STAT 927 Bayesian Statistics Final Project (HW4)
Hierarchical Bayesian Analysis of Double Pass Experiment

Lingqi Zhang (lingqiz@sas.upenn.edu)
Department of Psychology and Computational Neuroscience Initiative
University of Pennsylvania

Introduction

Visual perception is challenging because of the various form of variabilities and ambiguities involved due to the inverse-problem nature of the process. Furthermore, in the context of human and animal vision, the problem is worsen by the unreliability of the neural system: Even with fixed stimulus (retinal images), the responses of neurons are highly variable (e.g., Poisson-like firing in spike patterns), only bringing more uncertainty to the process.

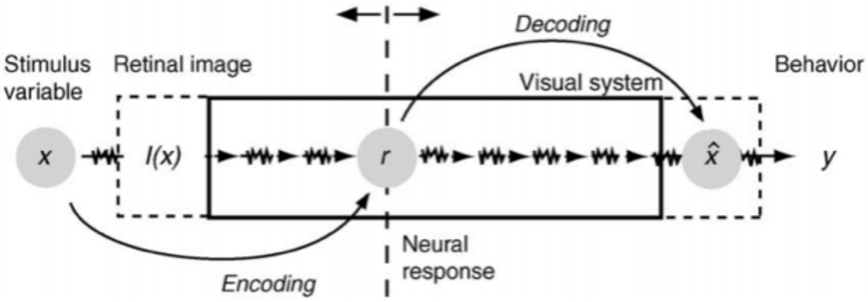

Figure 1: Overall process of visual processing, figure adopted from [5].

We can formulate these problems with a two-stage, encoding-decoding model of the visual system (Fig. 1). The goal of visual perception is to extract behaviorally relevant latent variable x (e.g., speed of a moving stimuli) from the noisy responses of neural population $\vec{r}$. The first problem, namely the nature of inverse-problem of vision, is captured by the *generative* model of images $p(I|x)$: Given a particular latent variable x, there are infinitely many possible images I correspond to it. For example, for $x = 5$ m/s, I could be different objects (cars vs. trains) moving at the same speed, or it could be the same moving object with different colors, or under different luminance conditions. These irrelevant, *nuisance* variations are what made vision particularly hard, and one goal of the visual system is to comp up with an *invariant* representation. In fact, for a noiseless system (e.g., convolutional neural network), we can simply describe it as a complex, nonlinear mapping $\vec{r} = f(I)$ that aims to reduce as much variability as possible. However, real neural system is unreliable and noisy, so given a particular retinal image, we also have a probability distribution over all possible firing patterns of the neural population $p(\vec{r}|I)$. The total uncertainty or "noise" for estimating a particular latent variable x can then be written as:

$$p(\vec{r}|x) = \int p(\vec{r}|I)p(I|x)dI \tag{1}$$

From the noisy response pattern $\vec{r}$, the visual system can then "decode", or form a percept $\hat{x}$ of the stimulus. Usually we assume the decoding process is deterministic (e.g., a MAP estimator). The uncertainty or "noise" in the final estimate can be represented as $p(\hat{x}|x)$, which is resulted from the joint effect of stimulus variation $p(I|x)$ (termed *external variability*) and response variation $p(\vec{r}|I)$ (termed *internal variability*) [2]. The goal of this data analysis is to measure the relative impact of internal and external variability on perceptual performance from human psychophysics data.

Experimental Method

The latent variable x we asked human subject to estimate in our experiment is binocular disparity, which is the slight difference in the perceived image by two eyes. It is primarily used by the brain for depth estimation [1]. On each trial of the experiment, the subject saw two images of different binocular disparity, and were asked which one has a greater perceived depth. The images were sampled from a large database and never repeat itself, however, in one block of the experiment, one of the image always corresponded to a fixed value of the latent variable (disparity levels), called "standard", and the other image can take on one of the nice different "comparison" disparities.

