CHALMERS

EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn			
DIT246	Empirical Software Engineering			
Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DIT246-0001-UXA		2023-08-22	15	5

* I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr		Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	X	8	
2	X	4	
3	X	10	
4	X	3	
5	X	4	
6	X	4	
7	X	4	
8	X	5	
9	X	4	
10	X	1,5	
11	X	6	
12	X	2,5	
13	X	11	
14	X	4	
15	X	3	
16			
17			
Bonus poäng			
Total examination points Summa poäng		74!	

[redacted]

DIT246-0001-UXA

Empirical Software Engineering

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

On September 11 at 10.15 you are welcome to Richard's office (4th floor in the Jupiter building at Campus Lindholmen) to complain about the grading. Before the meeting 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 10.15 on September 11, then I will not meet with you.

> Repetition is the mother of all knowledge.
>
> — Richard Torkar

Grade 3: 42 points; ~50%
Grade 4: 57 points; ~70%
Grade 5: 74 points; ~90%
Maximum: 82 points

DIT246-0001-UXA

Question 1 :

(**8p**) Over the years Bayesian data analysis has evolved and spread to all natural sciences. There are several reasons for this, e.g., a principled way to incorporate prior knowledge, increase in computational power making Markov chain Monte Carlo a viable option for sampling, and probabilistic programming languages gaining ground. However, not until lately have statisticians developed guidelines researchers can follow to systematically design Bayesian models.

In *your* opinion, what key steps are compulsory when conducting Bayesian data analysis? Please explain each step one takes when designing models so that we can place *some* confidence in the results.

You can either draw a flowchart and explain each step, or write a numbered list explaining each step.

(It's ok to write on the backside, if they haven't printed on the backside again. . .)

✓ 1. Clearly understand & define theoretical estimand. Build a DAG!

✓ 2. Design generative model to output synthetic data matching theory.

✓ 3. Start with template (null) model and (gradually) build in theory.

✓ 4. Perform prior predictive checks, to verify priors are sensible w.r.t. theory.

✓ 5. Run model with synthetic data given by 2. Analyze that model captures features of synthetic data.

✓ 6. Run with real data, look at diagnostics (e.g. effective samples, divergent transitions for MCMC). Analyze.

✓ 7. Repeat steps from 3 to make alternative models.

✓ 8. Compare models. If models are used to predict, compare out-of-sample predictive power, e.g. WAIC or PSIS.

8

DIT246-0001-UXA

Question 2 :

(**4p**) Underfitting and overfitting are two concepts not always stressed a lot with black-box machine learning approaches. In this course, however, you've probably heard me talk about these concepts a hundred times…

What happens when you underfit and overfit, i.e., **what would the results be**? What are some principled **ways to deal with** under- and overfitting?

Consider modelling with categories using no pooling:

$$R_i \sim \text{Normal}(\mu, \sigma)$$
$$\mu_i = \alpha$$
$$\alpha \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

this assumes all categories are the same. Under fits if false (likely).

Now with complete pooling:

$$R_i \sim \text{Normal}(\mu, \sigma)$$
$$\mu_i = \alpha_{category[i]}$$
$$\alpha_{category} \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

this assumes categories are fully independent. Overfits if false (likely).

Solve: Partial pooling: $\alpha_{category} \sim \text{Normal}(\bar{\alpha}, \tau)$
$\bar{\alpha} \sim \ldots$
$\tau \sim \ldots$

Also very wide priors can overfit by being too excitable about extremes And very tight priors can underfit by being too skeptical.

4

DIT246-0001-UXA

Question 3 :

(**10p**) To understand **how team size affects psychological safety** the following data was collected:

Team	Team size	SPI	Psychological safety
1	5	67%	High
2	15	33%	Low
3	11	49%	Low
4	7	90%	High
⋮	⋮	⋮	⋮

The experiment started with assuming that planning effectiveness and psychological safety has a very strong association. For planning effectiveness they used schedule performance indicator (SPI) as a stand-in variable.

Schedule Performance Indicator = (Completed points / Planned points)

Based on the result, if the SPI is more than 50% they are classified as a team with high psychological safety. If less than 50% they are classified as a team with low psychological safety.

With the above data, the firm wants to use your knowledge to understand the association between team size and psychological safety.

Write down the 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.

