Summary

- Based on what we have seen so far, at the moment renewal of your contract is more unlikely than likely, but we are giving you a chance to change this.
- The renewal decision will be made in April, that is about 6 months before the end of current contract, so that you will have time for arrangements if the contract will not be renewed. The decision will be heavily influenced by what you can show us until then.
- So far you have had two main tasks:
  1. help with LOO-uncertainty paper,
  2. first author the continuation paper.
- We have allowed you to have side projects, which seems to have been a mistake as you have now been doing many projects (many of them not related to the funding), but none of them demonstrate that you would be able to finish papers or thesis.
- Recently you have made it more clear that you are not motivated to do 2, and we did first discuss with you that you could do something slightly different. As we're now running out of time, we need to choose a topic that is familiar for you and related to the funding.
- We want you to write a "minipaper" that
  - showcases you can write a coherent "story" that makes sense from beginning till end,
  - quality- and structurewise should be like a paper,
  - from the amount of content need not be like a paper,
  - can have theory (proofs), but we highly recommend you focus on simulations as you already have done similar simulations and it seems like you can get stuck with the proofs,
  - ideally forms the basis for an actual paper.
- It does not matter if the results themselves are not interesting enough to be included in a paper.
- Before hiring you, we also discussed that you need to improve your English skills. This minipaper doesn't need to be perfect, but it should show that you have improved your English.
- You can add content to the outline after we discuss this and if we agree that it makes sense.
- If you have any questions or problems you can always contact me.

Comments

- Please be careful with the notation!
- Please be careful with the definitions!
- Please avoid unnecessary "Definition", "Lemma" or "Theorem" blocks!
- Please be careful to appropriately reference sources!
- Please be careful when you make assertions!
- Please explain everything as simply and explicitly as possible to avoid ambiguity!
- Please be very careful with mathematical proofs!
- Comprehensively describe experimental settings!

Overarching goal

- In LOO uncertainty paper, we say that if $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y){<}4$, then $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ can be underestimated and there may be significant skewness ignored by normal approximation, but we also say that then the models have still small difference.
  - demonstrate with simulation the maximum loss in predictive performance and parameter point estimation if either model is selected when $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y){<}4$,
  - possible sketch of the corresponding analytic solution,
  - the same simulations will also show that 4 is reasonable threshold,
  - analytic justification can be obtained by examining the KL-divergence from the true distribution to either model A or model B predictive distribution.
- Sometimes users also compute LOO-weights or LOO-BB-weights and get different result than with $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$. As analysis of LOO-BB-weights is more complicated, we simplify by looking at LOO-SE-weights in two model case (note that LOO-SE-weights in the stacking paper are wrong).
  - add LOO-weights and LOO-SE-weights to the simulation,
  - as SE can be underestimated, explain analytically how that underestimation affects $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$ and LOO-SE-weights, and how LOO-weights correspond to $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y){=}0$,
  - explain what special there is in LOO-weights and LOO-SE-weights if $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y){<}4$.

Based on the results should we recommend $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$ or LOO-SE/BB-weights? I think $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ and $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ together tell more than either $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$ or LOO-SE/BB-weight alone, and would assume examining $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ and $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$ is a good recommendation.

Empirically and (only if possible) theoretically validate and extend LOO recommendations that we can give users.

FOR NOW EVERYTHING ONLY FOR TWO MODELS

Recommendations should include what to do if

- $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ is small and
  - $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ is small or
  - $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ is large
- $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ is large and
  - $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ is small or
  - $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$ is large

and should include explanations why different methods may give different results.

You should compare

- in the setting of two (nested) models
- using simulations and theory (only if immediate!)
- different methods to weigh/rank/select models
  - $\widehat{P}(\mathrm{elpd}(M_A, M_B|y) > 0)$,
  - LOO-weights,
  - LOO-SE weights and
  - LOO-BB weights
- for different model sets (reuse your experiments from github!),
- and using different metrics
  - true expected log pointwise predictive density $\mathrm{elpd}(M_A|y)$ and
  - root mean square error (RMSE) of the parameter point estimate.