We can think of subjects' response as computing the "decision variable" $D = p(\hat{x}_{cmp}|x_{cmp}) - p(\hat{x}_{std}|x_{std})$ and compare it with 0, given a fixed standard and comparison value: If $D > 0$, then the comparison image has a greater depth, otherwise is the standard image that has a greater depth. Furthermore, we assume D is Gaussian distributed with some mean μ and variance σ_T^2. So across many different trials, the probability of choosing the comparison image (we denote this kind of choice as $+$, and conversely choosing standard image as $-$) to have greater depth is given by:

$$p(+|x_{std}, x_{cmp}) = \int_0^{+\infty} \mathcal{N}(D; \mu, \sigma_T^2) dD$$

Usually, given a reasonably behaved subject, $\mu = x_{cmp} - x_{std}$ (centered around the true difference of the latent variable). Here we do not make this assumption and measure μ directly from the data. So with a fixed x_{std} and different x_{cmp}, we can measure the *psychometric curve*, which is the value of $p(+)$ as a function of x_{cmp}:

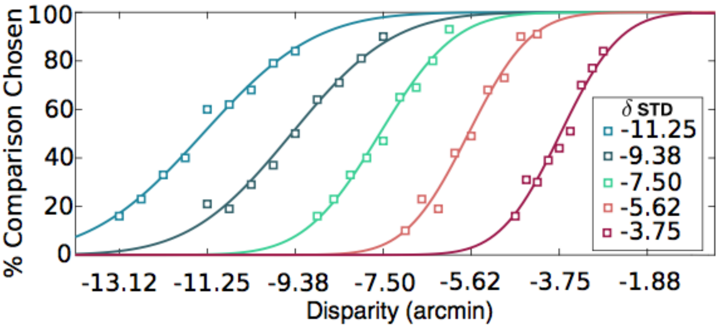

Figure 2: Psychometric curve of subject 1 in our experiment. Each line corresponds to a standard disparity condition, and there are 9 dots on each curve correspond to 9 different comparison disparities. Each dot contains 100 trials (responses). Figure courtesy of David White in our lab.

The slope of the curve is determined by σ_T^2, which is the total variability, or noise presented in the system. We can already measure the magnitude of σ_T^2 from the data above. However, as we discussed above, σ_T^2 represents the joint effect of external variability and internal variability. So we can further write $D = Z + W$, where $Z \sim \mathcal{N}(\mu, \sigma_E^2)$ representing external variability, and $W \sim \mathcal{N}(0, \sigma_I^2)$ representing internal variability. We can see $\sigma_T^2 = \sigma_E^2 + \sigma_I^2$, but how can we measure the relative contribution of these two components? The method we used is the "double-pass" experiment [7]: We repeat the experiment with the exact same set of images presented in the first pass, but with randomization and large number of trials to eliminate confounding factors such as memory. In this way, we can think that subject is under the influence of the same random variable resulted from external variability Z, while internal variability is independent of each other in these two passes: $D_1 = Z + W_1$, and $D_2 = Z + W_2$, where $Z \sim \mathcal{N}(\mu, \sigma_E^2)$ and $W_1, W_2 \sim \mathcal{N}(0, \sigma_I^2)$.

Interestingly, there is a direct relationship between the correlations of decision variables of the two passes $\mathrm{Corr}(D_1, D_2)$ and the relative influence of these two contributing sources of variability:

$$\rho = \frac{\mathrm{Cov}(D_1, D_2)}{\sqrt{\sigma^2_{D_1}\sigma^2_{D_2}}} = \frac{E[(Z+W_1)(Z+W_2)] - E[Z+W_1]E[Z+W_2]}{\sigma^2_D} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_I} \tag{2}$$

So to determine the relative impact of external and internal variability, we need to infer the correlation between decision variables D_1 and D_2 from choice data across two passes.

Statistical Model of the Data

The joint distribution of D_1, D_2 across two passes (for one fixed standard disparity and one fixed comparison disparity) can be written as:

$$\begin{pmatrix} D_1 \\ D_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right] \tag{3}$$

We are interested in inferring these parameters from the choice data across two passes. However, given only the choice pattern, the mean μ and variance σ^2 can always be arbitrarily scaled up or down together. So to make the model identifiable, we further assume that $\sigma^2 = 1$. With this assumption, we are effectively inferring the variance-normalized μ and a correlation level ρ.

