Empirical Software Engineering

Write your answers directly on these pages (there's always a risk that loose papers disappear)—use the back also if needed. I'll be at the written exam twice (first time after approximately one hour).

On November 10 at 13.30 you are welcome to Richard's office (4th floor in the Jupiter building at Campus Lindholmen) to complain about the grading. Before you come you *must* send Richard an email clearly pointing out where you think the error is, what you wrote, and why you believe the grading was not correct. If I don't receive such an email before 13.30 on November 10, then I will not meet with you.

There are no perfect models, but some models I'd take out for dinner.

Grade 3: 45 points; 50%
Grade 4: 63 points; 70%
Grade 5: 81 points; 90%
Maximum: 90 points

Question 1 :

(**3p**) What is **underfitting** and **overfitting** and why does it lead to poor prediction?

Write a **short example** (perhaps as a model specification?) and **discuss** what would be a possible condition for underfitting and overfitting.

Solution: One straightforward way to to discuss this is to show two simple models, one where we only have a general intercept α (complete pooling), and one where you estimate an intercept for each cluster in the data (varying intercept), but where we don't share information between the clusters (think of the pond example in the coursebook), i.e., no pooling.

This question can be answered in many different ways.

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha \\
\alpha &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

above complete pooling, below no pooling

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{CLUSTER}[j]} \\
\alpha_j &\sim \text{Normal}(0, 2) \qquad \text{for} j = 1, \ldots, n
\end{aligned}
$$

Question 2 :

(**3p**) What is **epistemological justification** and how does it differ from **ontological justification**, when we design models and pick likelihoods? Please provide **an example** where you **argue** epistemological and ontological reasons for selecting a likelihood.

Solution: Epistemological, rooted in information theory and the concept of maximum entropy distributions. Ontological, the way nature works, i.e., a physical assumption about the world.

One example could be the Normal() likelihood for real (continuous) numbers where we have additive noise in the process (ontological), and where we know that in such cases Normal is the maximum entropy distribution.

Question 3 :

(**2p**) What are the **two types** of predictive checks? **Why** do we have to do these predictive checks?

Solution: Prior predictive checks: To check we have sane priors.

Posterior predictive checks: To see if we have correctly captured the regular features of the data.

Question 4 :

(**6p**) When diagnosing Markov chains, we often look at **three** diagnostics to form an opinion of how well things have gone. Which three diagnostics do we commonly use? What do we **look for** (i.e, what thresholds or signs do we look for)? Finally, **what do they tell us**?

Solution:

Effective sample size: A measure of the efficiency of the chains. 10% of total sample size (after warmup) or at least 200 for each parameter.

$\widehat{R}$: A measure of the between and the within variance of our chains. Indicates if we have reached a stationary posterior probability distribution. Should approach 1.0 and be below 1.01.

Traceplots: Check if the chains are mixing well. Are they converging around the same parameter space?

Question 5 :

(**3p**) Quadratic approximation and grid-approximation works very well in many cases, however in the end we fall back on Hamiltonian Monte Carlo (HMC) at many times. **Why** do we use HMC, i.e., what **are the limitations** of the other approaches?

Additionally, **how** does HMC work conceptually, i.e., describe with your own words?

Solution:

Well, `quap` and other techniques assume a Normal likelihood. Additionally, they don't see levels in a model (they simply do some hill climbing to find the maximum a posterior estimate), i.e., when a prior is itself a function of parameters, there are two levels of uncertainty.

HMC simulates a physical system, i.e., the hockey puck, which it flips around the log-posterior (i.e., the valleys are the interesting things, and not the hills as in the case of `quap`).

	WAIC	pWAIC	dWAIC	weight	SE	dSE
m1	361.9	3.8	0.0	1	14.26	NA
m2	401.8	2.6	40.9	0	11.28	10.48
m3	405.9	1.6	44.0	0	11.66	12.23

Question 6 :

(**4p**) As a result of comparing three models, we get the above output. What does each column (WAIC, pWAIC, dWAIC, weight, SE, and dSE) **mean**? **Which** model would you **select** based on the output?

Solution:

WAIC: information criteria (the lower the better).

pWAIC: effective number of parameters in the model.

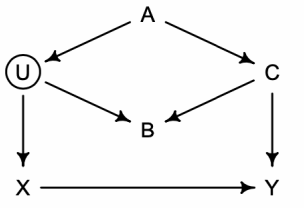
dWAIC: the difference in WAIC between the models.

weight: an approximate way to indicate which model WAIC prefers.

