Bayesian Stacking in Multilevel Models

Mingya Huang David Kaplan

University of Wisconsin–Madison

IMPS 2023

Model Uncertainty

George Box:

"All models are wrong, some are useful."

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

Bayesian Model Averaging (BMA)

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

Q: What is Δ?

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

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

(Bayes Rule)

Q: should this be $\tilde{y}$?

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

~~**Drawbacks:** BMA assumes $\mathcal{M}$-closed framework.~~ → Sameer's suggestion

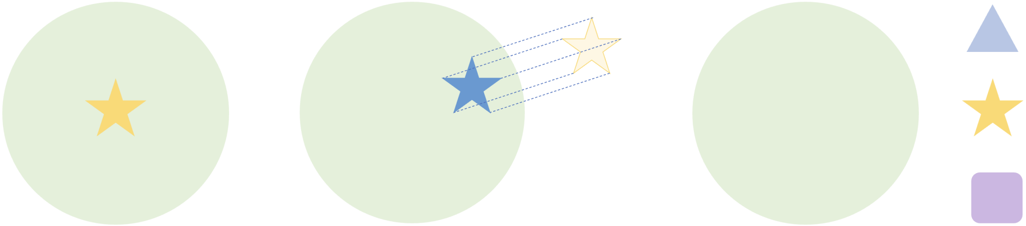
$\mathcal{M}$-Framework

- $\mathcal{M}$-**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 $p(\tilde{y} \mid y) = p(\tilde{y} \mid M_t, y)$ cannot be specified.

I like this figure :)



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

Bayesian Stacking (BS)

simple explanation like this

SUGGESTION: explain equation like this

- **BMA:** (weighted avg. of posteriors from each model)

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

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

$$p_{k,-i} = \int_{\Theta_k} \underbrace{\boxed{p\left(y_i \mid \theta_k, x_i, M_k\right)}}_{\text{likelihood for } y_i \text{ given..}} \underbrace{p\left(\theta_k \mid M_k, \{(x_{i'}, y_{i'}) : i' \neq i\}\right)}_{\text{posterior prob. for } \theta_k \text{, given } (x_i, y_i)\ i \neq i'} d\theta_k$$

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

Drawbacks: Bayesian Stacking use input independent weights.

Bayesian Hierarchical Stacking (BHS)

- **Bayesian Stacking**: (same stacking weight for all observations)

use color for w_k

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

- **Bayesian Hierarchical Stacking**: (observation specific weight).

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

use color for this.

Bayesian Hierarchical Stacking (BHS)

Make sure stuff in this slide is corrected.

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

$$p\left(\tilde{y}_i | x_i, M_k, \mathcal{D}'\right) = \int p\left(\tilde{y}_i | x_i, M_k, \theta_k\right) p\left(\theta_k | M_k, \mathcal{D}'\right) d\theta_k \tag{1}$$

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:

$$p\left(\tilde{y}_i | \mathrm{w}, x_i\right) = \mathbb{E}_{\mathcal{D}'}\left[p\left(\tilde{y}_i | \mathrm{w}, x_i, \mathcal{D}'\right)\right] = \sum_{k=1}^{K} w_k\left(x_i\right) \mathbb{E}_{\mathcal{D}'}\left(p\left(\tilde{y}_i | x_i, M_k, \mathcal{D}'\right)\right) \approx \sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i} \tag{2}$$

$$\log(p(\mathcal{D} | \mathrm{w})) \approx \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} w_k\left(x_i\right) p_{k,-i}\right) \tag{3}$$

Bayesian Hierarchical Stacking (BHS)

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

$$\log p(\mathrm{w}(\cdot) \mid \mathcal{D}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k \left(x_i\right) p_{k,-i} \right) + \log p^{\text{prior}} \left(\mathrm{w}\right) + \text{ constant}, \quad \mathrm{w}(\cdot) \in \mathcal{S}_K \tag{4}$$

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

$$w_{jk} = \frac{\exp\left(\alpha_{jk}\right)}{\sum_{k=1}^{K} \exp\left(\alpha_{jk}\right)}, 1 \leq k \leq K-1, 1 \leq j \leq J; \quad \alpha_{jK} = 0, 1 \leq j \leq J, \tag{5}$$