For each point on the psychometric curve, there are 100 pairs of responses r_i consist of $(++, -+, --, +-)$: Here $++$ means subject consistently choose the comparison stimulus to be of greater depth, $-+$ means on the first pass, the standard stimulus was chosen but on the second pass the comparison was chosen, etc. Each point on the psychometric curve corresponds to one μ_j and ρ_j, and there are $J = 9$ total conditions on one psychometric curve. The distribution of observed data in each condition follows a multinomial distribution which can be written as:

$$p(R_j|\mu_j, \rho_j) = \Pi_{i=1}^{N}\mathrm{Mult}(r_{ij}, (\pi_1, \pi_2, \pi_3, \pi_4)) \quad \text{where} \tag{4}$$

$$\pi_1 = \int_0^{+\infty}\int_0^{+\infty} \mathcal{N}\left[\begin{pmatrix} D_1 \\ D_2 \end{pmatrix}; \begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix}\right] dD_1 dD_2 \tag{5}$$

$$\pi_2 = \int_{-\infty}^{0}\int_0^{+\infty} \mathcal{N}\left[\begin{pmatrix} D_1 \\ D_2 \end{pmatrix}; \begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix}\right] dD_1 dD_2 \quad \text{etc.} \tag{6}$$

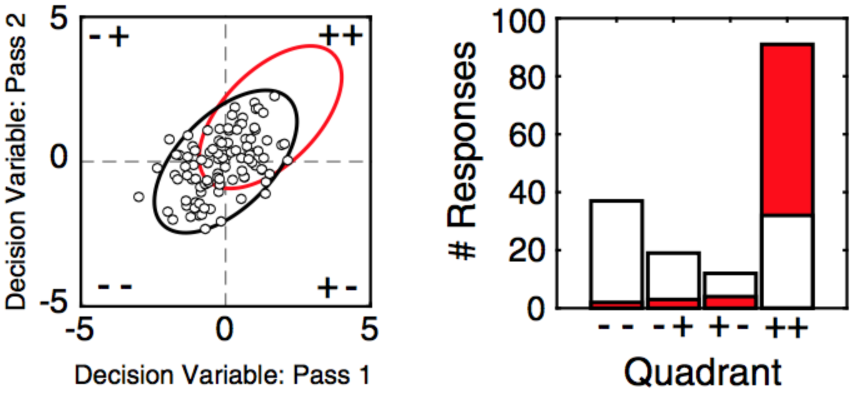

Figure 3: Model of the choice data: Multinomial with parameters as the probability mass of Gaussian distribution in each of the four quadrants. Figure courtesy of David White in our lab.

It turns out that the maximal likelihood estimator, argmax $\mathcal{L}(\mu_j, \rho_j; R_j) = \sum_{i=1}^{N} \log p(r_{ij}|u_j, \rho_j)$ is biased and unreliable in this case, for mainly two reasons: Each condition only had 100 trials, which is not enough for a reliable estimate. Furthermore, for conditions with large absolute value of μ, most of the data is either in the fist or the third quadrant (e.g., the distribution in red in Fig.3 above), only worsening the problem. So we first want to perform a Bayesian analysis to obtain the full posterior distribution $p(\mu_j, \rho_j|R_j)$, allowing us to take into the consideration the uncertainty of our estimates. Assume a non-informative prior over μ_j and ρ_j, the posterior is of the form:

$$p(\mu_j, \rho_j|R_j) \propto [\Pi_{i=1}^{N} p(r_{ij}|\mu_j, \rho_j)]p(\mu_j, \rho_j) \tag{7}$$

Which is a non-standard distribution. We can use methods such as Metropolis-Hastings algorithm to take samples directly from the posterior distribution. Alternatively, we can also implement this model with an argumentation with unobserved decision variable D:

$$p(D_j, \mu_j, \rho_j|R_j) = [\Pi_{i=1}^{N} p(r_{ij}|d_i)p(d_{ij}|\mu_j, \rho_j)]p(\mu_j, \rho_j) \tag{8}$$

And the Gibbs sampler for this posterior can be easily derived:

$$p(d_{ij}|\mu_j, \rho_j, r_{ij}) \propto p(r_{ij}|d_{ij})p(d_{ij}|\mu_j, \rho_j) \tag{9}$$

$$p(\mu_j, \rho_j|D_j, R_j) \propto p(\mu_j, \rho_j|D_j) \tag{10}$$

The first step means, for each response r_{ij}, we randomly sample one d_{ij} under the current parameters of Gaussian distribution. If d_{ij} is in the right quadrant (consistent with r_{ij}), we accept it, otherwise we sample d_{ij} again. The second step is the standard problem of inferring the parameters of a Gaussian distribution from random samples taken form it. We can choose an conjugate prior (such as Normal-Inverse-Wishart prior) to make this conditional posterior a standard distribution.