SE: standard error. How much noise are we not capturing.

dSE: the difference in SE between the models.

If you have approximately 4–6 times larger dWAIC than dSE you have fairly strong indications there is a 'significant' difference. Here $\mathcal{M}_1$ is, relatively speaking, the best.



Question 7 :

(**6p**) The DAG above contains an exposure of interest X, an outcome of interest Y, an unobserved variable U, and three observed covariates (A, B, and C). Analyze the DAG.

a) Write down **the paths** connecting X to Y (not counting $X \rightarrow Y$)?

b) Which must be **closed**?

c) Which variable(s) should you **condition** on?

d) What are these **constructs** called in the DAG universe?

- $U \rightarrow B \leftarrow C$
- $U \leftarrow A \rightarrow C$

Solution: First is, $X \leftarrow U \leftarrow A \rightarrow C \rightarrow Y$, and second is, $X \leftarrow U \rightarrow B \leftarrow C \rightarrow Y$.

These are both backdoor paths that could confound inference. Now ask which of these paths is open. If a backdoor path is open, then we must close it. If a backdoor path is closed already, then we must not accidentally open it and create a confound.

Consider the first path, passing through A. This path is open, because there is no collider within it. There is just a fork at the top and two pipes, one on each side. Information will flow through this path, confounding $X \rightarrow Y$. It is a backdoor. To shut this backdoor, we need to condition on one of its variables. We can't condition on U, since it is unobserved. That leaves A or C. Either will shut the backdoor, but conditioning on C is the better idea, from the perspective of efficiency, since it could also help with the precision of the estimate of $X \rightarrow Y$ (we'll give you points for conditioning on A also).

Now consider the second path, passing through B. This path does contain a collider, $U \rightarrow B \leftarrow C$. It is therefore already closed. If we do condition on B, it will open the path, creating a confound. Then our inference about $X \rightarrow Y$ will change, but without the DAG, we won't know whether that change is helping us or rather misleading us. The fact that including a variable changes the $X \rightarrow Y$ coefficient does not always mean that the coefficient is better now. You could have just conditioned on a collider!

First construct is a collider. If you condition on B you open up the path. The second construct is a confounder (fork). If you condition on A you close the path.

Question 8 :

(**8p**) Conducting Bayesian data analysis requires us to do **several things** for us to trust the results. Can you **describe the workflow** and what **the purpose is with each step**?

Solution: Start with a null model, do prior checks, check diagnostics, do posterior checks, then conduct inferential statistics if needed. Also, it would be good if you mention **comparisons** of models and that it is an **iterative** approach.

Question 9 :

(**5p**) To set correct insurance premiums, a car insurance firm is analyzing past data from several drivers. An example dataset is shown below:

Driver	Age	Experience	Num. incidents
1	45	High	1
2	29	Low	3
3	26	Medium	1
4	50	Medium	0
⋮	⋮	⋮	⋮

With this data, the firm wants to use Bayesian data analysis to **predict the number of accidents** that are likely to happen given the **age** of the driver and their **experience** level (low, medium, and high).

- Write down a **mathematical model** definition for this prediction using any variable names and priors of your choice.
- State the **ontological** and **epistemological** reasons for your likelihood.

Remember to **clearly state and justify** the choices and assumptions regarding your model.

Solution:

A math model (following the math notation with sub $_i$ sprinkled over the model) should be given.

They'll hopefully go for a Poisson or Negative-Binomial/Gamma-Poisson.

Question 10 :

(**4p**) What is the **purpose** and **limitation** of using field study and field experiment as a research strategy?

Solution: Field studies: Specific and real-world setting (natural settings).

Unobtrusiveness: Researcher does not change/control parameters or variables

Goal: Understand phenomena in concrete and realistic settings. Explore and generate new hypothesis.

Low statistical generalizability (claim analytical generalizability!)

Low precision of the measurement of the behavior

If the data is collected in a field setting that does NOT necessarily imply that a study is a field study. Examples of research methods: Case study, Ethnographic study, Observational study. Mostly qualitative but may include quantitative data.

Field experiment: Specific and real-world setting (natural settings).

Obtrusiveness: Researcher **manipulates** some variables/properties to observe an effect. Researcher does not control the setting.

