Comparing Bayesian post-estimation variable selection methods

projection predictive inference versus spike-and-slab priors

Kim (Zhipei) Wang

Introduction

- We live in a data-rich era, but prediction challenges remain:
  - **Noisy data**
  - **Poor generalization**
  - **High-dimensional datasets**

Less is more

- This leads to a critical problem in methodology: **Variable Selection**.
- My thesis compares:
  - **Spike-and-Slab Variable Selection (SSVS)** George and McCulloch (1993); Mitchell and Beauchamp (1988).
  - **Projection Predictive Variable Selection (PPVS)** Piironen, Paasiniemi, and Vehtari (2020).

Spike-and-Slab Variable Selection (SSVS)

- SSVS assigns a **probability** for each variable:
  - Probabilty of being included or discarded.
- Widely used in Bayesian statistics.
- Serves as a benchmark in my study.

What is Projection Predictive Variable Selection (PPVS)?

- PPVS takes a **different approach** to variable selection:

1. Fit a comprehensive **reference model**.
2. Identify a minimal **submodel** that retains predictive accuracy.
3. **Project** the reference model onto the submodel.

plot(cvvs)

How is PPVS Implemented?

- Use the `projpred` R package (Piironen et al. 2023):

1. Fit the **reference model** using standard Bayesian libraries.
2. Use `cv_varsel()` to build the **solution path**:
   - Adds variables one by one.
3. Retrieve solution path using `solution_terms()`.
4. Use `suggest_size()` to decide submodel size.
5. Retrieve the submodel and compute the **projected distribution**.

alternatively

Example workflow in r code chuncks:

Simulation Study:

Design (Bainter et al. 2023):

Variable	Levels	Values
Sample size	2	100, 400
Predictors (10 true effects)	1	50
Regression coeff. (β)	3	{0.1, 0.3, 0.5}
Correlation (σ)	3	{0, 0.4, 0.8}
True effect pattern	2	{mixed, clustered}

Metrics:

True and false inclusion rates.

What to Expect?

- PPVS is expected to be better for **high-dimensional data**:
  - Expected to outperform SSVS in these settings.
- In lower-dimensional scenarios (50 predictors, n = 100 or 400):
  - Performance differences may be minimal.

Future Work:

- Extend simulations to **higher dimensions**.
- Explore **complex scenarios** to find where PPVS excels.

Conclusion

- My thesis compares two Bayesian variable selection methods:
  - **SSVS:** Established and widely used.
  - **PPVS:** Promising, efficient for high dimensions.
- Simulation studies will reveal:
  - Strengths and weaknesses of PPVS across conditions.
  - Practical guidance for researchers using Bayesian methods.

Citations

Bainter, Sierra A., Thomas G. McCauley, Mahmoud M. Fahmy, Zachary T. Goodman, Lauren B. Kupis, and J. Sunil Rao. 2023. "Comparing Bayesian Variable Selection to Lasso Approaches for Applications in Psychology." *Psychometrika* 88 (3): 1032–55. https://doi.org/10.1007/s11336-023-09914-9.

George, Edward I., and Robert E. McCulloch. 1993. "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association* 88 (423): 881–89. https://doi.org/10.1080/01621459.1993.10476353.

Mitchell, T. J., and J. J. Beauchamp. 1988. "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association* 83 (404): 1023–32. https://doi.org/10.1080/01621459.1988.10478694.

Piironen, Juho, Markus Paasiniemi, Alejandro Catalina, Frank Weber, and Aki Vehtari. 2023. "Projpred: Projection Predictive Feature Selection." https://mc-stan.org/projpred/.

Piironen, Juho, Markus Paasiniemi, and Aki Vehtari. 2020. "Projective Inference in High-Dimensional Problems: Prediction and Feature Selection." *Electronic Journal of Statistics* 14 (1). https://doi.org/10.1214/20-EJS1711.