Results for Each Condition Separately

We implemented both the grid method and a sampling algorithm to obtain the posterior distribution over μ and ρ. Below we show two example conditions (3 and 7). We take $2 * 10^4$ samples from the posterior and thinning is done by taking every 10 samples.

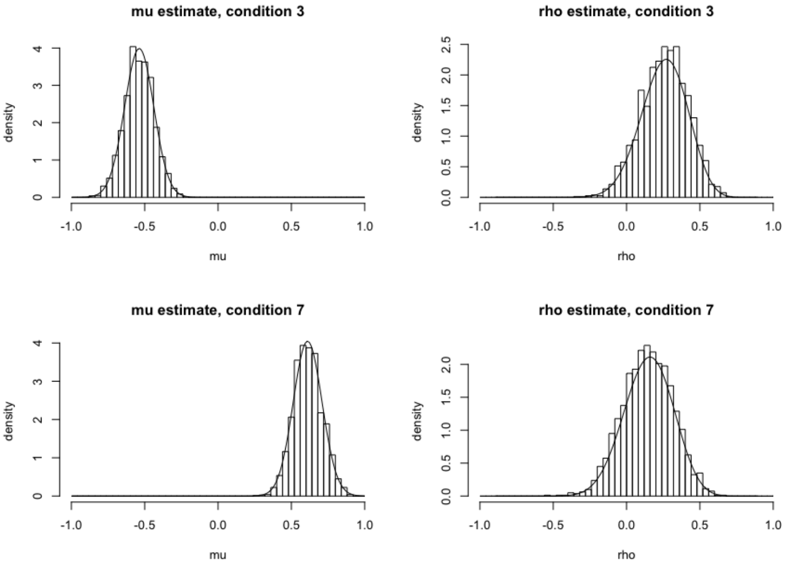

Figure 4: Posterior of μ, ρ from the gird method (numerical integration) and sampling method.

Note that first, the posterior obtained from the sampling algorithm is in-line with the grid implementation. Secondly, although we got a pretty good estimate of μ_j in both cases, there is still a large degree of uncertainty in the estimation of ρ_j. We repeated this analysis for all nine conditions along the psychometric curve, and the figure shown below is the mean and 95% posterior interval of these estimates:

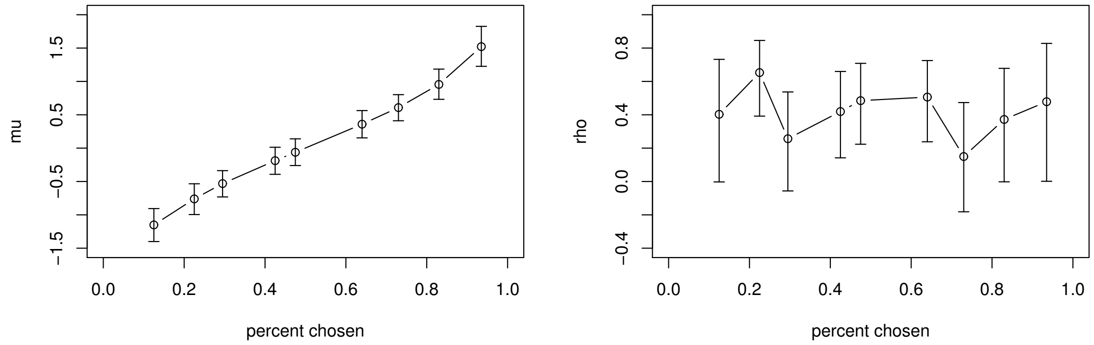

Figure 5: Mean estimates and 95% posterior interval across all 9 conditions for μ_j's and ρ_j's

The estimates of μ are highly reliable across all conditions. The mean estimates of ρ however, fluctuated around 0.4 with wide posterior intervals. We can also see that the three conditions in the middle (with percent comparison chosen around 0.5) gave us the most reliable estimates, since the responses are more evenly distributed among four types in these conditions. The conditions towards the lower or upper ends however, are associated with the highest degree of uncertainty.

