Bayesian Stacking in Multilevel Models

Mingya Huang David Kaplan

University of Wisconsin–Madison

IMPS 2023

Model Uncertainty

"Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."[1]

- Structural uncertainty: functional form f and the data D.
- Parametric uncertainty: parameter of interest θ.
- Blackbox methods.

[1] Hoeting, et al., 1999

Bayesian Model Averaging (BMA)

For multiple candidate models $\mathcal{M} = (M_1, \ldots, M_K)$, the posterior probability of the interested quantity Δ given observed response y can be derived as:

$$p(\Delta \mid y) = \sum_{k=1}^{K} p\left(\Delta \mid M_k, y\right) p\left(M_k \mid y\right)$$

$$p\left(M_k \mid y\right) = \frac{p\left(y \mid M_k\right) p\left(M_k\right)}{\sum_{k=1}^{K} p\left(y \mid M_k\right) p\left(M_k\right)}$$

Bayes Theorem

$$p\left(y \mid M_k\right) = \int p\left(y \mid \theta_k, M_k\right) p\left(\theta_k \mid M_k\right) d\theta_k.$$

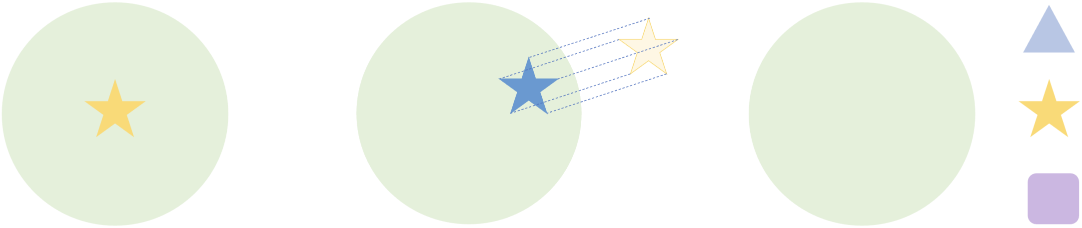
Drawbacks: BMA assumes $\mathcal{M}$-closed setting.

$\mathcal{M}$-Framework

- **$\mathcal{M}$-closed**: the true data generating model is within the candidate model list. That is, the true model M_t is one of the $M_k \in \mathcal{M}$.

- **$\mathcal{M}$-complete**: the true model exists but is not within the model list.

- **$\mathcal{M}$-open**: not only the true model M_t is not in $\mathcal{M}$ but also the explicit form $p(\tilde{y} \mid y) = p\left(\tilde{y} \mid M_t, y\right)$ cannot be specified.



(a) M-closed (b) M-complete (c) M-Open

Bayesian Stacking (BS)

- **BMA: Weighted average over candidate models**

$$p(\tilde{y} \mid \tilde{x}, \mathrm{w}, \text{ model averaging}) = \sum_{k=1}^{K} w_k p\left(\tilde{y} \mid \tilde{x}, M_k\right), \quad \mathrm{w} \in \mathcal{S}_K$$

- **BS: Leave-one-out crossed validation (LOO-CV):**

$$p_{k,-i} = \int_{\Theta_k} p\left(y_i \mid \theta_k, x_i, M_k\right) p\left(\theta_k \mid M_k, \{(x_{i'}, y_{i'}) : i' \neq i\}\right) d\theta_k$$

$$\hat{\mathrm{w}}^{\text{stacking}} = \arg\max_{\mathrm{w}} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k p_{k,-i}\right), \text{ such that } \mathrm{w} \in \mathcal{S}_K$$

Drawbacks: Bayesian Stacking use input-independent weights.

Bayesian Hierarchical Stacking (BHS)

- **Bayesian Stacking**[2]:

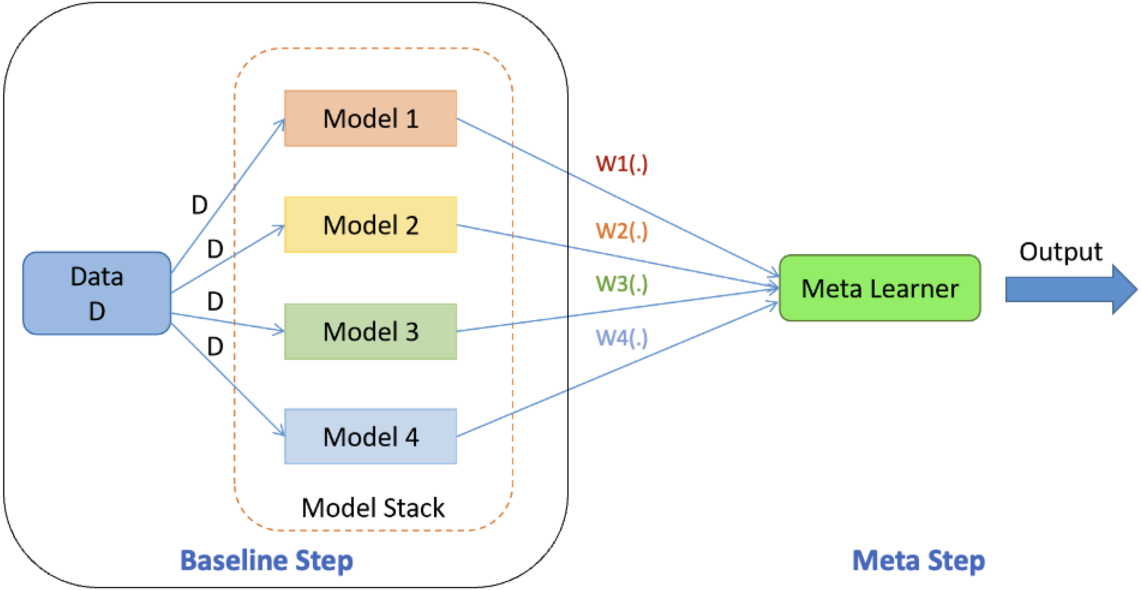
$$\hat{w}^{\text{Stacking}} = \arg\max_{w} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k p_{k,-i} \right), \text{ such that } w \in \mathcal{S}_K$$

- **Bayesian Hierarchical Stacking**:[3]

$$\hat{w}^{\text{Hierarchical Stacking}} = \arg\max_{w} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k (x_i) p_{k,-i} \right) + \log p^{\text{prior}}(w) + \text{ constant}$$

[2]Yao, et al., 2018
[3]Yao, et al., 2022

Bayesian Hierarchical Stacking (BHS)



Bayesian Hierarchical Stacking (BHS)

- **Replace w with $w(.)$:**

$$\log p(\mathrm{w}(\cdot) \mid \mathcal{D}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k \left(x_i\right) p_{k,-i} \right) + \log p^{\text{prior}} (\mathrm{w}) + \text{ constant}, \quad \mathrm{w}(\cdot) \in \mathcal{S}_K \tag{1}$$

- **Softmax transformation for categorical x in J categories:**

$$w_{jk} = \frac{\exp\left(\alpha_{jk}\right)}{\sum_{k=1}^{K} \exp\left(\alpha_{jk}\right)}, 1 \leq k \leq K-1, 1 \leq j \leq J; \quad \alpha_{jK} = 0, 1 \leq j \leq J, \tag{2}$$

- **Partial pooling with priors and hyper-priors:**

$$\alpha_{jk} \mid \mu_k, \sigma_k \sim \mathcal{N}\left(\mu_k, \sigma_k\right), k = 1, \ldots, K-1, j = 1, \ldots, J \tag{3}$$

$$\mu_k \sim \mathcal{N}\left(\mu_0, \tau_\mu\right), \quad \sigma_k \sim \mathcal{N}^+\left(0, \tau_\sigma\right) \tag{4}$$

Simulation Study

1. Based on the parameters and distribution of the PISA 2018 data (OECD), data are generated using a full model with y = reading scores and two random effects.
2. Four reduced models are fitted:

Models	Covariates	Random Effects
Model 1	Female, Escs, Staffshort	Random intercept for Staffshort
Model 2	Joyread, Pisadiff, Staffshort	Random intercept and slope for Staffshort
Model 3	Gfofail, Mastgoal, Adaptivity, Compete, Public	Random intercept for Public
	Metasum, SWBP, Workmast	
Model 4	Perfeed, Belong, Public	Random intercept and slope for Public