Please correctly introduce (with references where appropriate) and explain (as much as needed)

- true expected log pointwise predictive density $\mathrm{elpd}(M_A|y)$,
- pairwise elpd difference $\mathrm{elpd}(M_A, M_B|y)$,

- LOO-estimate of the elpd $\widehat{\text{elpd}}_{\text{LOO}}(M_A|y)$,
- LOO-estimate of the elpd difference $\widehat{\text{elpd}}_{\text{LOO}}(M_A, M_B|y)$,
- standard error estimate of the LOO-estimate of the elpd difference $\widehat{\text{SE}}_{\text{LOO}}(M_A, M_B|y)$,
- estimated probability of true elpd_diff > 0 $\widehat{P}(\text{elpd}(M_A, M_B|y) > 0)$,
- LOO-weights,
- LOO-SE weights,
- LOO-BB weights,
- differences and connections between the different weights/rankings
- reasons why different methods give different rankings

and anything else which will be used / needed.

As discussed previously, if you notice any pattern that you believe could hold more generally, write this down as a conjecture! There is no need to prove them, especially if this may take an excessive amount of time!

Practical LOO for model comparison

Introduction

Methods

- Models:
  - Experiments from your github (these could be enough)
- Conjectures (simulations reveal a pattern which we might believe holds generally)
  - Unless proof is immediate, skip it

Results

- $\text{elpd}(M_A, M_B|y) < 4$: Models are similar. What does this imply? ($\text{elpd}(M_A|y)$ and RMSE)

What are the cases when the difference between two models is small anyways, even if e.g. $\text{P}(\text{elpd}(M_A|y) > 0)$ is small.

- No stacking
- $\widehat{\text{elpd}}_{\text{LOO}}(M_A|y)$ ranking vs $\text{sv\_elpd}_{\text{LOO}}(M_A, M_B|y)$ weights/ranking
- if something is defined in the loo-uncertainty draft you can use it

- Validate and extend LOO recommendations
    - Questions and answers provided by Aki
- You don't need to invent anything
- If you can find additional patterns/connections that is good but not necessary.
- SNR + horsehoe + detection threshold (Juho Piironen)

Discussion

TBD

Sources with topics

Authors in parenthesis are not necessarily originators of ideas

"To select or not to select" (Aki)

WILL BE REMOVED FOR ASAEL

Compare

- using
    - simulations or
    - theory where possible
- for non-zero but potentially small model difference β
- different methods to combine a given set of model candidates, including using weights from
    - BMA,
    - BMA+,
    - stacking,
    - model selection using different criteria, e.g. use the smaller model
        - always,
        - if BF < delta (**maybe**),
        - if ...
        - never,
- for different model candidate sets, including
    - y ∼ 1 vs y ∼ x with
        - wide prior on model difference or

        - (R)HS prior on model difference
    - y ∼ x1+x2+x3+x4+x5 vs y ∼ x1+x2+x3+x4+x5+x6
    - y ∼ x vs y ∼ s(x)
    - y ∼ x vs y ∼ x + (xg)
    - y ∼ x with
        - family=normal vs family=t or
        - family=poisson vs family=negbin
    - more than two model candidates (**later**), including
        - y ∼ 1 vs y ∼ x1 vs y ∼ x2 vs y ∼ x1 + x2 with correlating x1 and x2,
        - y ∼ x1 vs y ∼ x2 vs y ∼ x1 + x2 vs y ∼ x1 + x2 + x1*x2 with an interaction term which correlates with main effects,
    - other models as e.g. in [Sivula20],
- using different metrics, including
    - (loss of) predictive accuracy as measured by the true expected log pointwise predictive density elpd($M_A|y$),
    - root mean square error (RMSE) of parameter estimates and/or other metrics
- visualized with (x,y,color) corresponding to e.g.
    - (beta, metric, method) and more.