Hierarchical Bayesian Model of the Data

The analysis above already provided us with a pretty good idea about the correlation level ρ in the experiment. However, since the psychometric curve is a local measurement, we would expect the level of external and internal variability to stay relatively constant. In another word, the correlation ρ should be approximately the same across all nine conditions. To model this assumption, and effectively aggregate and share information across all conditions, we build a hierarchical model in which all ρ_j's came the same distribution: $\rho_j \sim p(\rho_0, \sigma_0^2)$. ρ_0 and σ_0^2 follow a non-informative prior distribution. The posterior distribution over all unknown parameters is:

$$p(\vec{\mu}, \vec{\rho}, \rho_0, \sigma_0^2 | R) \propto \{\Pi_{j=1}^{J}[\Pi_{i=1}^{N} p(r_{ij}|\mu_j, \rho_j)] p(\rho_j|\rho_0, \sigma_0^2)\} p(\mu, \rho_0, \sigma_0^2) \tag{11}$$

The Gibbs sampler for this posterior distribution is [4]:

$$p(\mu_j, \rho_j | \rho_0, \sigma_0^2, R) \propto [\Pi_{i=1}^{N} p(r_{ij}|\mu_j, \rho_j)] p(\rho_j|\rho_0, \sigma_0^2) \tag{12}$$

$$p(\rho_0 | \vec{\mu}, \vec{\rho}, \sigma_0^2, R) = \mathcal{N}(\frac{1}{J}\sum_{j=1}^{J} \rho_j, \sigma_0^2/J) \tag{13}$$

$$p(\sigma_0^2 | \vec{\mu}, \vec{\rho}, \rho_0, R) = \text{Inv-}\mathcal{X}^2(J-1, \frac{1}{J-1}\sum_{j=1}^{J}(\rho_j - \rho_0)^2) \tag{14}$$

The first conditional posterior distribution is still non-standard. We implemented this Gibbs sampler with Metropolis-Hastings nested for taking samples from $p(\mu_j, \rho_j | \rho_0, \sigma_0^2, R)$. We took $2 * 10^3$

samples from the joint posterior distribution, and below (Fig. 6) is the marginal distribution for hyper-parameter ρ_0 and σ_0:

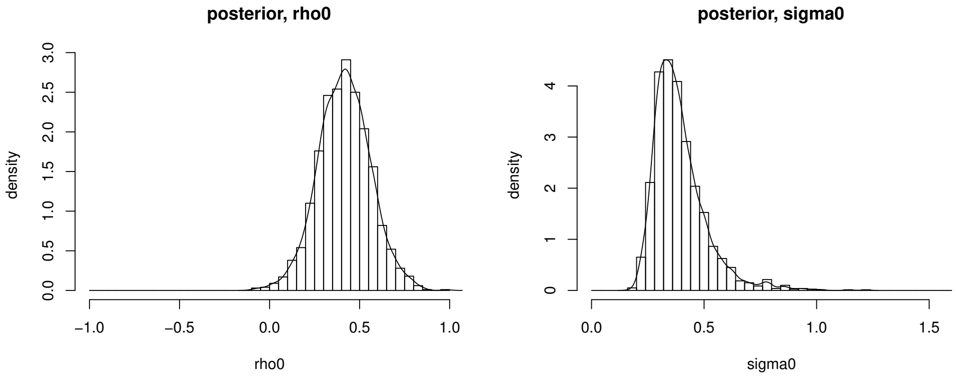

Figure 6: Marginal posterior distribution $p(\rho_0|R)$ and $p(\sigma_0|R)$, the continuous outline is obtained with kernel density estimation.

We can already use ρ_0 as our final estimation of the true underlying ρ directly:
The mean of $\rho_0 \sim p(\rho_0|R)$ is 0.415, the mode is 0.42, and the 95% posterior interval is $[0.127, 0.703]$.
The mean of $\sigma_0 \sim p(\sigma_0|R)$ is 0.392, the mode is 0.33, and the 95% posterior interval is $[0.235, 0.696]$.