Table: Model Specification

3. Weakly informative priors.

Results: Model Weights

Ratio	Sample Sizes (n_i*n_j)	Methods	Model 1	Model 2	Model 3	Model 4
1:1	100 (10*10)	BS	0.169	0.670	0.061	0.101
		BHS	0.222	0.550	0.124	0.104
	4900 (70*70)	BS	0.169	0.824	0.006	0.000
		BHS	0.217	0.702	0.059	0.023
1:5	500 (10*50)	BS	0.202	0.636	0.048	0.115
		BHS	0.235	0.560	0.088	0.117
	4500 (30*150)	BS	0.166	0.820	0.011	0.003
		BHS	0.231	0.669	0.059	0.041
5:1	500 (50*10)	BS	0.150	0.849	0.002	0.000
		BHS	0.195	0.769	0.023	0.013
	4500 (150*30)	BS	0.151	0.847	0.002	0.000
		BHS	0.133	0.813	0.045	0.009

Table: Model weights for each model across different sample sizes

Results: Predictive Performance with Kullback-Leibler (KL) Divergence

Ratio	Sample Sizes (n_i*n_j)	Methods	KLs
1:1	100 (10*10)	BS	0.156
		BHS	0.114
	4900 (70*70)	BS	0.127
		BHS	0.110
1:5	500 (10*50)	BS	0.239
		BHS	0.200
	4500 (30*150)	BS	0.204
		BHS	0.172
5:1	500 (50*10)	BS	0.086
		BHS	0.078
	4500 (150*30)	BS	0.079
		BHS	0.077

Table: Predictive performance of BS and BHS across different sample sizes

Conclusion & Discussion

- BHS outperforms BS in multilevel models, especially when the sample size is small.
- The effects of between-group and within-group variation on prediction are left to investigate.

Conclusion & Discussion

- BHS outperforms BS in multilevel models, especially when the sample size is small.
- The effects of between-group and within-group variation on prediction are left to investigate.

Thank you!

Email: mhuang233@wisc.edu
GitHub: https://github.com/huskyh233/BayesStacking-Multilevel
Website: https://mhuang233.github.io
Preprint: 10.31234/osf.io/e9m6x

Bayesian Hierarchical Stacking (BHS)

1. **Baseline Step:** For a hold out dataset D' to $D = \{y_i, x_i\}_{i=1}^{n}$,

$$p\left(\tilde{y}_i | x_i, M_k, \mathcal{D}'\right) = \int p\left(\tilde{y}_i | x_i, M_k, \theta_k\right) p\left(\theta_k | M_k, \mathcal{D}'\right) d\theta_k \tag{5}$$

2. **Meta Step:** Plug in x_{is} and y_{is}, calculate the probability using LOO: By integrating out the holdout dataset $\mathcal{D}'$, the likelihood can be obtained:

$$p\left(\tilde{y}_i | \mathrm{w}, x_i\right) = \mathbb{E}_{\mathcal{D}'}\left[p\left(\tilde{y}_i | \mathrm{w}, x_i, \mathcal{D}'\right)\right] = \sum_{k=1}^{K} w_k\left(x_i\right) \mathbb{E}_{\mathcal{D}'}\left(p\left(\tilde{y}_i | x_i, M_k, \mathcal{D}'\right)\right) \approx \sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i} \tag{6}$$

$$\log(p(\mathcal{D} | \mathrm{w})) \approx \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i}\right) \tag{7}$$

Continuous and Hybrid Inputs

The weights can be modeled additively to allow for more structure:

$$w_{1:K}(x) = \text{softmax}\left(w^*_{1:K}(x)\right), \tag{8}$$

where

$$w^*_k(x) = \mu_k + \sum_{m=1}^{M} \alpha_{mk} f_m(x), k \leq K-1, w^*_K(x) = 0, \tag{9}$$

and where $\{f_m : \mathcal{X} \to \mathbb{R}\}$ are m distinct features, $w^*_k(x)$ is the combination of the prior mean μ_k, and the additive functions $\alpha_{mk} f_m(x)$. The final joint posterior density will be:

$$\begin{aligned}\log p(\alpha, \mu, \sigma | \mathcal{D}) = &\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i} \right) \\ &+ \sum_{k=1}^{K-1} \sum_{j=1}^{J} \log p^{\text{prior}}\left(\alpha_{jk} | \mu_k, \sigma_k\right) \sum_{k=1}^{K-1} \log p^{\text{hyper prior}}\left(\mu_k, \sigma_k\right)\end{aligned} \tag{10}$$

