Audio-visual causal inference, behavioral model technical report

For outlining the specific modeling steps that have been taken for translating the causal inference model of multisensory perception (initially proposed by kording et al 2007 and Sato et al 2007) into a primate behavioral paradigm. This document is somewhere between a working notebook and a first draft.

Key points I want to focus on in this paper:

1. Does CI provide a good explanation of behavior when the causality decision is an explicit feature of the task (rather than assuming unified, for instance) and both targets must be indicated (rather than only one)?
2. Does an optimal form of CI accurately describe responses, or is there some other non- optimal solution like model selection?

**Long term needs:**

* Intro and conclusion
* Replace equations with non-pictures
* Rework into a story
* Formatting
* Compare CI and alternative models

**Title: Primate behavior in a multisensory localization task is consistent with human models of optimal causal inference**

**Introduction**

There has been extensive modeling of causal inference in human subjects, suggesting that humans are able to optimally determine when two cues are originating from the same or separate sources in the world. Here we find that these same models provide good fits of non-human primates performing a similar task, which requires both a judgement about the number of targets and an inference about the location of those targets.

-contrast with integration condition

- set up CI models as a useful step forward in understanding multisensory perception and perceptual decision making

- motivate use of monkeys for addressing this question

- benefits of using saccades as a report

**Methods**

**Behavioral Paradigm**

Subjects are seated in an anechoic chamber at a distance of XXX m from a row of speakers and LEDs located on the horizontal plane. Eye movements were monitored via magnetic eye coil (Riverbend) or video eye tracker (Eyelink XXX). While fixating at a central point, subjects were presented with either a light (green LED), sound (white noise), or both at one of 10 visual (+- 6-30 degrees in 6 degree increments) or 4 auditory (+- 6 and 24 degrees) locations. After a brief delay (600-900 ms) Subjects reported percepts by making saccades to the perceived location in space and then maintaining fixation at that target location. On conditions with multiple targets, subjects are required to make sequential saccades to each target in any order. The timing of the task is such that subjects must make both saccades in rapid succession, and so cannot adopt a strategy of waiting until the reward is delivered (or not) before making a decision about the second saccade. This means that subjects are making both an explicit causal inference judgement (number of saccades) and an implicit judgement (location of stimuli) on each multisensory trial.

**Trial filtration and saccade detection**

Trials are included as long as the subject held fixation through the go cue, and then made at least one saccade, without enforcing any restrictions on saccade accuracy. For multi-stimulus trials, trials which ended within 600 ms of the go cue (the minimum duration for a successfully completed trial, see timing section [xxx]) were also excluded. This was done to minimize the number of trials which ended before the subjects full response could be reported, and ensure that single saccades were indicative of a unified percept rather than a lapse.

Saccades were defined as any eye movement exceeding 50 degrees per second and followed by at least 30 ms of very little eye movement (max velocity <25 deg/s). Saccades of less than 3 degrees were considered corrective [XXX] and were not included as responses in subsequent analyses.

**Models of Causal inference**

We build upon the standard model of multisensory causal inference [kording ref]. This includes both the ideal observer model, as well as adapted versions which are sub-optimal but for which there is support in the behavioral literature. In order to compare these models with behavior, it is first necessary to compute the response probabilities for the various conditions under each of the generative models. A brief overview of how these probabilities are computed is given below.

**Explicit causal inference (Unity judgement)**

On each multisensory trial, the subject must explicitly respond whether they perceive a single unified target (C=1) or multiple discrete targets (C=2). The subject must make this judgement by relying on noisy sensory information ( and ) as well as (potentially) prior information about the structure of the task. Below, we consider XXX strategies for how this judgement could be performed, as previously described in [XXX].

*Bayesian**CI*: A Bayesian observer performs causal inference by determining the posterior of a common cause, as specified by:

Which, for the single cause case, can be evaluated as

*A full derivation of p(c=k|xa,xv) can be found here XXX*

Here we assume that the subject reports a single target if the posterior probability is greater than 0.5, with some lapse probability ,

Where is the Iverson bracket which is 1 if the argument inside is true and 0 otherwise [ref].

*Fixed criterion:* ***TBD***

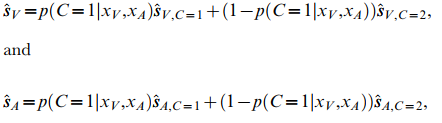
*Forced fusion*: The final possibility is that subjects are not performing causal inference at all, and are instead integrating (or segregating) stimuli on every trial regardless of perceived separation between stimuli. We consider a generalized version of this model which performs fusion or segregation randomly (stochastic fusion) at some rate . This would be analogous to learning the overall distribution of one cause and two cause trials and then randomly reporting one or two causes in agreement with that rate but without consideration of any stimulus feature.

**Implicit Causal inference (location judgement)**

There are two strategies for optimally estimating the location of and , the source of the auditory and visual stimuli, respectively, conditional on whether C=1 or C=2. For C=1, it is a problem of Bayesian integration, which in this case takes the form:

For the C=2 case, there should be two independent source estimates*,* one for each modality. These take the form:

These estimates do not yet address causal inference, and assume perfect knowledge of the causal structure of the stimuli. Therefore in order to compare with real data (which will have a mix of one and two saccade responses for a given A and V source pair), it is necessary to combine the estimates according to some judgement of causal inference. This was done using the weighted model averaging strategy that has been reported as the most common perceptual strategy used by humans in a similar task [rohe 2015, lacerbi 2018, XXX], although probability matching and model selection strategies have been proposed elsewhere [wozny 2010, XXX]. This model averaging strategy takes the form:



This strategy reweights each of the two sensory estimates (integrated or segregated) according to the value of the posterior on common cause for that target condition. This automatically results in two separate sensory estimates (one auditory and one visual), which correspond to the responses provided by subjects. Because it is permissible for subjects to make only a single saccade on a given trial when perceiving the targets as fused, these sensory distributions will not have a one-to-one mapping to behavior. However, assuming that subjects make single saccade reports only when these estimates are close together, and after correcting for difference in number of saccades across trials (see below), the average behavioral response distribution should still be well described by these estimates.