We can use the individual sample μ_j, ρ_j as marginal distribution for the estimates of each condition $p(\mu_j|R), p(\rho_j|R)$ separately. We also computed the mean and 95% posterior interval of μ_j and ρ_j. However, there seems to be no noticeable effect of shrinkage on the individual posterior distributions of μ_j's and ρ_j's (so we choose to not show it here since it will look exactly the same as Fig. 5). Presumably this is because of the wide distribution over ρ_0 and σ_0. To enforce more shrinkage and further stabilize the individual posterior, we did a empirical Bayes analysis by using the a point estimate $\hat{\sigma}_0^2$ of the actual σ_0^2. The Gibbs sampler for posterior with fixed $\hat{\sigma}_0^2$ is:

$$p(\mu_j, \rho_j|\rho_0, R; \hat{\sigma}_0^2) \propto [\Pi_{i=1}^N p(r_{ij}|\mu_j, \rho_j)]p(\rho_j|\rho_0, \hat{\sigma}_0^2) \tag{15}$$

$$p(\rho_0|\vec{\mu}, \vec{\rho}, R; \hat{\sigma}_0^2) = \mathcal{N}(\frac{1}{J}\sum_{j=1}^{J}\rho_j, \hat{\sigma}_0^2/J) \tag{16}$$

Here we use the mode: $\hat{\sigma}_0^2 = 0.33$. And as before, in the first step we did the Metropolis-Hastings algorithm to sample (μ_j, ρ_j), and the second step is to sample from a normal distribution with known mean and variance.

Results for Empirical Bayes Analysis

We first compare the posterior distribution over the hyper-parameter ρ_0 (Fig. 7 below).

We can see the posterior of ρ_0 with a fixed σ_0^2 is narrower:
The mean of $\rho_0 \sim p(\rho_0|R; \hat{\sigma}_0^2)$ is 0.416, the mode is 0.407, and the 95% posterior interval is $[0.176, 0.653]$ (whereas before it is $[0.127, 0.703]$).

Next, same as before, we plotted below the marginal distribution for the estimates of each condition separately with the fixed $\hat{\sigma}_0^2$ (Fig. 8 below):

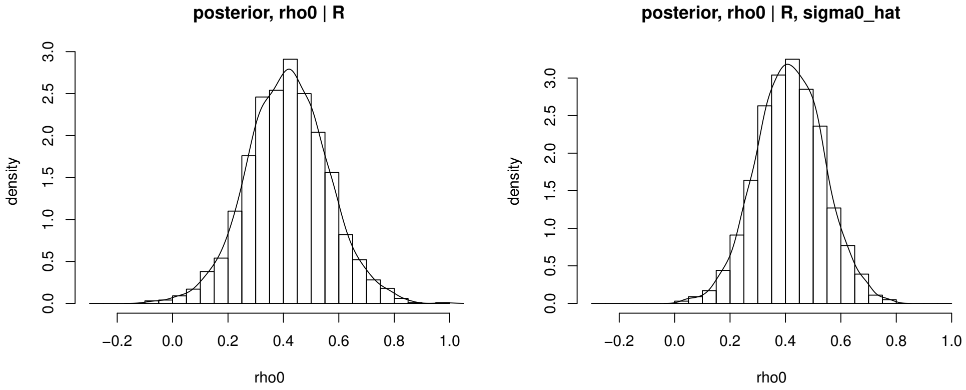

Figure 7: Left: $\rho_0 \sim p(\rho_0|R)$, Right: $\rho_0 \sim p(\rho_0|R, \hat{\sigma}_0^2 = 0.33)$