Personal communcations (Aki)

- Plots (row, col, x, y, color):
    - (N/A, $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$, $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$, weight, method)
    - (N/A, metric (elpd($M_A|y$)/rmse), beta, metric, method)

“Using uncertainty for model comparison” (Asael)

- estimated probability of true elpd_diff > delta $P(\mathrm{elpd}(M_A, M_B|y) > \delta)$
    - student-t approach
- show that w_a < w_a+ for w_a < 1/2 if w_a+ from normal approximation (**don’t include**)
    - what if w_a+ from BB? (**don’t include**)

“Practical recommendations for considering the uncertainty in Bayesian model comparison with leave-one-out cross-validation” (Tuomas, Mans, Aki)

Introduction

- when can LOO model comparison be trusted?
- small number of models
- contrast LOO with BMA
- use sv_elpd$_{\mathrm{LOO}}(M_A, M_B|y)$ weights

Practical recommendations

- Recommendations to assess whether LOO estimates are reliable
- theory and experiments => recommended thresholds
- $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ assumed to be exactly computed
- $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y) < 4$:
  - LOO estimates likely to have bias and/or high variance/skew
  - LOO can provide no reliable assessment
- $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y) > 4$:
  - assess diagnostics (k_hat, PPC, LOO-PIT) and sample size
  - if diagnostics for better model are fine, it’s probably safe to pick (bad diagnostics usually lead to overoptimistic $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A|y)$ estimates)

Bayesian model averaging

- Introduce PBMA as an approximation to BMA and expression for weights
- Introduce PBMA+ ([Yao18])
- Introduce sv_elpd$_{\mathrm{LOO}}(M_A, M_B|y)$ weights and $P(\mathrm{elpd}(M_A, M_B|y) > \delta)$ (with $\delta = 0$)

Connection to BMA

- Quality of exposition degrades

Analysis of LOO-BB

- Discussion of plots (row, col, x, y, color):
  - (beta, n, w_a, w_a+ (BB), point density)
  - (beta, n, $P(\mathrm{elpd}(M_A, M_B|y) > \delta)$, w_a+ (BB), point density)

    - (beta, n, w_a, w_a+ (BB), point density)

“practical loo for model comparison” (Oriol, Osvaldo)

- How to select models?
    - no SBC, “just” simulations
    - effect of noisy data
    - how often do we pick which model as a function of
        - effect size,
        - sample size and more,
    - evaluate/compare behavior of using elpd($M_A|y$)/BMA/stacking:
        - is one method always superior/inferior?
        - does this depend on the goal?
        - “error” of choosing
            - the more complex model,
            - the model with best elpd($M_A|y$),
            - model based on BF,
            - weights using BMA,
        - evaluate “selection performance”
            - can a hard threshold be defended? (unlikely)
        - evaluate predictive performance
        - when to use CV/predictive methods for model comparison?
            - m-open/-closed/-complete
            - examples when LOO works or does not work,
            - rule of thumb?
        - LOO diagnostics in practice?
            - bad k_hats?
        - LOO vs LOGO vs k-fold?
        - discuss (briefly) how LOO compares to BF
        - other scoring rules?

“loo subworkflow” (Oriol, Osvaldo)

- PPC to discard grossly misspecified models (**don’t include**)
- Sometimes CV is not needed. When, when not? (**don’t include**)
- Large k-hat values? (**don’t include**)
- Model expansion (Poisson=>negative binomial, Gaussian=>student t, pooledunpooled=>hierarchical)
- When to choose simpler (special case of bigger) model? (**don’t include**)
- Should LOO only be used for small number of models with clear difference? (**don’t include**)
- SBC for $\widehat{\text{elpd}}_{\text{LOO}}(M_A|y)$ (**don’t include any of this**)
  - Investigate impact of k-hat distribution on reliability of rankings
  - elpd$(M_A, M_B|y)$ rule of thumb? (e.g. elpd$(M_A, M_B|y) > 4$)
  - LOO vs LOGO vs k-fold
- LOO (**don’t include any of this**)
  - Pitfalls/limits? How to fix/circumvent?:
    - Sample size?
    - Non-robust models?
    - BF estimates?
  - Strengths
    - MCMC draws variation has little impact
    - built-in failure diagnostics
    - tool for model exploration
  - How do k-hats change when model complexity increases?
  - Plots (col,x,y,color):
    - color scale, elpd_loo_i, elpd_psisloo_i, k_hat_i
    - color scale, elpd_psis_loo_i, ml_smc(?), k_hat_i
    - y, k_hat_i, elpd_psis_loo_i or elpd_loo_i, None