**Model fitting**

The model is fit by maximizing the likelihood (minimizing the log likelihood) of the actual saccade distributions under the probability distribution described by the model. Parameter optimization was done using a nonlinear simplex search algorithm as implemented in Matlab’s fminsearch function (Matlab 2016a). The full CI model is initialized with 5 free parameters. These are the prior on common cause (p(C)), visual sensory variance (σV), auditory sensory variance (σAn for near targets and σAf for distant targets), and centrality bias (σc). Each of these parameters was given an initial starting point determined by a prior grid search. Likelihood for each set of parameters was obtained by evaluating the probability density function at the location of each real saccade endpoint in the training dataset, and summing these values together. Likelihood was combined across all target pair conditions for a given parameter set, which enforced the assumption that the same parameter values were used for each condition (rather than each target pair having its own variance parameter, for example). In order to compensate for overweighting of dual saccade trials in the model fitting step (because these trials would contribute two saccades to the dataset, rather than the single contribution of one saccade trials), all trials were subsampled such that they only contributed a single saccade to the final distribution.

This procedure was repeated for simplified models that assume the targets are always at the same location (integrate model) and models that assume there are always two locations (segregated model). These models have identical structure to the causal inference model where p(C) is set to 1 or 0, respectively.

**Model Comparison**

In order to compare the CI and alternative models, we used a 10-fold cross validation strategy. For each fitting step, 1/10th of the data was held out and used as a testing set while the remaining 9/10th was used for parameter fitting. The total negative log likelihood across conditions was calculated for the testing set, and then the procedure was repeated with a different subset of trials held out. For the model comparison reported below, the negative log likelihoods were summed across folds. Each subject was assigned a “final” set of parameters which is equal to the mean of the fits acquired in each fold. These mean parameter values were used to generate plots for comparison with behavior, as well as when calculating the range of parameters for comparison across subjects.

**Perceptual Modeling – Comparison with behavior – This procedure has changed significantly, rework XXX**

In order to directly compare the model results with behavior, we adopted a generative approach to create synthetic saccade distributions from the fit model parameters. First, the model assumes that the one (C=1) or two (C=2) cause conditions are determined by drawing from a binomial distribution with the common source prior P(C=1) = pcommon. The model also assumes that there exists a normal prior distribution of possible target locations described by N(mup, sigp). In this iteration the value mup = 0, reflecting a distribution centered on the midline. When in the C=1 condition, a single target location (SAV) is drawn from the prior distribution and used to generate estimates of both auditory (xA) and visual (xV) percepts by sampling from a normal distribution centered on SAV with standard deviations specified by σA (currently there are two σA parameters, one for close targets and one for distant targets, this might change) and σV respectively. These draws are meant to reflect sensory noise corrupting the target estimate. The model then uses the Bayesian integration strategy described above to produce a single saccade estimate for that trial. In the C=2 condition, two target locations are drawn independently (SA and SV) and estimates xA and xV are then drawn centered on these different locations.

**Deviations from previous modeling (Mostly leaving this for the end XXX)**

List of changes:

-Saccades (analog signal)

- different likelihood calculation (because using analog)

- extra aud variance term

- holding out data

The biggest difference between this paper and any previous work on perceptual causal inference is the reporting of both auditory and visual perceived locations on all trials. This is beneficial because it incorporates both an explicit causal inference report (number of saccades) as well as an implicit one (influence of non-localized modality on saccade endpoint). However, this requires several adjustments in order to disentangle sensory and motor effects.

One of the major differences from previous models (kording) is the inclusion of a second auditory variance term. This was done because the variance on very lateral auditory targets was much lower than the variance on very medial targets. There are two reasons for this difference, the first is that the auditory discrimination is more challenging for the middle targets (12 degrees from nearest aud target, 6 deg off fixation) than on the lateral targets (18 degrees from nearest aud target, 24 degrees from fixation). The second is that the monkey has been trained to make a saccade of at least 6 degrees to either the left or right on every trial. This biases the monkey towards making large saccades, even when the target is actually “perceived” near the midline. We think it’s possible that when a monkey is making a difficult discrimination (i.e. is this target to the left or right of center?) they bias their reports to locations either on the left or right of the fixation light, rather than the true perceptual location near the midline. This would result in an overdispersion of responses, with a bimodal distribution rather than a purely unimodal one (XXX is this the case in the data? Also could be tested by changing the prior from a central one to a bimodal one).

