Bayesian Inference and Adaptive Estimation in General Recognition Theory

Joni Pääkkö

Abstract

BACKGROUND: Bayesian adaptive estimation has been applied to many problems with good results. However not in this context, no! General Recognition Theory is a multidimensional generaliztion of Signal Detection Theory.

Aims (which are all intertwined):

- To develop General Recognition Theory by integrating it more with the theory of psychometric function.

    – To argue for abandoning the ubiquitous categorization task, and to instead use a discrimination task (at least in audition)

    – To study YesNo and 2I-4AFC tasks.

- To evaluate the widely used Bayesian adaptive procedure in the context of General Recognition Theory.

    – Does it work? Are there some additional difficulties, or something to take into account?

    – How effective is it? How well are generating parameters recovered? How many trials are needed?

- To facilitate the use of Bayesian inference in GRT.

    – Easy-to-use models are made publicly available

    – Theory is (somewhat) developed

METHOD: Simulation studies and psychophysical studies.
RESULTS: TBA

Contents

1 Simulations

I will be considering two main questions:

1. How much more efficient the adaptive algorithm is in relation to sampling stimuli from a fixed grid–if at all?

2. How well can generating parameters be recovered?

These questions are closely related, since relative efficiency of the algorithms (Question 1) is defined here by the quantities that are also used to evaluate Question 2.

There are two quantities of interest. First is defined by taking the means of the marginal posterior distributions as point estimates and calculating the squared differences to the generating parameters. This quantifies squared bias and variance of these estimators. The second quantity is the standard deviations of the marginal posterior distributions. This is used as a measure of how much uncertainty about the parameters is left after the data collection process. The goal, as already stated, is to minimize uncertainty about the parameters.

Question 1 is evaluated by inspecting if there are differences between the algorithms in how quickly the aforementioned quantities approach zero. Question 2 is answered by looking at the same quantities, but the focus is on the overall performance, not differences. The first question is more closely related to the topic of this thesis, but the second question has more general value regarding the estimation of GRT models for which reason it can't be ignored.

1.1 Methods

The general method was the following: first a set of generating parameters for the simulated observer were drawn randomly, and then either of the algorithms (adaptive/non-adaptive) described earlier were run. This was done for both the Yes/No and 21-4AFC procedure, resulting in four different conditions[1].

I have chosen the number 800 fairly arbitrarily; it represents a number of trials that, I think, could still be administered relatively continuously to a participant, without taking into account non-stationarities induced by a multi-session design. I have tried to pick a number that would b large enough; if one thinks that the number is lower, one can still get that information from the graphs.

Prior distributions   Prior distributions and distributions from which generating parameters for the simulations were drawn from are shown in Figure 1 and in tabular form in Table 1. The same prior and generating distribution is used for both dimensions. Note that the scale for criterion is given in false alarm probabilities for easier interpretation.

Priors for the parameters were chosen based on prior information from Silbert, Townsend, and Lentz (2009) and from pilot testing.

Prior for σ was chosen to be fairly vague to reflect the possibility of widely differing thresholds.

Generating parameters for the simulations were drawn from bimodal distributions. The idea was to draw values that are covered by the prior, but which do not necessarily correspond with the mode

---

[1]There was a third algorithm, that consisted of sampling stimuli randomly from an adaptive grid, but results from this condition are not reported here, since it doesn't directly answer the main question posed in this thesis. The interested reader can find the data and R code to plot the results from: `https://github.com/joanpaak/Master_thesis`