Goal: **Investigate**/**evaluate**/**compare** techniques in concrete and realistic settings.

Low statistical generalizability (claim analytical generalizability!)

Low precision of the measurement of the behavior (confounding factors)

Not the same as an experimental study. Changes are made and results observed. But does not have control over the natural setting. Correlation is observed. But not causation, generally speaking. Examples of research methods: Evaluative case study, quasi-experiment, action research. Both qualitative and quantitative data.

Question 11 :

(**6p**) Read parts of the paper appended to this exam (you should know which parts to focus on and **not read everything**!) and answer the questions,

- Which of the eight research strategies presented in the ABC framework does this paper fit? Justify and argue!
- Can you present an overview of the research method employed?
- Can you argue the main validity threats of the paper?
  - It would be very good if you can list threats in the four common categories we usually work with in software engineering, i.e., internal, external, construct, and conclusion validity threats.

Solution: The students can argue for several research strategies since the authors of the paper could've been more explicit...

Once they've argued for one or a few strategies, they need to explain what the strategy they picked does.

Finally, they should think about threats to this study (we've covered internal, external, conclusion, and construct validity threats for different research strategies/approaches).

Question 12 :

(**20p**) A common research method in software engineering that is used to complement other methods is survey research. Here follows a number of questions connected to survey research:

1. A survey can be supervised or unsupervised. What's the difference? (2p)
2. Surveys can either intervene at one point, or repeatedly over time. What are these two types of surveys usually called? (1p)
   (a) When we do intervene repeatedly over time we often see two common study designs being used (tip: it has to do with the sample you use). (1p)
3. Discuss the weaknesses and strengths with convenience sampling. (2p)
4. We often differ between reliability and validity concerning surveys,
   (a) What is the difference between the reliability and validity in survey design? (2p)
   (b) Name and describe at least two types of reliability in survey design. (3p)
   (c) Name and describe at least two types of validity in survey design. (3p)
5. Even if you measure and estimate reliability and validity you still want to *evaluate* the survey instrument. Which are the two (2) common ways of evaluating a survey instrument? Explain their differences. (3p)
6. Likert scale questions are common in survey designs. If you want to use such a variable as a predictor in your statistical model then you need to set a suitable prior, e.g., not a Normal prior. Which one should you preferably use and why? (3p)

Solution: 1. Supervised: Assigning a survey researcher to one or a group of respondents. Unsupervised: E.g., mailed questionnaire.

2. Cross-sectional vs. longitudinal

2a. Cohort and panel studies

3. Strength: It's convenient and we usually get decent sample sizes. Weaknesses: It's not random so confounders will be lurking.

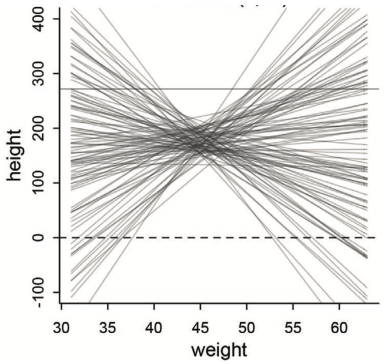
4a. Reliability: A survey is reliable if we administer it many times and get roughly the same distribution of results each time. Validity: How well the survey instrument measures what it is supposed to measure?

4b. e.g., Test-retest (intra-observer), alternate form reliability, internal consistency, inter-observer (Krippendorff's α).

4c. e.g., content validity, criterion validity, construct validity.

5. **Focus groups:** Assemble a group of people representing respondents, experts, beneficiaries of survey results. Ask them to complete the survey questionnaire. The group identifies missing or unnecessary questions, ambiguous questions and instructions. **Pilot studies:** Uses the same procedures as planned for the main study but with smaller sample. Targets not only problems with questions but also response rate and reliability assessment.

6. A Dirichlet(α) prior. The Dirichlet distribution is the multivariate extension of the Beta distribution. The Dirichlet is a distribution for probabilities, values between zero and one that all sum to one. The beta is a distribution for two probabilities. The Dirichlet is a distribution for any number. This way we model the probability separately for each category in our Likert question and we do not assume a Gaussian distribution.