**Results**

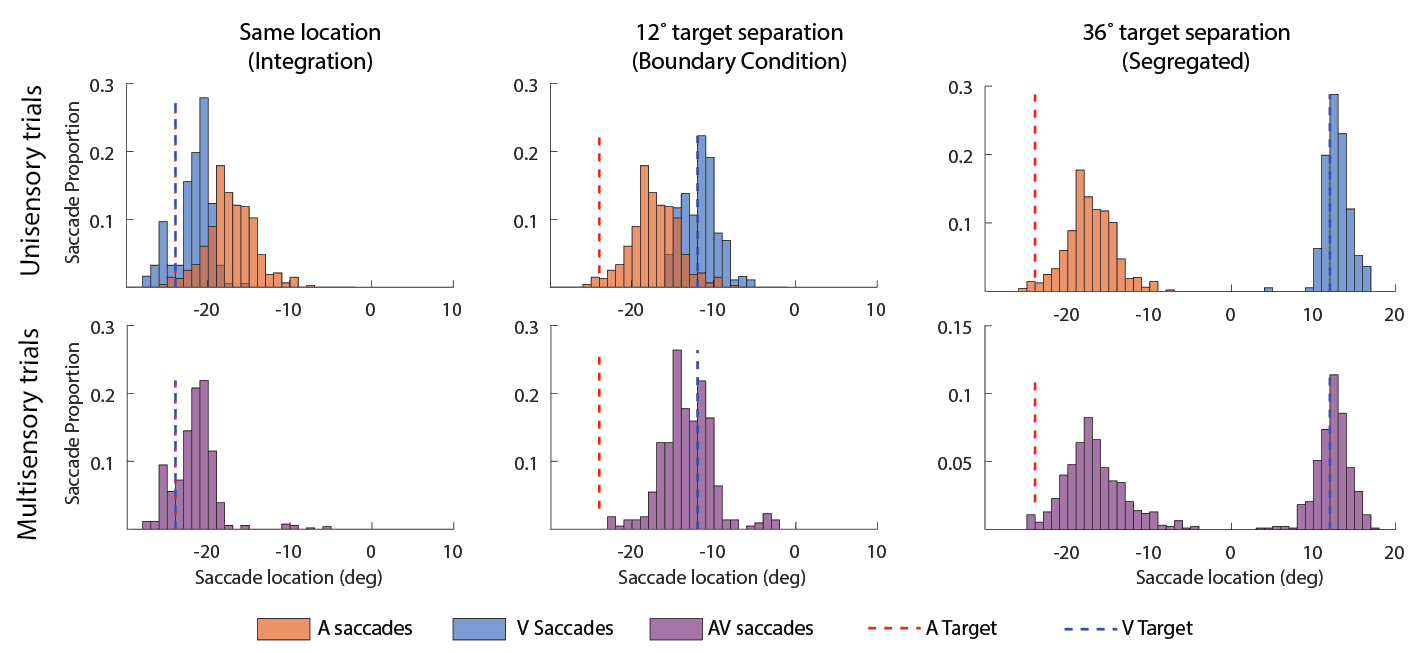
* Does the model capture behavior well
* How does the model compare with other possibilities
* What are some things we can pull out from looking at the model fits (accuracy, bias, priors)
* How do monkeys and humans compare in this paradigm
* What does this tell us about CI as a strategy used in the brain

**Humans and Monkeys are able to localize and report multiple targets in multimodal trials**

Monkeys could be trained to sequentially report both an auditory and visual stimulus, even when presented at the same time. In conditions where the targets were in the same spatial location, subjects almost exclusively made single saccades which were well described as the unification of visual and auditory distributions, with the visual dominating (figure 1, left). In conditions which had ambiguous target separation (typically 6 or 12 degrees, figure 1, center) subjects made a mixture of one and two saccade reports. Importantly, single saccade trials were distributed in between the visual and auditory target locations (consistent with integration) while multi-saccade trials were more consistent with a segregation strategy (figure S1). Finally, when targets were well separated subjects were easily able to discriminate and report both target locations, in a manner consistent with their unisensory estimates (figure 1, right).

Monkeys (XXX true for humans as well?) did not show a strong preference for order between stimuli. Although first target selection was not random within conditions, monkeys appeared to switch preferences between conditions, possibly preferring saccade patterns that were easier to generate from a motor perspective rather than preferring one stimulus type over another. This suggests that one modality does not universally dominate the other, at least with respect to this behavioral task.

XXX need to include something about the number of saccades vs target sep (in figure 2 below)