of the prior distribution. Another motivation was to have qualitatively different simulated observers: some that have high values for some of the parameters and others that have low values.
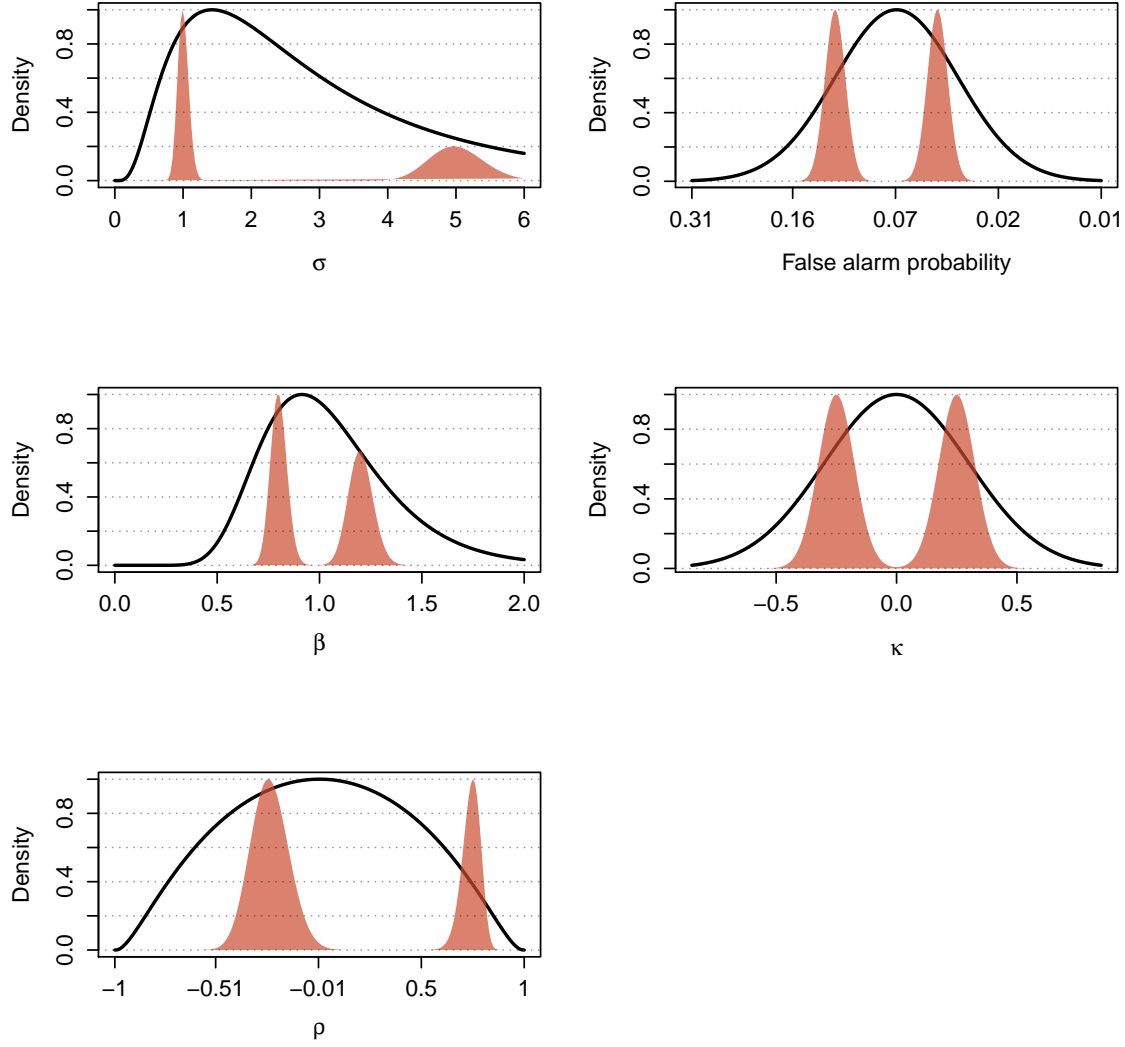

Figure 1: Prior distributions for the parameters of the model (solid black lines) and distributions for generating parameters for the simulations (regions shaded with red). Note that all of the densities are normalized to have maximum value of 1.0.

Table 1: Parameters used for prior distributions and distributions of generating parameters. M^l and M^u, respectively, are for the lower and upper peaks of bimodal distributions.