SPI_i : $SPI \in [0,1]$ for team i. X_i: Team size for team i

$$SPI_i \sim \text{Normal}(\mu, \sigma) \quad (1)$$
$$\mu_i = \alpha + \beta X_i$$
$$\alpha \sim \text{Uniform}(0, 1) \quad (2)$$
$$\beta \sim \text{Normal}(0, 0{,}15) \quad (3)$$

(1) Turning continuous variable (0..1) into binary discards a silly amount of information, treating team with 0% and 50% score as the same but 50% and 51% as distinctly opposite. Thus SPI. Ontologically I justify likelihood by most teams likely performing similar, with fewer teams at more extreme SPI/size. Thus, Normal. Epistemologically I justify it with Gaussian being a maximum entropy distribution. → ✓

10p

(2) Here, a prior restricting SPI to {0,1] (e.g. uniform), especially with lots of data seems sensible. Maybe Normal is better: it will give some invalid values but, encompasses the idea of SPI's at the extremes being more unlikely (due to how we design tests.) Initially, I chose Normal (0.5, 0.2) for this reason, but then changed my mind to avoid model being able to predict nonsense, e.g. -10%, or 120%..

(3) Allow for negative and positive association for β. wide prior (+/- 30% with one extra/less team member considered plausible.) I chose it wide because I have no real domain knowledge

Pooling? chose no pooling because I assume:

1. We are looking for average.
2. Teams will only contribute exactly 1 sample.
3. Not interested (per question story) in predicting for individual teams.

DIT246-0001-VXA

Question 4 :
(**3p**) What is **epistemological justification** and how does it differ from **ontological justification**, when we design models and choose likelihoods? Please provide **an example** where you **argue** epistemological and ontological reasons for selecting a likelihood.

✓ Ontological: This is what we experience/observe.

✓ Epistemological: Information theory, maximum entropy.

E.g.: Investigate start of some animal's mating season's association with "available food".

✓ Ontologically we argue this phenomenon is Normally distributed. If there is a causal link we expect most individuals to behave similarily with some noise/deviation.

Epistemologically we argue Gaussian is a maximum entropy Exponential distribution, well equipped to let the data "speak".

fair enough, but think once again about the outcome I underlined above

3

DIT246-0001-UXA

Question 5 :

(**4p**) When diagnosing Markov chains, we often look at several diagnostics to form an opinion of how well things have gone. Name four diagnostics we commonly use? What do we **look for** in each diagnostics (i.e, what thresholds or signs do we look for)? Finally, **what do they tell us**?

$\hat{R}$ (R-hat): Looking for convergence. Want to see value approaching 1. Run more/~~debug~~ if >1.01. ✓

Traceplot: Observe how chains explore, want to see "fuzzy caterpillars", i.e. they all converge. ✓

N-eff Effective number of samples. If far less than actual samples: (e.g. 10%), then samples are auto correlated. ✓

Divergent Transitions: Transitions that were dropped. Want this to be 0. ✓ If muliple, consider if there is some "narrow valley" in probability space.

4

DIT246-0007-UXA

Question 6 :

(**4p**) Explain the four main benefits of using multilevel models.

1. Better estimates if we repeat ✓ sample individuals*

2. Better estimates if we have imbalanced ✓ number of entries from different individuals* in sample. ✓

3. Estimate variation between categories. ✓

4. Avoid averaging variation. ✓

4

DIT246-0007-VXA

Table 1: Output from running WAIC on three models.

	WAIC	SE	dWAIC	dSE	pWAIC
m1	127.6	14.69	0.0	NA	4.7
m3	129.4	15.10	1.8	0.90	5.9
m2	140.6	11.21	13.1	10.82	3.8

Question 7 :

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

WAIC: Estimated out-of-sample predictive power score. ✓

SE: standard error of WAIC. ✓

dWAIC: Difference of row model's WAIC and "best" model's WAIC. ✓

dSE: standard error of dWAIC. ✓

pWAIC: Number of effective parameters in model. ✓

No model stands out strongly (e.g. no dWAIC 4-6 times larger than dSE). Would report all models & the comparison. ✓ At gunpoint, I'd choose m1.

4

DIT246-0001-UXA

Question 8 :

(**5p**) Write an example mathematical model formula for a Poisson regression model with two different kinds of varying intercepts, also known as a cross-classified model.

$$R_i \sim \text{Poisson}(\lambda)$$

$$\log(\lambda_i) = \alpha_{CAT1[i]} + \gamma_{CAT2[i]}$$

$$\alpha_{CAT1} \sim \text{Normal}(\bar{a}, \sigma)$$

$$\bar{a} \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

$$\gamma_{CAT2} \sim \text{Normal}(0, T)$$

$$T \sim \text{Exponential}(1)$$

5

DIT246-0001-VXA

Question 9 :

(**4p**) Explain the terms in your own words:

- prior
- posterior
- information entropy
- instrumental variable

Consider Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$!