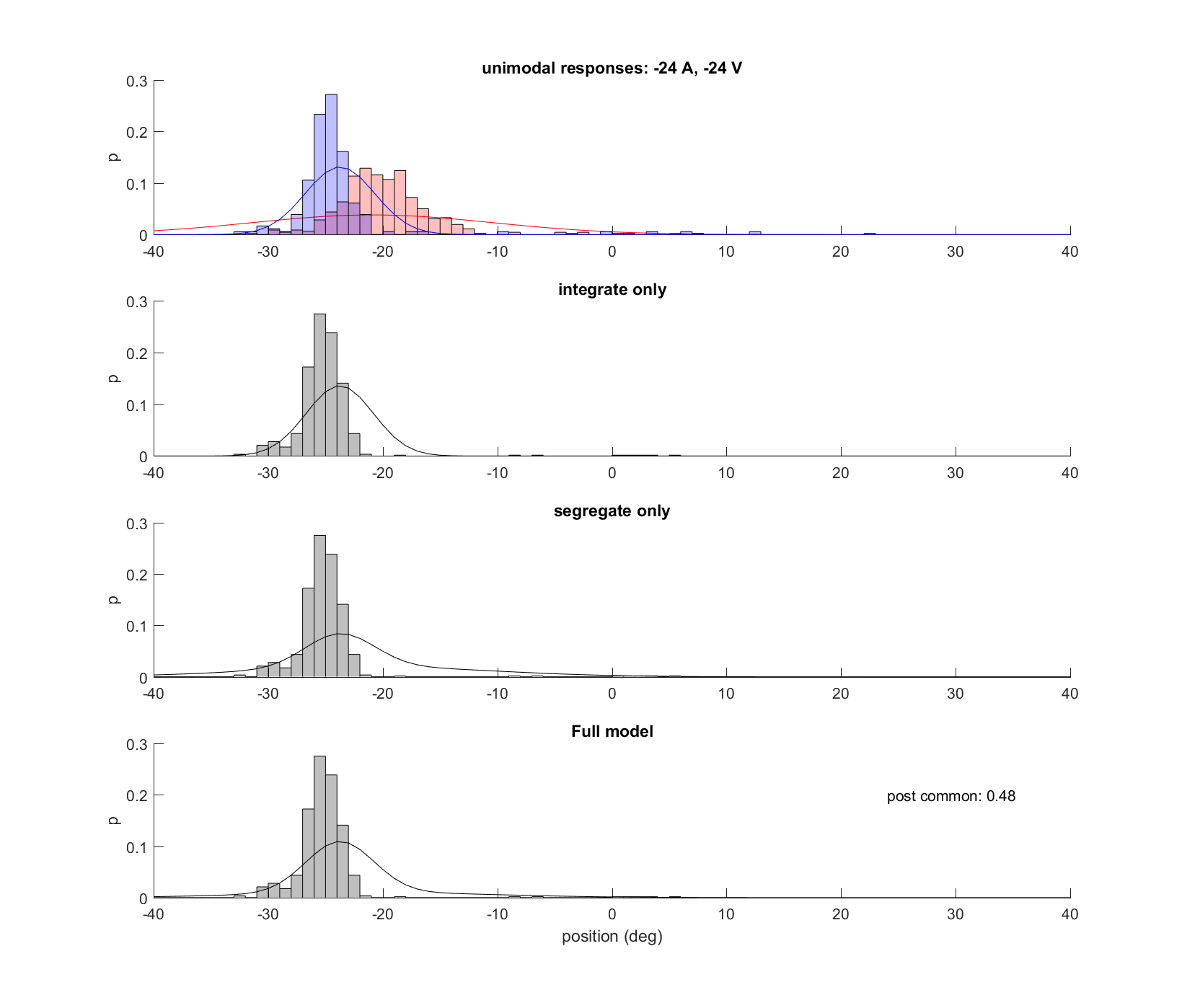
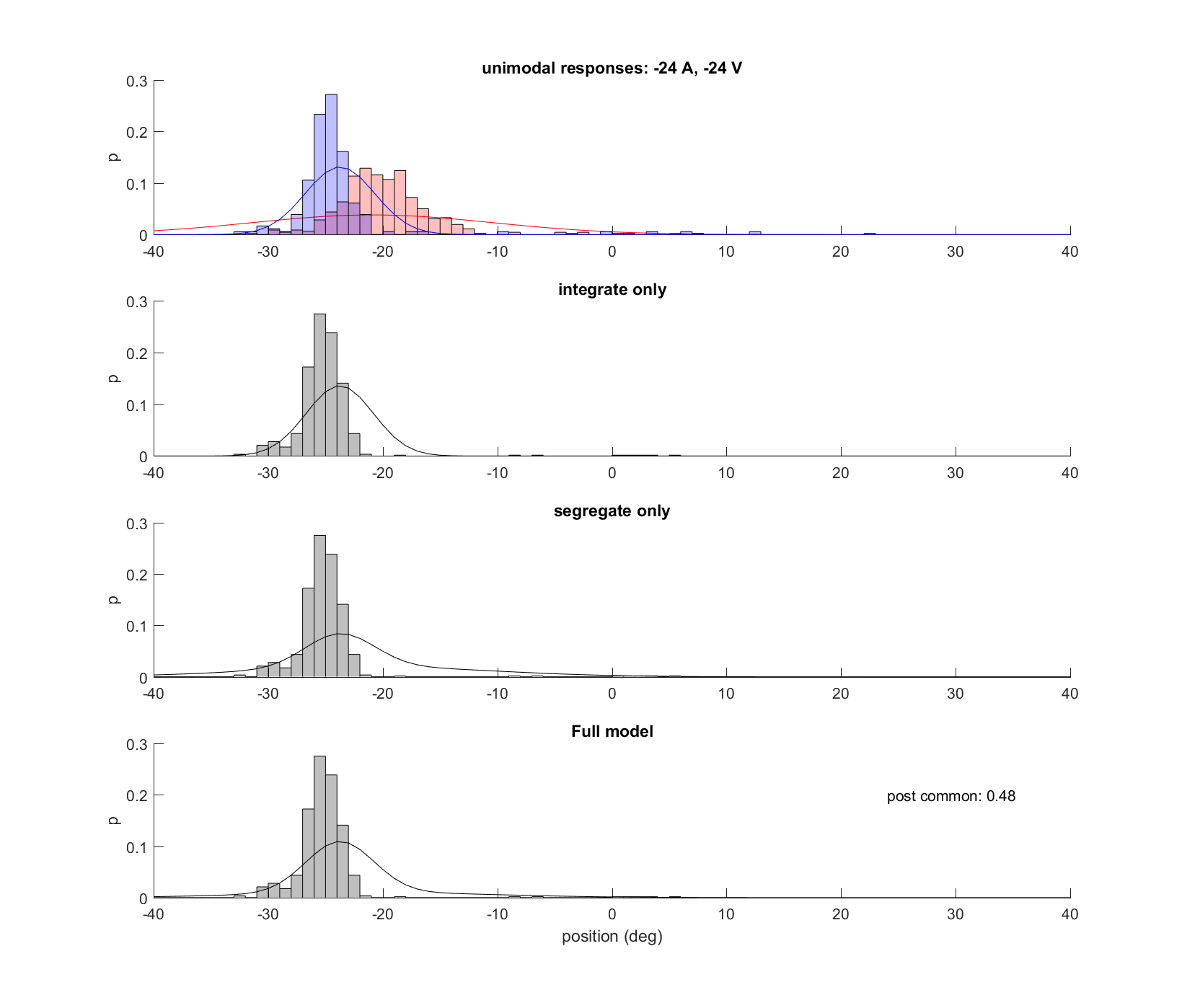