	Prior		Generating		
	M	SD	M^l	M^u	SD
σ	$log(2.5)$	0.75	$log(1)$	$log(5)$	0.08
C	1.5	0.3	1.2	1.7	0.05
β	$log(1)$	0.3	$log(0.8)$	$log(1.2)$	0.05
κ	0	0.3	-0.25	0.25	0.075
ρ	0	0.7	$atanh(0.75)$	$atanh(-0.25)$	0.1

1.2 Results

I will present trial-by-trial estimates and a more detailed analysis of differences during the last trial.

Trial-by-trial estimates Trial-by-trial results from the simulations are summarized in Figures 2 to 9. The figures show squared errors in relation to generating parameters and marginal standard deviations after N trials, all the way to trial number 800. In all plots black color is used for the randomly sampled stimuli while red is used for adaptively sampled stimuli. The shaded regions indicate 50%-quantiles (from 25% to 75%); solid lines indicate medians.
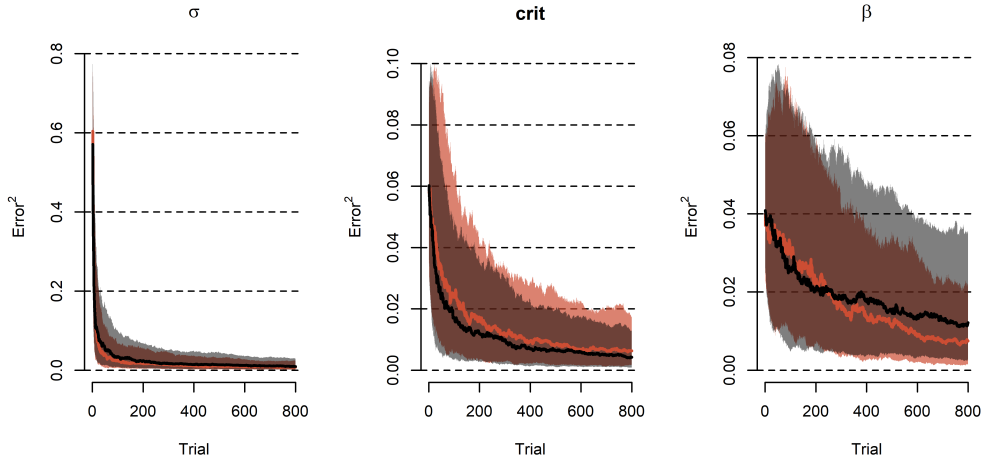

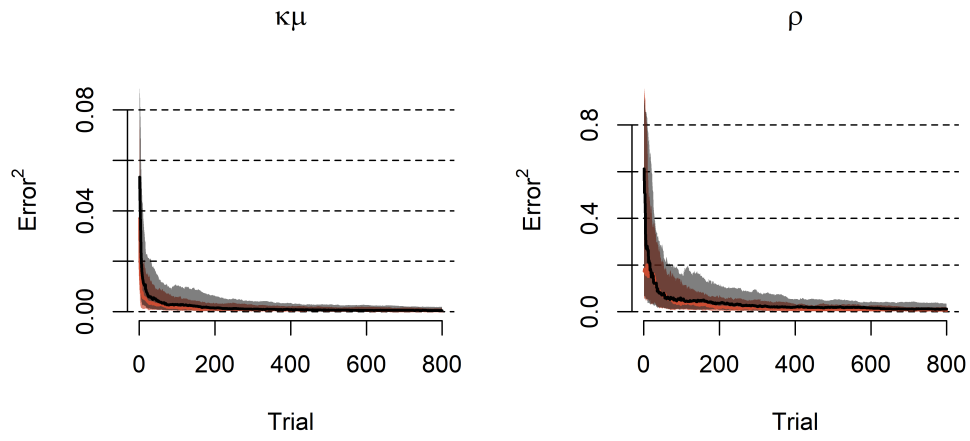Figure 2: Yes/No, sensory parameters, squared error.