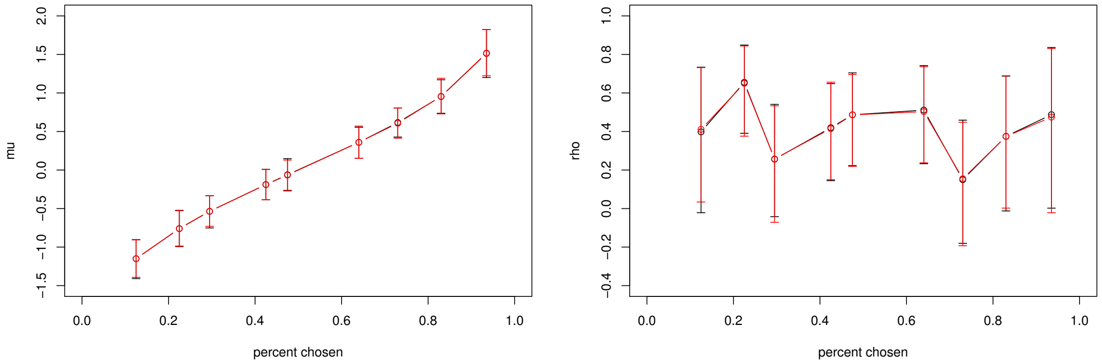

Figure 8: Mean estimates and 95% posterior interval across all 9 conditions for μ_j's and ρ_j's. We plotted the posterior under current model (red) on top of the posterior under the non-hierarchical model (black) for comparison

We can observe a small effect of shrinkage only towards the lower and upper end. We suspect that with the variance term $\hat{\sigma}_0^2$ set to 0.33, the posterior over ρ_0 is still wide (especially for a narrow domain only on $[-1, 1]$), so the posterior for individual μ_j's and ρ_j's is primarily driven by the 100 observed data point.

Conclusion

In this project, we performed a full Bayesian analysis of human psychophysics data for inferring correlation between "decision variables" in a double pass experiment, in order to determine the relative contributions of external variability and internal variability in perceptual tasks. The advantages of full Bayesian analysis is two fold:

1. Compute the full posterior distribution $p(\mu_j, \rho_j|R)$ for each condition, allowing us to take into consideration the uncertainty of the estimates, instead of relying on a single point estimate from the maximum likelihood procedure.

2. By building a hierarchical model with common prior distribution $\rho_j \sim \mathcal{N}(\rho_0, \sigma_0^2)$, we effectively allowed aggregation of data across all conditions, and $p(\rho_0|R)$ can be used as our final estimate directly. Furthermore, although there is no salient effect of shrinkage in our

particular case, there are some other conditions in our data set where the posterior $p(\rho_j|R)$ is extremely wide (range from -1 to some positive value). We would expect that applying hierarchical model to those data set will provide more advantages.

The correlation level we estimated (from $p(\rho_0|R, \hat{\sigma}_0^2)$ has a mean of 0.416, with 95% posterior interval of $[0.176, 0.653]$. This means even for complex task with huge natural stimulus variation, the amount of external and internal variation is still at a comparable level. Why biological system has such large amount of internal noise is still largely unknown (there have been some interesting proposals though, such as resource limitations on the number of neurons and total firing rate [3, 6]), but overall it is an intriguing open question that requires our further investigation.

Acknowledgement

The double pass experiment is designed by graduate student Ben Chin, David White, and Professor Johannes Burge in our lab. And thanks David for generously sharing his data.

References

[1] Johannes Burge and Wilson S Geisler. “Optimal disparity estimation in natural stereo images”. In: *Journal of vision* 14.2 (2014), pp. 1–1.

[2] Johannes Burge and Wilson S Geisler. “Optimal speed estimation in natural image movies predicts human performance”. In: *Nature communications* 6 (2015), p. 7900.

[3] Deep Ganguli and Eero P Simoncelli. “Efficient sensory encoding and Bayesian inference with heterogeneous neural populations”. In: *Neural computation* 26.10 (2014), pp. 2103–2134.

[4] Andrew Gelman et al. *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL, 2014.

[5] Nicole C Rust and Alan A Stocker. “Ambiguity and invariance: two fundamental challenges for visual processing”. In: *Current opinion in neurobiology* 20.3 (2010), pp. 382–388.

[6] Xue-Xin Wei and Alan A Stocker. “A Bayesian observer model constrained by efficient coding can explain’anti-Bayesian’percepts”. In: *Nature neuroscience* 18.10 (2015), p. 1509.

[7] David R Williams et al. “Double-pass and interferometric measures of the optical quality of the eye”. In: *JOSA A* 11.12 (1994), pp. 3123–3135.