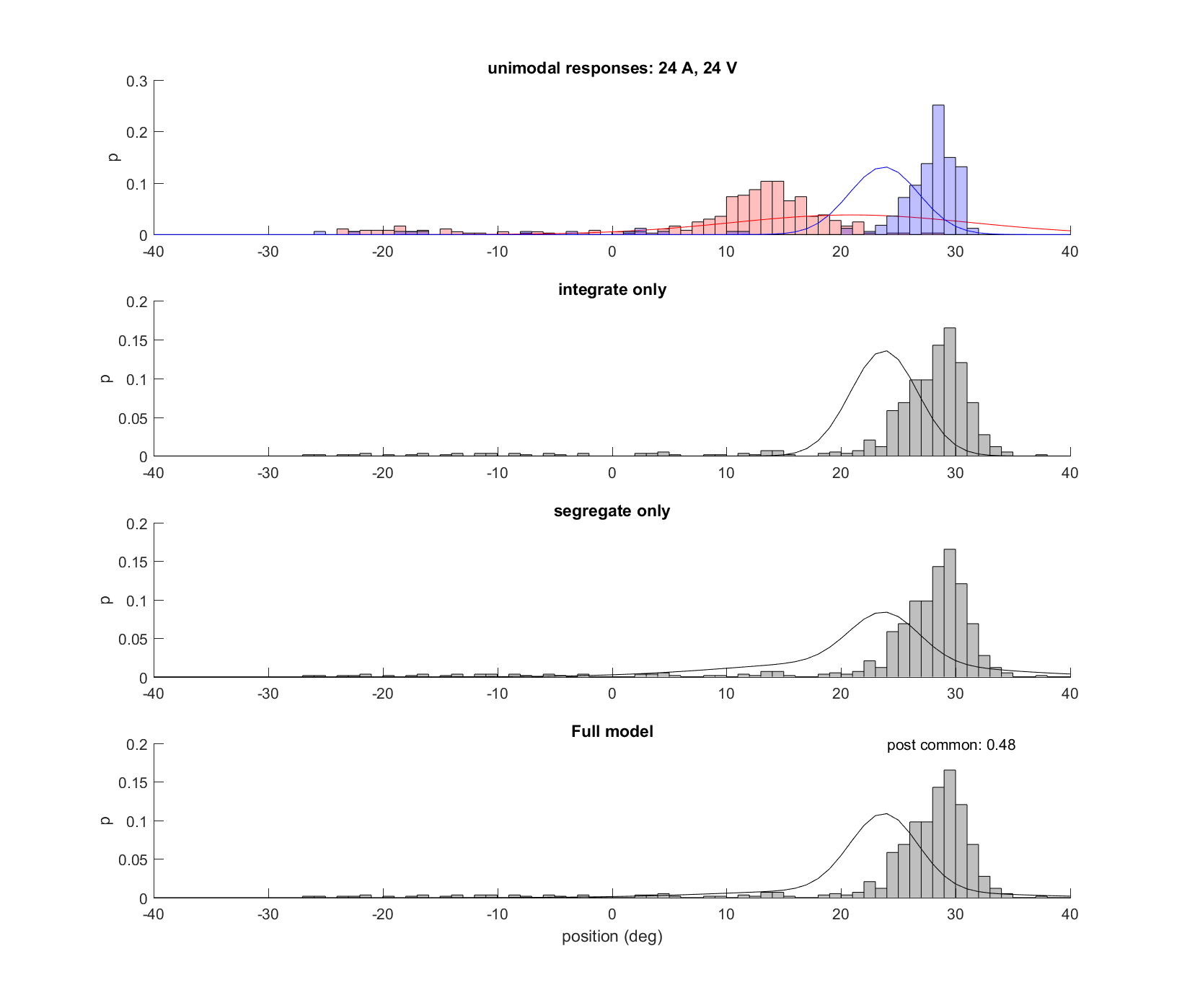
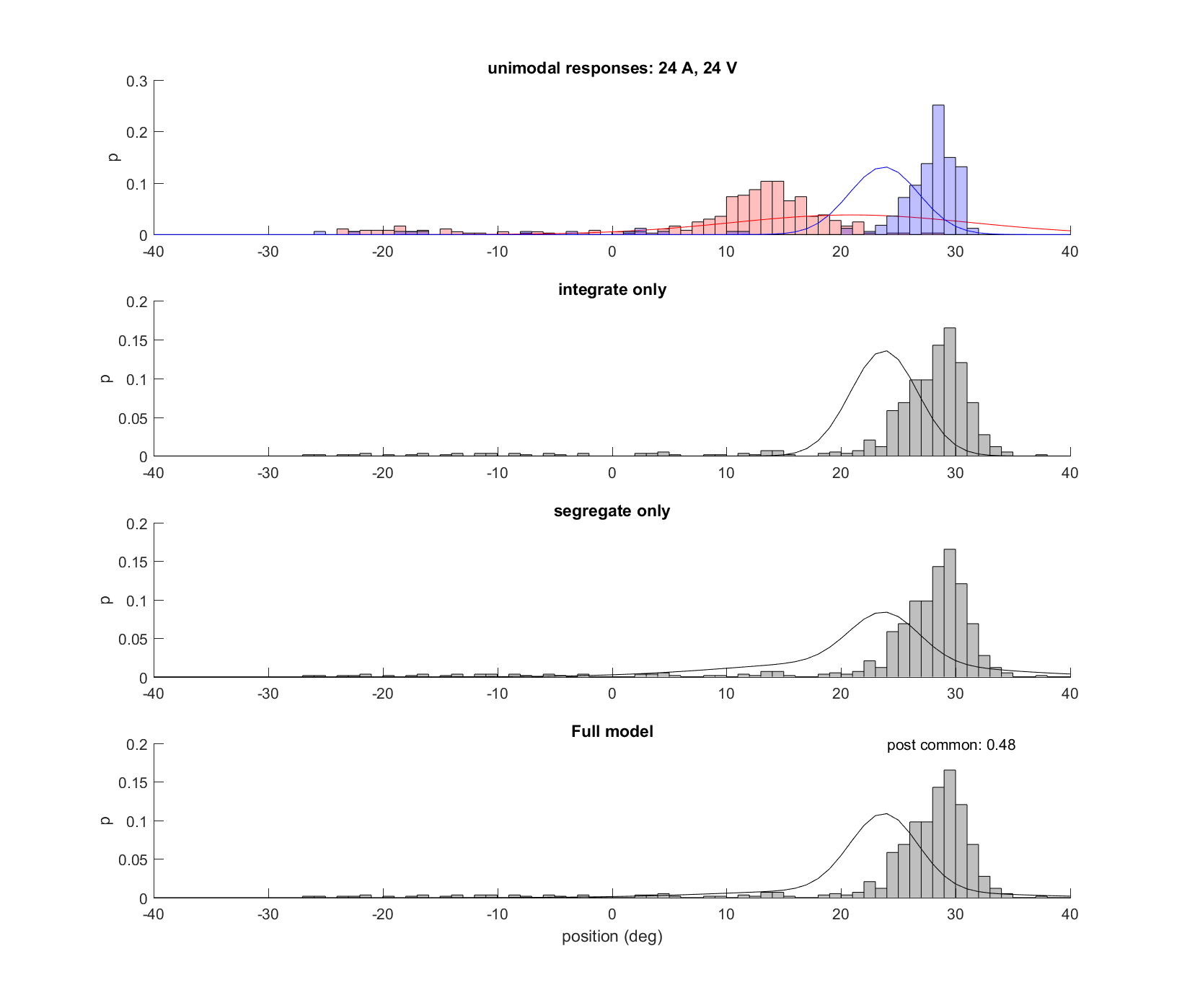


