Recommendations for Bayesian hierarchical model specifications for case-control studies in Mental Health

Vincent Valton, Toby Wise, Oliver J. Robinson

University College London

Motivation

Hierarchical model fitting has become commonplace for case-control studies of cognition and behaviour in mental health (Brown et al., 2020). However, these techniques require us to formalise assumptions about the data-generating process at the group level, which may not be known (T. Maia & M. Frank, 2011). Specifically, researchers typically must choose whether to assume all subjects are drawn from a common population, or to model them as deriving from separate populations (Aylward et al., 2019). These assumptions have profound implications for computational psychiatry, as they affect the resulting inference (latent parameter recovery) and may conflate or mask true group-level differences.

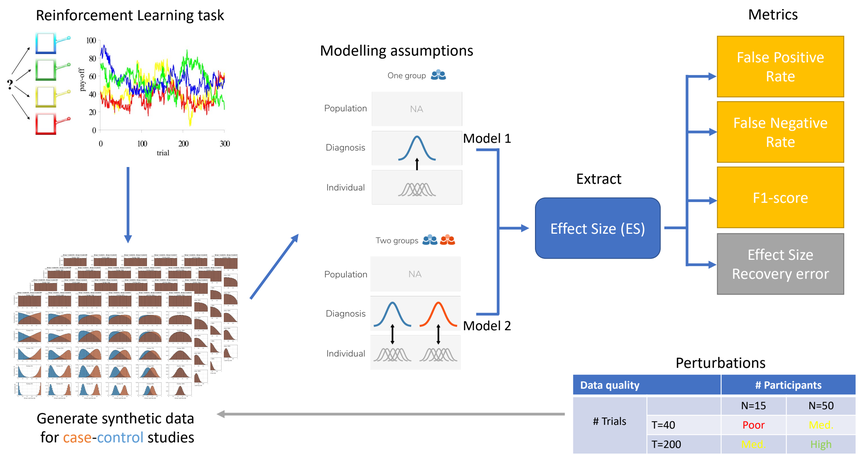
To test these assumptions we ran systematic simulations on synthetic multi-group behavioural data from a commonly used multi-armed bandit task (reinforcement learning task - Seymour et al., 2012). We then examined recovery of group differences in latent parameter space under the two commonly used generative modelling assumptions: (1) modelling groups under a common shared group-level prior (assuming all participants are generated from a common distribution, and are likely to share common characteristics); (2) modelling separate groups based on symptomatology or diagnostic labels, resulting in separate group-level priors.

We evaluated the robustness of these approaches to variations in data quality and prior specifications on a variety of metrics and found that fitting groups separately (assumptions 2), provided the most accurate and robust inference across all conditions.

Synthetic data as benchmark

Both methods (modelling assumptions) have been used on real datasets in the past (e.g. Aylward et al., 2019, Mrktchian et al., 2017), but the biases and limitations introduced by each method were unknown because researchers typically cannot access the ground truth parameters from real datasets. Our study bridges this gap by performing a robust analysis of the two methods on synthetic data where the ground truth of parameters is known and can be used to benchmark each method.

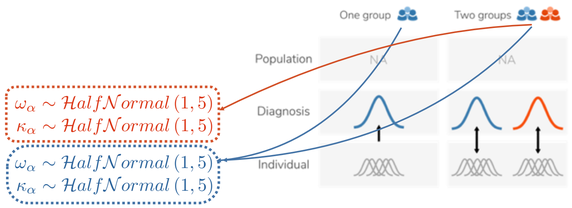
Study workflow



Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

Reinforcement Learning model specifications

Reinforcement Learning Model **Prior specifications**

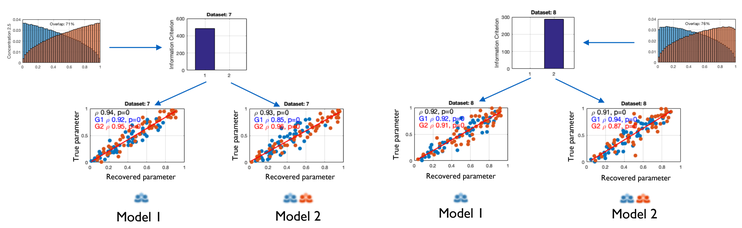


$$\alpha \sim beta\,(\omega_\alpha, \kappa_\alpha)$$
$$\tau \sim beta\,(\omega_\tau, \kappa_\tau)$$
$$V\,(s) = V\,(s) + \alpha\,(r - V\,(s))$$
$$P\,(a|\tau, V) = softmax\,(V|\tau)$$
$$choice \sim categorical\,(P\,(a|\tau, V))$$

