Summary

- Based on what we have seen so far, renewal of your contract is more unlikely than likely.
- The renewal decision will be made in April heavily influenced by what you can show us until then.
- 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.
- 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 be careful with appropriately referencing sources!
- Please be very careful with mathematical proofs!
- Comprehensively describe experimental settings!

Practical LOO for model comparison

NOTATION MAY HAVE TO BE ADAPTED!

FOR NOW EVERYTHING ONLY FOR TWO MODELS

Introduction

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

- 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 $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A|y)$,
- LOO-estimate of the elpd difference $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$,
- standard error estimate of the LOO-estimate of the elpd difference $\widehat{\mathrm{SE}}_{\mathrm{LOO}}(M_A, M_B|y)$,
- random variable modeling the true elpd difference using LOO estimates $\mathrm{sv\_elpd}_{\mathrm{LOO}}(M_A, M_B|y)$ including
  - the normal ansatz and
  - the Bayesian Bootstrap ansatz,
- stacking (of predictive distributions, see [Yao18] and [Yao21]),
- results from [Sivula20], [Yao18] and [Yao21],
- connections between $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ and $\mathrm{sv\_elpd}_{\mathrm{LOO}}(M_A, M_B|y)$ weights,
- reasons why $\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(M_A, M_B|y)$ ranking can differ from $\mathrm{sv\_elpd}_{\mathrm{LOO}}(M_A, M_B|y)$ ranking,

and anything else which will be used / needed.

Methods

TBD

Results

TBD

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 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)

- probability of elpd_diff > delta p_δ
    - 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 $\widehat{\text{elpd}}_{\text{LOO}}(M_A|y)$ estimates)

Bayesian model averaging

- Introduce PBMA as an approximation to BMA and expression for weights
- Introduce PBMA+ ([Yao18])
- Introduce sv_elpd$_{\text{LOO}}(M_A, M_B|y)$ weights and p_δ (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_δ, 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