**Figure 1:** *Behavior shifts with increasing target separation.* Same location conditions (right) produce exclusively single saccade trials. In close target conditions (12 degrees of separation, center), monkeys report a mixture of fused and separated percepts (indicated by the bimodality of the purple distribution). When the two distributions are easily separable, monkeys make two saccades approximately matching those made in the single modality condition (right). This data is from 15 combined days of monkey J behavior.

**Qualitative evaluation of model XXX (lots of work needed to improve fits, see eval of fits s2 below)**

In line with previous studies, the causal inference model was able to grossly capture response locations of monkeys in both single saccade (figure 2A) and multi saccade (figure 2B) conditions, including the presence of both average and winner take all saccades in the multi-saccade case. Additionally the model was able to capture the relationship between target separation and number of saccades, which was never explicitly fit in the model (figure 2C).

**A)**

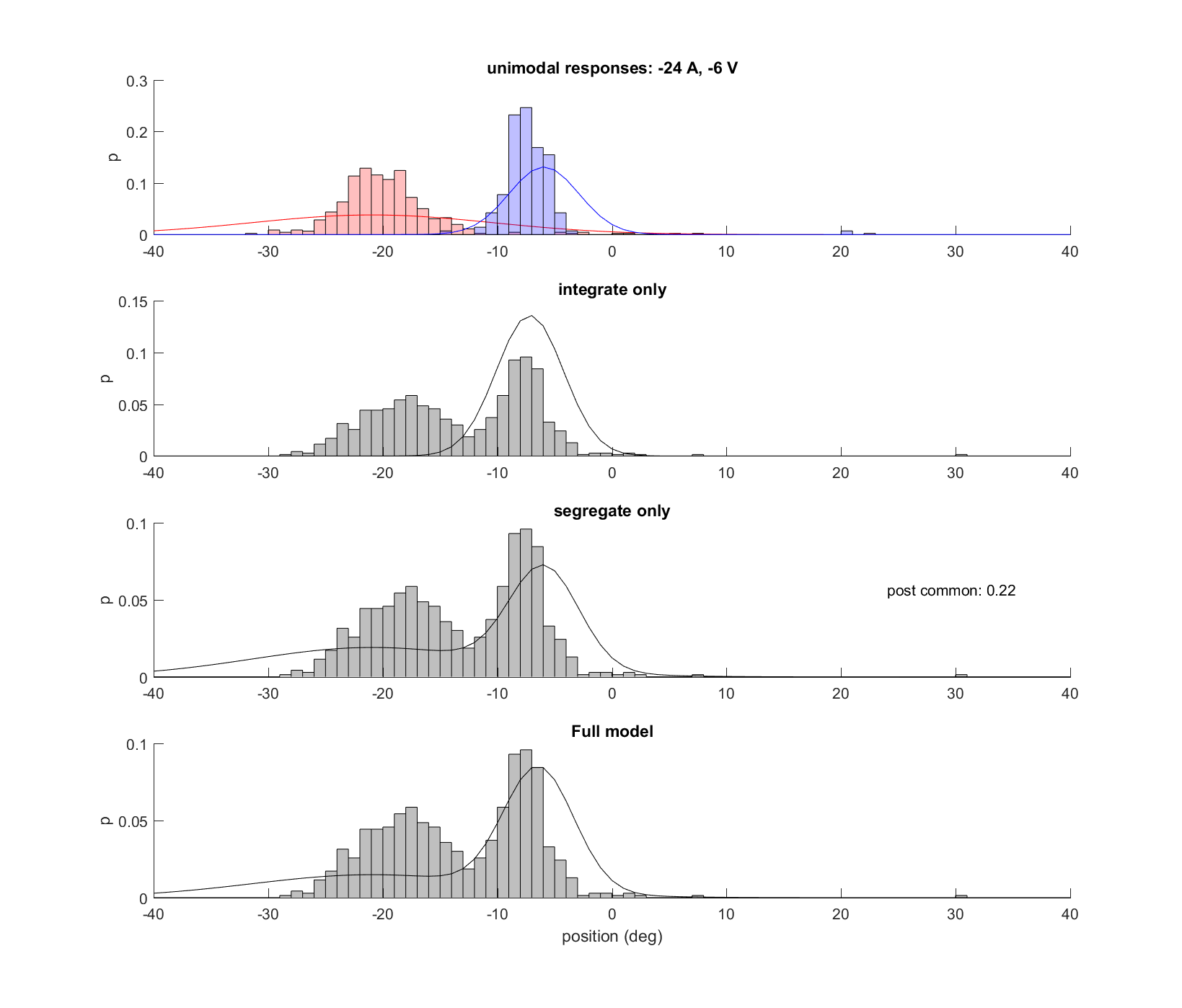
  

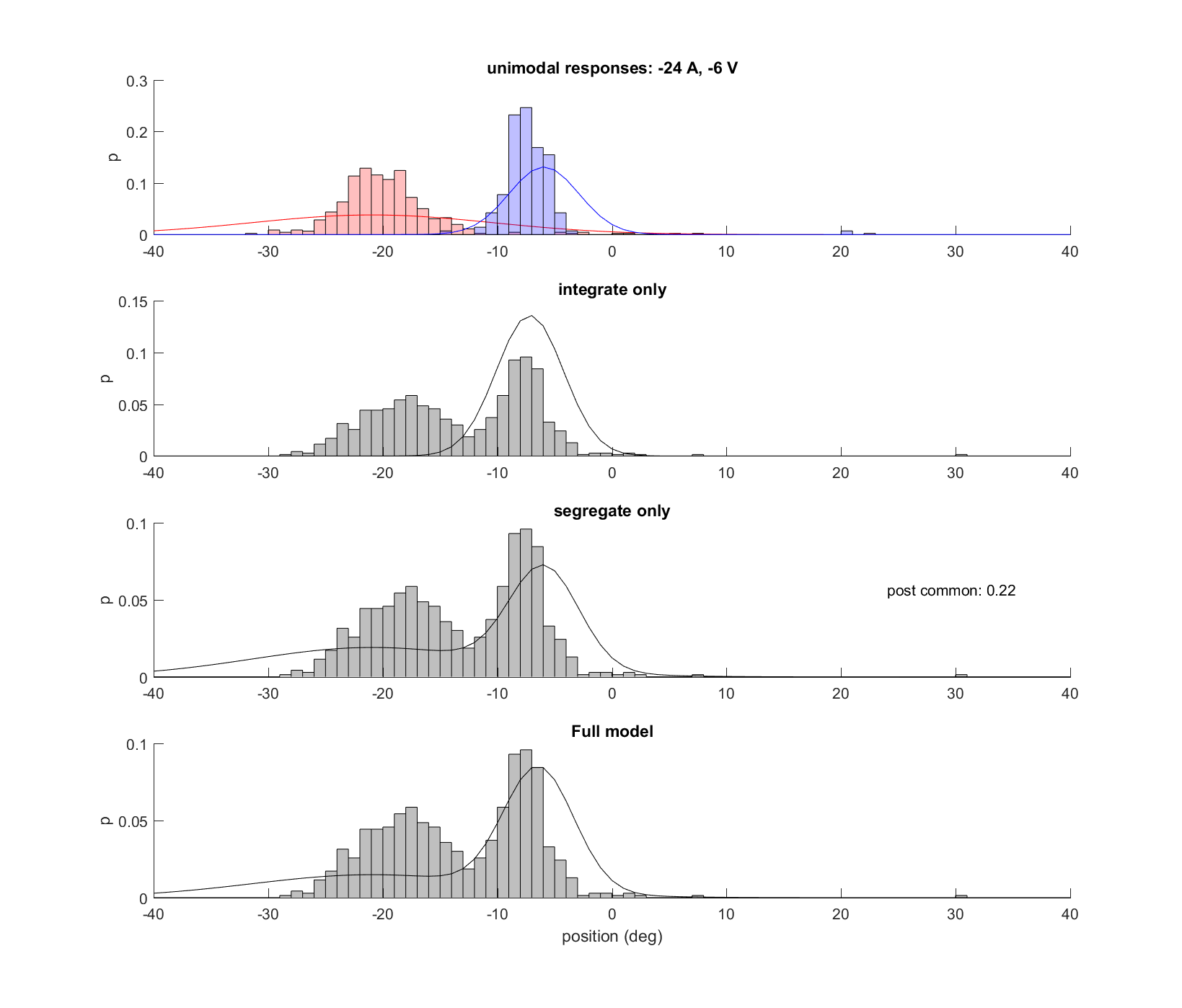
A trial saccades

V trial saccades

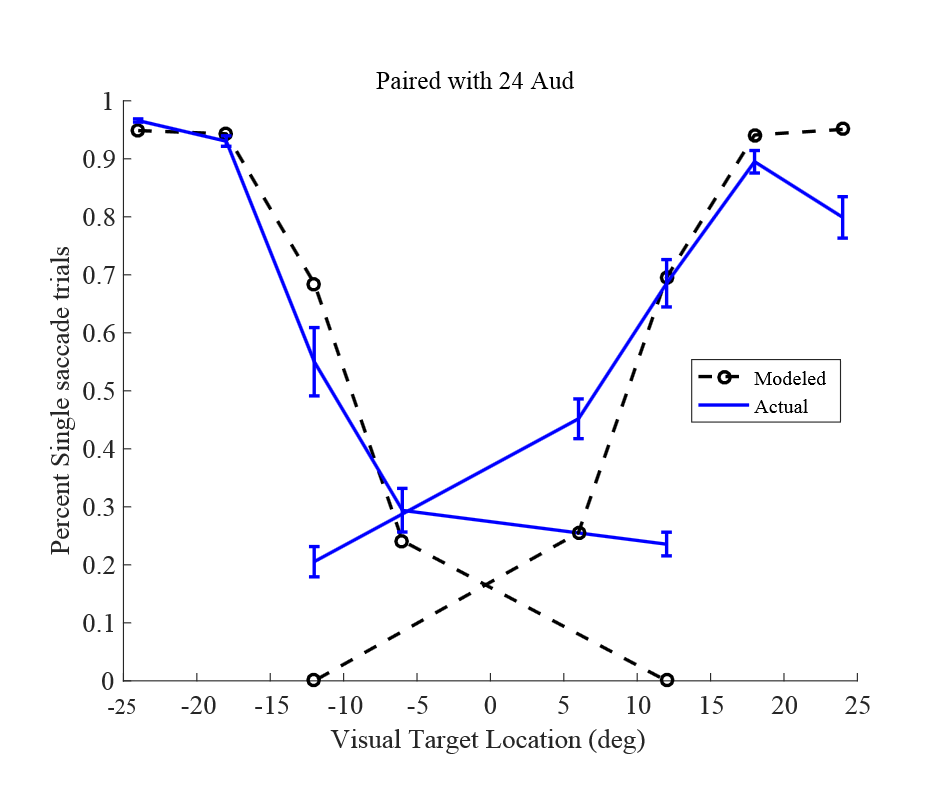
AV trial saccades

**B)**



**C)**

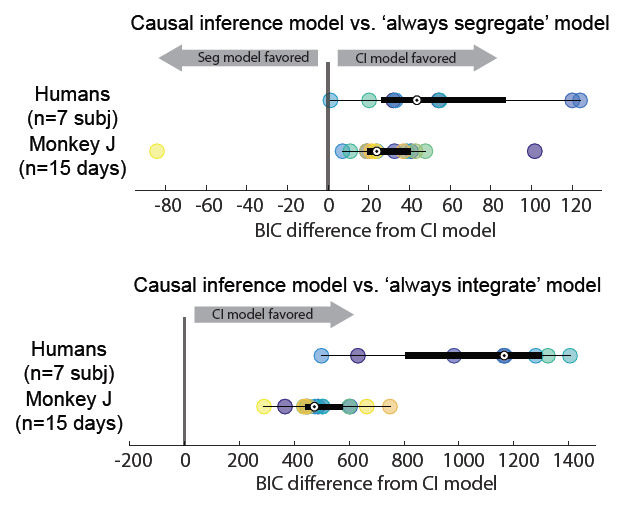


**Figure 2 XXX ugly placeholder:** *results of CI model fitting* (A) Model fits (lines) vs actual behavior (histograms) for conditions with matching auditory andvisual stimuli (B) model fits for a well separable stimulus pair. Note the bimodality of the modeled distribution. (C) number of saccades produced by generative model compared with actual behavior. This feature was not fit by the model but instead falls out automatically from the causal inference judgement

Qualitatively, these results suggest that monkeys are both integrating and segregating stimuli in a