Question 13 :

(**7p**) **Finish** the following statements,

1. Crap! I got 15 divergent transitions, now I probably need to...
2. No! I have $\widehat{R} > 1.6$, now I need to...
3. My posterior predictive check shows that I don't capture the regular features of the data. I probably need to...
4. My traceplots look like a mess. They are all mixed up and look like hairy caterpillars. I guess that...
5. My DAG has a pipe between X (intervention) and Y (outcome). I should probably...
6. The posterior is proportional to the ...
7. (see figure above) Hmm... my prior predictive check looks a bit funky. I probably need to...

Solution:

1. ...reparameterize the model or be more careful with PriPC
2. ...run the chains for more iterations
3. ...take into account more aspects (dispersion?) by extending the model
4. ...it's ok!
5. ...**not** condition on the thing in the middle, i.e., the 'pipe'
6. ...prior times the likelihood
7. ...tighten my prior [redo prior predictive check]

Question 14 :

(**3p**) **Rewrite** the following model as a **multilevel** model.

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{GROUP}[i]} + \beta x_i \\
\alpha_{\text{GROUP}} &\sim \text{Normal}(0, 1.5) \\
\beta &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$

Solution: All that is really required to convert the model to a multilevel model is to take the prior for the vector of intercepts, α_{GROUP}, and make it adaptive.

This means we define parameters for its mean and standard deviation. Then we assign these two new parameters their own priors, *hyperpriors*. This is what it could look like,

$$
\begin{aligned}
y_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{GROUP}[i]} + \beta x_i \\
\alpha_{\text{GROUP}} &\sim \text{Normal}(\bar{\alpha}, \sigma_\alpha) \\
\beta &\sim \text{Normal}(0, 1) \\
\bar{\alpha} &\sim \text{Normal}(0, 1.5) \\
\sigma_\alpha &\sim \text{Exponential}(1)
\end{aligned}
$$

The exact hyperpriors you assign don't matter here. Since this problem has no data context, it isn't really possible to say what sensible priors would be.

Question 15 :

(10p or even −10p) Let us finally end with some multiple choice questions! You get one point if you answer a question correctly (one or more choices can be correct!) You get −1 points if you fail to answer the question correctly (if you don't answer you get 0 points). Simply put circle(s) around what you think is/are the correct answer(s)!

Q1: What is/are the premises for epistemological justifications when we design our models?

a) Information theory b) Maximum entropy c) None of the answers

Q2: Adding predictors and parameters to a model can have which of the following impact(s)?

a) Complex models b) Models that overfit c) Improved estimates

Q3: Why do we use posterior predictive checks?

a) Check that model captures regular features of data b) Check that our priors are sane c) Check that we have a perfect fit d) Check the effect priors have on outcome

Q4: Which of the following strategies are in a non-empirical setting?

a) Sample studies b) Computer simulations c) Field studies d) Laboratory experiments

Q5: In the model definition below, which line is the likelihood?

$$\mu \sim \text{Normal}(0, 10) \quad (1)$$
$$y_i \sim \text{Normal}(\mu_i, \sigma) \quad (2)$$
$$\sigma \sim \text{Exponential}(1) \quad (3)$$

a) Line 1 b) Line 2 c) Line 3 d) None of the answers

Q6: In the model definition below, how many parameters are in the posterior distribution?

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim \text{Normal}(0, 10)$$
$$\beta \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Exponential}(2)$$

a) 1 b) 2 c) 3 d) 4

Q7: Which mechanisms can make multiple regression produce false inferences about causal effects?

a) Confounding b) Multi-collinearity c) Not conditioning on post-treatment variables

Q8: Which of the below desiderata are key concerning information entropy?

a) The measure of uncertainty should not be continuous
b) Measure of uncertainty should increase as the number of possible events increases
c) The measure of uncertainty should be multiplicative

Q9: What is the difference between an ordered categorical variable (OCV) and an unordered categorical variable (UCV)?

a) For UCV we often see 'Likert' questions
b) For OCV the distances between the values are not necessarily the same
c) For UCV the values merely represent different discrete outcomes
d) None of the answers

Q10: Which of the following priors will produce more *shrinkage* in the estimates?

a) $\alpha_{\text{TANK}} \sim \mathsf{Normal}(0, 1)$
b) $\alpha_{\text{TANK}} \sim \mathsf{Normal}(0, 2)$
c) None of the answers

Certum est.

Solution:

1. a & b
2. a & b & c
3. a
4. b
5. b
6. c
7. a & b
8. b
9. b & c
10. a