Model selection

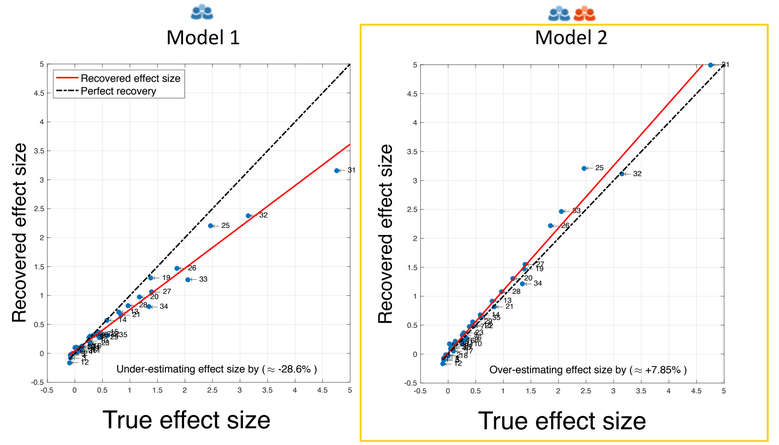
Failed to choose one consistent model spec. when fits are equally good

Model comparison (WAIC)



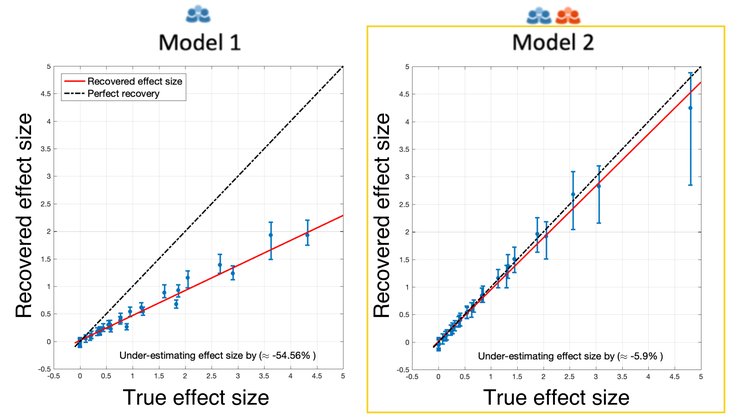
Results: Effect size recovery (using sampling)

Model 1 underestimates effect sizes, Model 2 overestimates effect sizes (#datasets= 36, high data quality)



Robustness: False Pos., False Neg. (using VB)

Model 1 vastly underestimates effect sizes, Model 2 slightly underestimate effect sizes (#datasets= 36,000 — high data quality)



Robustness: F1-Score, and data perturbations

Table 1: Model accuracy: False positive rate, false negative rate, F1-Score, and 95% CI

50 subj. 200 trials	False Pos. Rate (%)	False Neg. Rate (%)	F1-Score (%)
Model 1	0.48 [± 0.07]	6.03 [± 0.24]	96.73 [± 0.23]
Model 2	2.66 [± 0.16]	**1.75** [± **0.13**]	**98.26** [± **0.17**]

Table 2: E.S. error (%) - positive and negative sign denote over/under estimation

Data perturb.	Effect size recovery error (%) *50 subj. 200 trials*	*15 subj. 200 trials*	*50 subj., 40 trials*	*15 subj. 40 trials*
Model 1	-28.60	-17.74	-63.81	-64.46
Model 2	**+7.85**	**+9.41**	**+28.12**	**+29.09**

Conclusions

We provide here the first quantitative assessment comparing two commonly used modelling assumptions for case-control studies in mental health research.

We find that when dealing with data from multiple clinical groups, researchers should analyse patient & control groups separately as it provides the most accurate and robust recovery of the parameters of interest.

References

Aylward et al. (2019). “Altered learning under uncertainty in unmedicated mood and anxiety disorders” In: Nature Human Behaviour. 3(10):116-1123

Brown et al., (2020). “Improving the reliability of computational analyses” In: Biological Psychiatry. Cognitive neuroscience and neuroimaging 5(6):601-609.

Mkrtchian et al., (2017). “Modelling avoidance in mood and anxiety disorders using reinforcement learning” In: Biological Psychiatry. 82(7):532-539.

Seymour et al. (2012). “Serotonin selectively modulates reward value in human decision making”. In: Journal of Neuroscience 32(17): 5833-5842

T. Maia, M. Frank (2011). “From reinforcement learning models to psychiatric and neurological disorders”. In: Nature Neuroscience 14(2):154-162

ML4H Workshop

NeurIPS 2020