fashion that is consistent with causal inference. Monkeys display a mix of integration, visual capture of auditory stimuli, and complete segregation, and these behaviors depend on the degree of separation between targets. However, some important deviations are obvious. First the model greatly overestimates the variance of the combined stimulus condition in both the integrate and segregate conditions. This can be seen in the difference between predicted distributions (black lines) and actual distributions (grey histograms) in figure 2. The most likely explanation for this is the presence of erratic saccade behavior on a subset of trials (see examples in -6a-12v, where purple distribution has a very long tail). The way that models are fit disproportionately emphasizes these points, and getting rid of them is an obvious way to improve the current model. Another major difference is the disagreement between optimal and observed behavior for explicit CI as reported by number of saccades. An optimal observer would have conditions in which 2 saccades were made 100% of the time, whereas the monkey actually makes only one saccade in at least 20% of trials regardless of stimulus configuration (see figure 2C, dashed vs solid blue line). This is likely because of a behavioral lapse, in which the monkey did not initiate the second saccade quickly enough or was not paying close attention on that trial. This could potentially be compensated for by including a lapse parameter in the model fits.

**Quantitative evaluation of model XXX**

In order to determine if the CI model performed better than either of the potential simpler models considered (pure integration of stimuli or pure unisensory target estimation), a direct comparison using BIC was performed (figure 3). From this it can be seen that the CI model provides only a modest improvement over the complete segregation model (which assumes that the