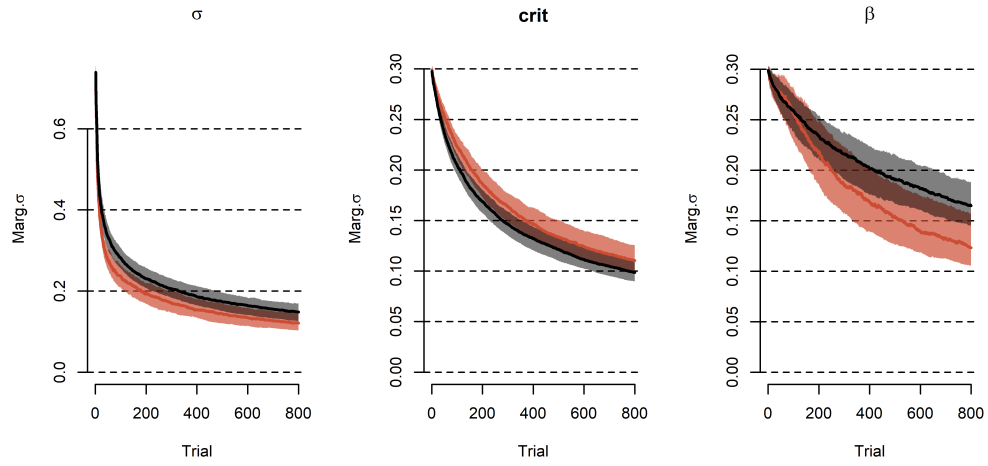Figure 3: Yes/No, interaction parameters, squared error.

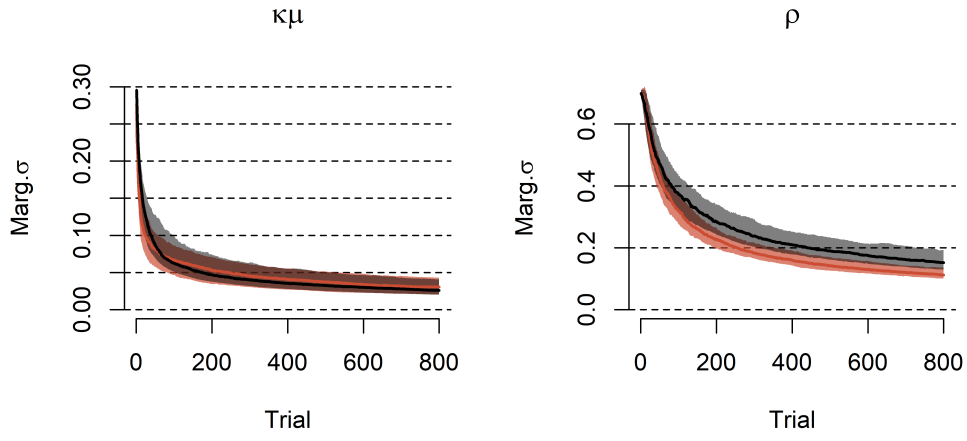Figure 4: Yes/No, sensory parameters, standard deviation.

κμ

ρ

Figure 5: Yes/No, interaction parameters, standard deviation.
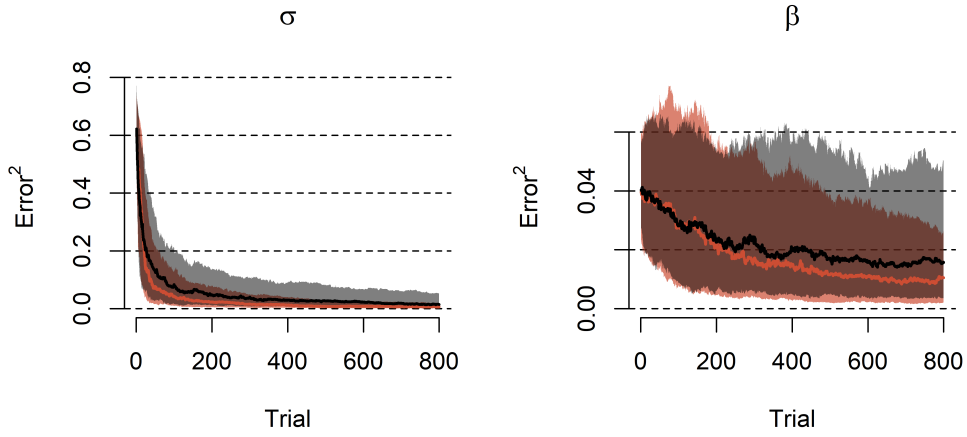
σ