Model Specification

Full Model:

$$y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_1 Female + \beta_2 escs + \beta_3 joyread + \beta_4 pisadiff + \\ \beta_5 staffshort + \beta_6 metasum + \beta_8 gfofail + \beta_9 mastgoal + \beta_{10} swbp + \beta_{11} workmast + \\ \beta_{12} adaptivity + \beta_{13} compete + \beta_{14} Public + \beta_{15} perfeed, \sigma^2)$$

$$\beta_{00} \mid \gamma_{00}, \gamma_{05,014}, \tau_{00} \sim N(\gamma_{00} + \gamma_{05} staffshort + \gamma_{014} Public, \tau_{00}^2)$$
$$\beta_5 \mid \gamma_{50}, \gamma_{51}, \tau_{50} \sim N(\gamma_{50} + \gamma_{51} staffshort, \tau_{50}^2)$$
$$\beta_{14} \mid \gamma_{140}, \gamma_{141}, \tau_{140} \sim N(\gamma_{140} + \gamma_{141} Public, \tau_{140}^2)$$

Model Specification

Model 1:

$$y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_1 Female + \beta_2 Escs, \sigma^2),$$

$$\beta_{00} \mid \gamma_{00}, \tau_{00} \sim N(\gamma_{00} + \gamma_{01} * staffshort, \tau_{00}^2), \quad \beta_{ps} \mid \gamma_{ps}, \tau_{ps} \sim N(\gamma_{ps}, \tau_{ps}^2), \quad \sigma \sim Cauchy^+(0, 2.5)$$

$$\gamma_{00} \sim N(400, 1)$$
$$\gamma_{ps} \sim N(0, 1)$$
$$\tau_{ps} \sim Cauchy^+(0, 2.5)$$

Model Specification

Model 2:

$$y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_3 joyread + \beta_4 pisadiff + \beta_{50} staffshort, \sigma^2)$$

$$\beta_{00} \mid \gamma_{00}, \tau_{00} \sim N(\gamma_{00} + \gamma_{01} * staffshort, \tau_{00}^2), \quad \beta_{3-4} \mid \gamma_{30-40}, \tau_{30-40} \sim N(\gamma_{30-40}, \tau_{00-40}^2),$$
$$\beta_{50} \mid \gamma_{5s}, \tau_{50} \sim N(\gamma_{50} + \gamma_{51} * staffshort, \tau_{50}^2), \quad \sigma \sim Cauchy^+(0, 2.5)$$

$$\gamma_{00} \sim N(400, 1)$$
$$\gamma_{ps} \sim N(0, 1)$$
$$\tau_{ps} \sim Cauchy^+(0, 2.5)$$

Model Specification

Model 3:

$$y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_7 metasum + \beta_8 gfofail + \beta_9 mastgoal + beta_{10} swbp + \beta_{11} workmast + \beta_{12} adaptivity + \beta_{13} compete + \beta_{14} Public, \sigma^2),$$

$$\beta_{00} \mid \gamma_{00}, \tau_{00} \sim N(\gamma_{00} + \gamma_{01} * Public, \tau_{00}^2), \quad \beta_{ps} \mid \gamma_{ps}, \tau_{ps} \sim N(\gamma_{ps}, \tau_{ps}^2), \quad \sigma \sim Cauchy^+(0, 2.5)$$

$$\gamma_{00} \sim N(400, 1)$$
$$\gamma_{ps} \sim N(0, 1)$$
$$\tau_{ps} \sim Cauchy^+(0, 2.5)$$

Model Specification

Model 4:

$$y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_{15} perfeed + \beta_{16} belong + \beta_{17} public, \sigma^2)$$

$$\beta_{00} \mid \gamma_{00}, \tau_{00} \sim N(\gamma_{00} + \gamma_{01} * Public, \tau_{00}^2), \quad \beta_{15-16} \mid \gamma_{150-160}, \tau_{150-160} \sim N(\gamma_{150-160}, \tau_{150-160}^2),$$
$$\beta_{17} \mid \beta_{17s}, \tau_{170} \sim N(\gamma_{170} + \gamma_{171} Public, \tau_{170}^2), \quad \sigma \sim Cauchy^+(0, 2.5)$$

$$\gamma_{00} \sim N(400, 1)$$
$$\gamma_{ps} \sim N(0, 1)$$
$$\tau_{ps} \sim Cauchy^+(0, 2.5)$$

Analysis Measures

Kullback-Leibler Divergence Score (KL):

$$\mathrm{KLD}(f, g) = \int p(y) \log \left(\frac{p(y)}{g(y \mid \theta)} \right) dy$$