation time
- In practice, integration over the models is sometimes difficult, and then we can ignore models, which predictive performance is clearly inferior
  - non-nested

Model comparison and selection

Slide 50

- SAT example
- Model choices
  - separate model
    - knowing results in 7 schools would not affect the estimate on the 8th school
  - pooled model
    - all coaching equally good
  - hierarchical model
    - differences between schools, but common population prior

Model comparison and selection

- SAT example
- What is the goal of the inference?
  - are we predicting for these 8 schools?
    - y^{rep} for these schools
  - are we predicting for other schools?
    - y^{rep} for new schools

Slide 51

- DIC corresponds to a situation, where y^{rep} for these schools

Deviance Information Criterion (DIC)

- SAT example

Model	$\bar{D}(\mathrm{E}_\theta[\theta])$	p_{eff}	DIC
separate ($\tau = \infty$)	54.6	7.8	70.4
pooled ($\tau = 0$)	59.3	1.0	61.3
hierarchical	57.4	2.8	62.9

Slide 52

- Is there significant differences between models?
  - separate model is different (difference>6)
  - no significant difference between pooled and hierarchical model

Summary

- Full (BMA) model after checking is the best choice
- Predictive performance can be estimated
  - partial validation robust, if lot of data
  - DIC fast and easy
  - others (e.g. cross-validation not in this course)

Slide 53

- Based on predictive performance estimate th application expert can say whether model might be useful in practice at all
- Full model can be reduced and predictive performance of the reduced models can be estimated
  - need to make model easier to explain
  - need to educe measurement cost for covariates
  - need to reduce computation time