β

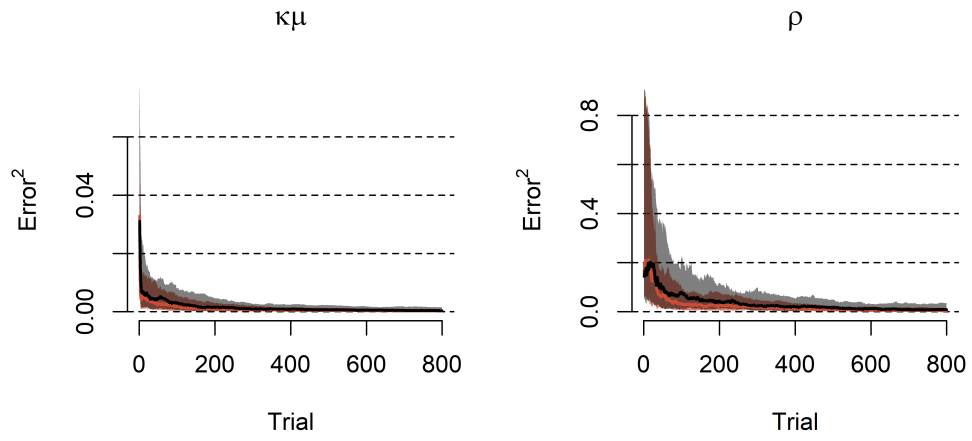Figure 6: AFC, sensory parameters, squared error.

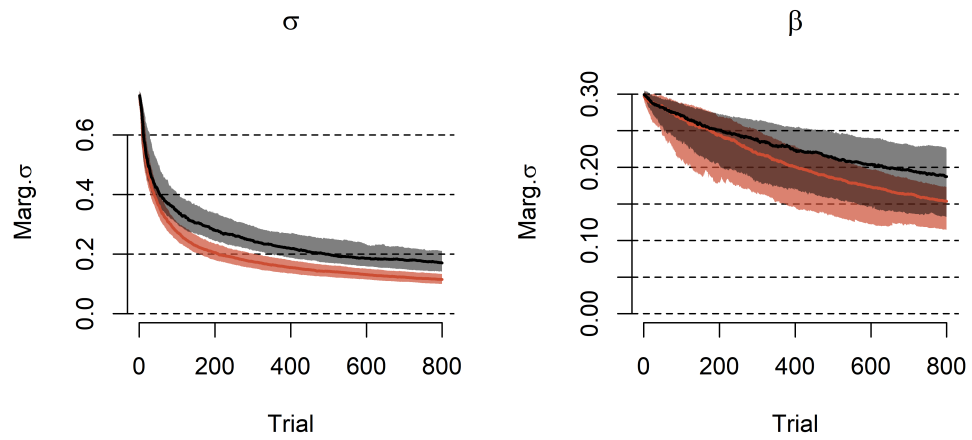Figure 7: AFC, interaction parameters, squared error.

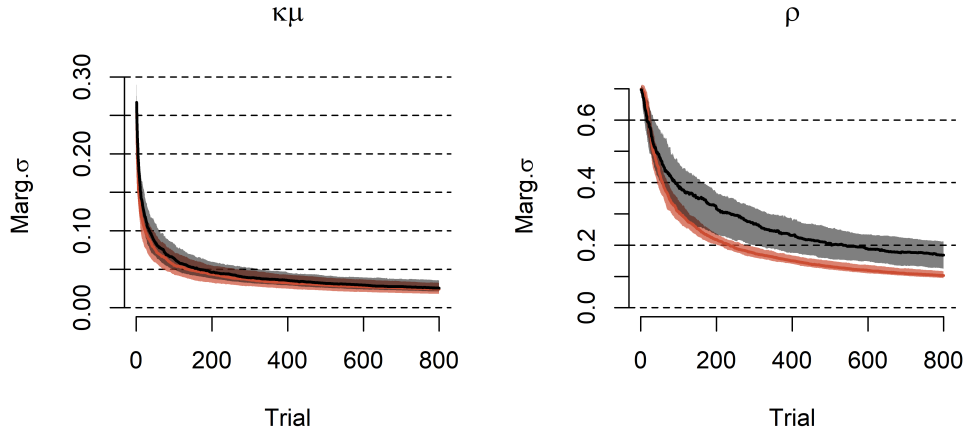Figure 8: AFC, sensory parameters, standard deviation.