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

$$\alpha_{jk} \mid \mu_k, \sigma_k \sim \mathcal{N}\left(\mu_k, \sigma_k\right), k = 1, \ldots, K-1, j = 1, \ldots, J \tag{6}$$

$$\mu_k \sim \mathcal{N}\left(\mu_0, \tau_\mu\right), \quad \sigma_k \sim \mathcal{N}^{+}\left(0, \tau_\sigma\right) \tag{7}$$

Simulation Study: $\mathcal{M}$-complete setting

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

Make a table like this

$$\textbf{Model 1: } y \mid \beta_s, \sigma \sim N(\beta_{00} + \beta_1 Female + \beta_2 Escs, \sigma^2),$$

$$\beta_{00} \sim N(\gamma_{00} + \gamma_{01} * staffshort, \tau_{00}^2), \quad \beta_{ps} \sim N(\gamma_{ps}, \tau_{p0s}^2), \quad \sigma \sim Cauchy(0, 2.5)$$

	Fixed eff	Random eff
Model 1	Female, Escs	staff short
Model 2	joyread, pisadiff, staff short	staff short

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

Simulation Study: $\mathcal{M}$-complete setting

Model 2:

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

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

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

Simulation Study: $\mathcal{M}$-complete setting

Model 3:

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

$$\beta_{00} \mid \gamma_{00}, \tau_{00} \sim N(\gamma_{00} + \gamma_{01} * Public, \tau_{00}^2), \quad \beta_{7-11} \mid \gamma_{70-110}, \tau_{70-110} \sim N(\gamma_{70-110}, \tau_{70-110}^2),$$
$$\beta_{14} \mid \gamma_{14s}, \tau_{140} \sim N(\gamma_{140} + \gamma_{141} Public, \tau_{140}^2), \quad \sigma \sim Cauchy(0, 2.5)$$

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

Simulation Study: $\mathcal{M}$-complete setting

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

measures how different two distributions are

- **Kullback-Leibler Divergence Score (KLD):**

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

- **Log Predictive Density Score (LPD):**

$$\sum_i \log\left[p\left(\tilde{y}_i \mid x, y, \tilde{x}_i\right)\right]$$

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

Results: Relative Bias

Ratio	Sample Sizes (n_i*n_j)	Model 1	Model 2	Model 3	Model 4
1:1	100 (10*10)	0.002	0.001	0.002	0.002
	4900 (70*70)	0.003	0.002	0.004	0.004
1:5	500 (10*50)	0.002	0.001	0.003	0.002
	4500 (30*150)	0.002	0.001	0.003	0.003
5:1	500 (50*10)	0.003	0.001	0.004	0.004
	4500 (150*30)	0.003	0.002	0.004	0.004

Results: Predictive Performance

Ratio	Sample Sizes (n_i*n_j)	Methods	KLDs	LPDs
1:1	100 (10*10)	BS	0.156	-0.958
		BHS	0.114	-0.934
	4900 (70*70)	BS	0.127	-1.058
		BHS	0.110	-1.008
1:5	500 (10*50)	BS	0.239	-0.947
		BHS	0.200	-0.893
	4500 (30*150)	BS	0.204	-1.035
		BHS	0.172	-0.976
5:1	500 (50*10)	BS	0.086	-1.053
		BHS	0.078	-1.023
	4500 (150*30)	BS	0.079	-1.061
		BHS	0.077	-1.024

Conclusion & Discussion

- BHS outperforms BS under $\mathcal{M}$-complete setting, especially for the small sample size.
- The effects of between-group and within-group variation on prediction are left to investigate.

Conclusion & Discussion

- BHS outperforms BS under $\mathcal{M}$-complete setting, especially for the small sample size.
- The effects of between-group and within-group variation on prediction are left to investigate.

~~Thanks, y'all!~~ Thank You!

Email: mhuang233@wisc.edu
GitHub: https://https://github.com/mhuang233/BS_BHS
Website: https://mhuang233.github.io

MAKE THIS BLACK IF YOU CAN.

keep this close to the bottom if you can

use the \vspace command.