Prior: P(A). what we "believe"* before ✓ update with (new) data.

Posterior: P(A|B). what we "believe"* ✓ after update with data.

* and how "strongly", i.e. distributions an P(B)?

Informational Entropy: $H = -\sum_{i=1}^{n} p_i \log(p_i)$. ✓

Quantification of reduced uncertainty when learning event.

Instrument Variable: Variable acting as ✓ natural experiment on exposure, independent of outcome. Useful to deal with unobserved confound.

4

~~DIT24600~~

DIT246-0001-VXA

Question 10 :

(**2p**) What are the two kinds of varying effects? Explain the effect they have on a statistical model.

Varying Intercept & Varying Slope.

They are used for partial pooling. Learning new data point for category i, updating intercept/slope also affects intercept & slope "belief" for category j, i ≠ j,

1,5

DIT246-0001-JXA

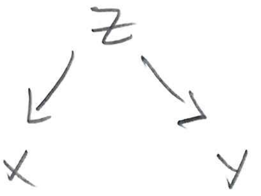
Question 11 :

(**8p**) What are the four elemental confounds on which any Directed Acyclic Graph can be explained?

Please draw the four confounds. Explain what they mean (preferably by explaining if one should condition or not on certain elements).

when estimating Y ~ X:

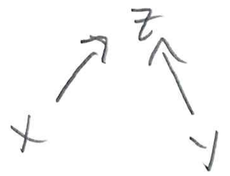
The Fork:



Condition on Z, to remove spurrious association between X & Y. ✓

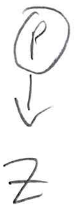
The Pipe: $X \to Z \to Y$. Don't condition on Z, will block path between X & Y. ✓

The Collider:



Don't condition on Z, will make X & Y appear associated. ✓

The Descendent:



✓

If P is a confound, conditioning on Z will run into the same problems but weaker. Especially dangerous if P is unobserved. ✓

6

DIT246-0001 -VXA

Question 12 :

(**4p**) What is the **purpose** and **limitations** of using laboratory experiments and experimental simulations as a research strategy?

Experimental simulation: ✓

Artificial setting. Natural actors.

Laboratory experiment. ✓

Artificial setting. Natural actors.

But short/induced trials.

Diffrence: e.g. greehouse vs test-tube.

Purpose: Control variables.

Limitation: Even if actors are natural their behavior may be affected by setting.

2,5

DIT 246-0001-UXA

Question 13 :

(**11p**) 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. 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)
2. 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)

1. a) Reliability: How similar results will be if performed repeatedly. ✓
Validity: How well instrument test/captures effect. ✓

b) Test-retest: Give test participant test again sometime later & check for positive correlation (0.7-1). ✓
Internal Consistency: Ask multiple differing questions about the "same" thing, and see if they are answered similarly ✓

c) Construct validity: experts review instrument ✓
Criterion validity: compare to other instruments ✓

2. Focus groups & Pilot studies. ✓
In focus groups you find a representative set of individuals to test & feedback. ✓
In pilot studies you launch the real thing but with much smaller sample. ✓

11

DIT 246 - 0001 - VXA

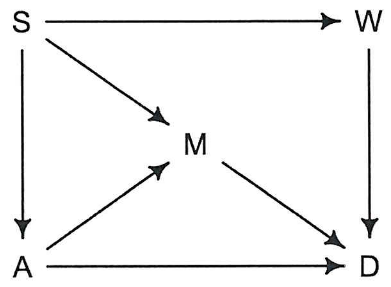


Figure 1: A messy Directed Acyclic Graph.

Question 14 :

(**6p**) Look at the DAG above. We want to estimate the total causal effect of W on D. Which variable(s) should we condition on and why?

Backdoor criterion:

Direct path: W → D

Backdoor paths: D ← A ← S → W,

D ← M ← A ← S → W,

D ← M ← S → W

Adjustment set to close path:

{S} (minimal)

Thus, condition on S. OR {A, M}

4

DIT246-0001-UXA

Question 15 :

(**5p**) You get one point if you answer a question correctly. Simply write your answer below.

1. We should always start the Bayesian data analysis by designing a . . . .
2. Adding predictors to a model can lead to several things. Two common things are . . .
3. My $\widehat{R}$ value is 1.04. I should . . .
4. We can quantify . . . using Kullback-Leibler divergence.
5. I am first and foremost always interested in propagating . . . while doing BDA.

1. Null model (although McElreath might argue to start with clearly defining ✓ theoretical estimand)

TRUE!

2. Overfitting. Introduction of confounds. ✓
3. Run the model more. ✓

3

Certum est.