Figure 9: AFC, interaction parameters, marginal standard deviations.

Hierarchical model A hierarchical model was fit to both the marginal standard deviations and squared errors at the last trial to get a more quantitative understanding of the differences at that point. A common dummy coded linear model with Gaussian errors was used. The slopes, intercepts and standard deviations were pooled inside each condition, which means that e.g. all of the marginal standard deviations in the Yes/No condition with adaptively selected stimuli were given a common hyperprior. The models were fit using Stan, the model code can be found online from `https://github.com/joanpaak` ERROR.

Data was standardized before fitting. Individual β coefficients of the linear model are shown in Figures from 10 to 13. In all plots the thicker parts show 50% and narrower lines 95% confidence intervals. Positive values indicate better performance by the adaptive algorithm.
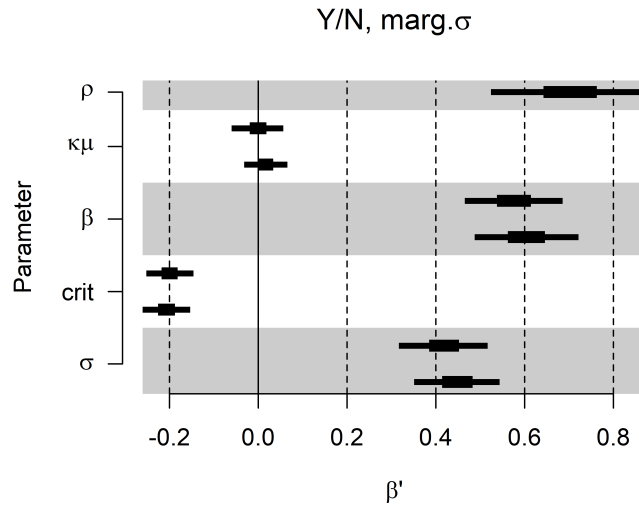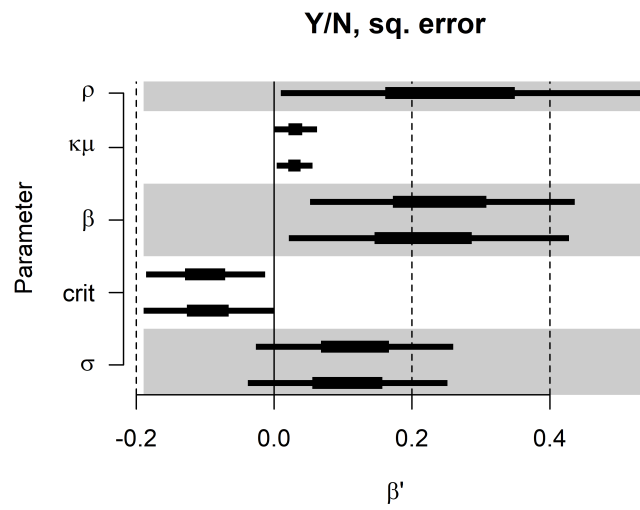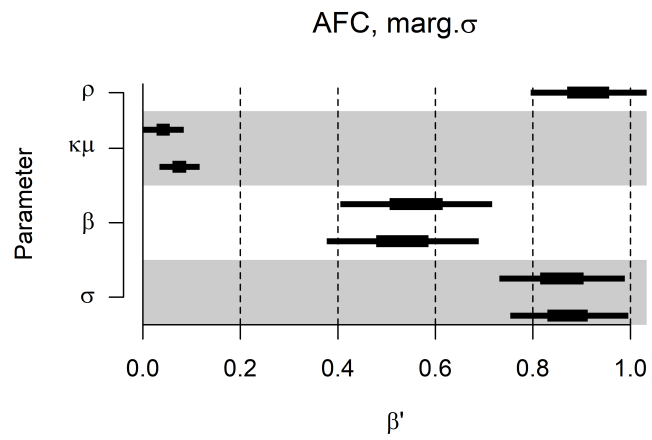

Figure 10: YN SD
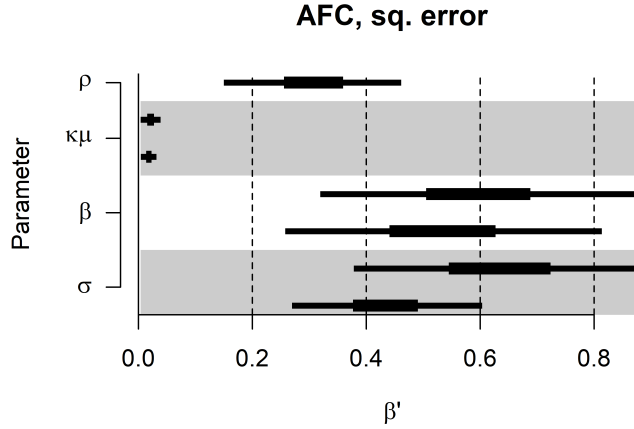
Figure 11: YN SQ Error

Figure 12: AFC SD

Figure 13: AFC SQ Error

1.3 Discussion

Question 1: is the adaptive algorithm more efficient? Judging from Figures 10 and 12 by the time of 800 completed trials the adaptive algorithm has managed to reduce marginal standard deviations more for parameters σ, β and ρ in both conditions. Effect sizes range from around 0.4 to almost 1.0. However, as can be seen from Figures 4, 5, 8 and 9, differences in raw scores are fairly modest, around 0.05 to 0.20.

There doesn't seem to be that much difference in κ_μ parameters, which is probably explained by the fact that, as can be seen from Figures 9 and 5 the marginal standard deviations for these parameters reduce fairly quickly.

The most surprising result is that random sampling seems to be more effective in reducing uncertainty about the criteria (Figure 10).

Similar pattern can be observed for the squared errors (Figures 11 and 13) but the differences are swamped by lot more variability.

Main conclusion is that in terms of squared error and marginal standard deviations, adaptive estimation doesn't seem to offer drastic enhancements over random sampling.

Question 2: how well are generating parameters recovered? It seems that the majority of of improvement happens before 400 trials. After that, information gain seems to slow down considerably. It would seem that there would still be room for improvement in the estimates even after 800 trials, however analysis of such multi-session designs is beyond the scope of this work, since one would have to be more serious about non-stationarities induced by breaking the observation periods into multiple sessions.

The most problematic parameters would seem to be the β parameters. This contrasts results in for example Kontsevich and Tyler (1999), and could indicate that when one attemps to do inference on interactions, inferences about the non-linearity of the d' function become more uncertain.

What is interesting to note is that variance of error in estimating the β parameter increases somewhat during the first 100 trials (figures 2 and 6), implying that for some simulations the

posterior means drift further from the generating parameters.

With regard to the interaction terms, squared errors and marginal standard deviations for the κ_μ parameters seem to approach fairly quickly, most of the improvements seems to have happened well before 100 trials, but the ρ parameters seem more problematic. Marginal standard deviations for the ρs don't get, on average, much smaller than .2 which still implies a lot of uncertainty–remembering that correlation coefficients are bound between -1 and 1.

References

Kontsevich, L., & Tyler, C. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16).

Silbert, N., Townsend, J., & Lentz, J. (2009). Independence and separability in the perception of complex nonspeech sounds. *Attention, Perception, & Psychophysics*, *71*(8), 1900 - 1915.