subject always makes two saccades, one to each target location). This is in line with the qualitative observation that the distributions generated by the CI and segregation models appear similar, with only a modest ‘filling in’ between two closely spaced unisensory distributions in the case of the CI model. However this model comparison depends only on the location report distributions, and does not take into consideration the explicit report of one vs two saccades. When comparing the number of saccades by target separation, it is clear that the CI model can provide an accurate account of the explicit causal inference, even when the implicit (localization) effects are less obvious.



**Figure 3:** *Results of model comparison against always integrate and always segregate strategies* BIC values were calculated for each model and compared with the BIC value for the CI model for each subject (humans, individual dots) or each individual day (monkey J, individual dots). Positive BIC values indicate that the CI model provides a better fit for the data after accounting for differences in number of parameters.

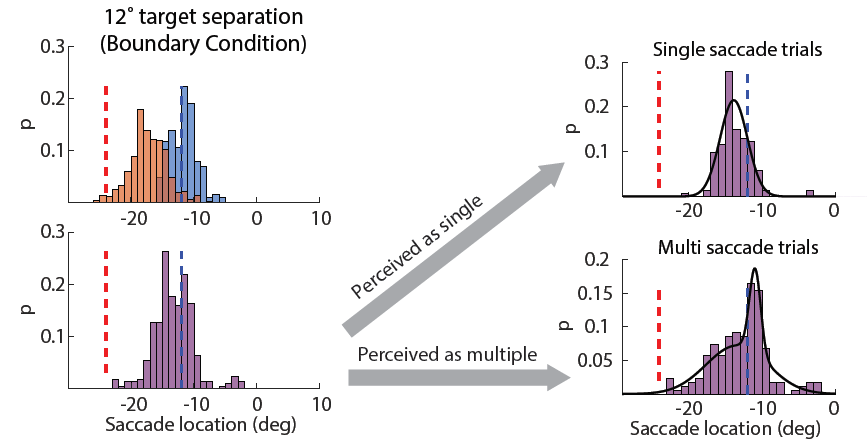
**Shortcomings of paradigm XXX**

**…** some of the things I would consider shortcomings are actually in previous sections of the results…

It is worth noting that because the visual and auditory stimuli are both well above perceptual threshold, the effects of multisensory integration are expected to be relatively weak (due to the principle of inverse effectiveness ref). Therefore the explicit causal inference, as reported through the number of saccades, is more informative than the degree of implicit causal inference (i.e., the amount of visual capture in the single saccade trials) when judging the differences between bound and separate percepts.

**Conclusion XXX**

**S1 Splitting the ‘ambiguous’ separation case by number of saccades**



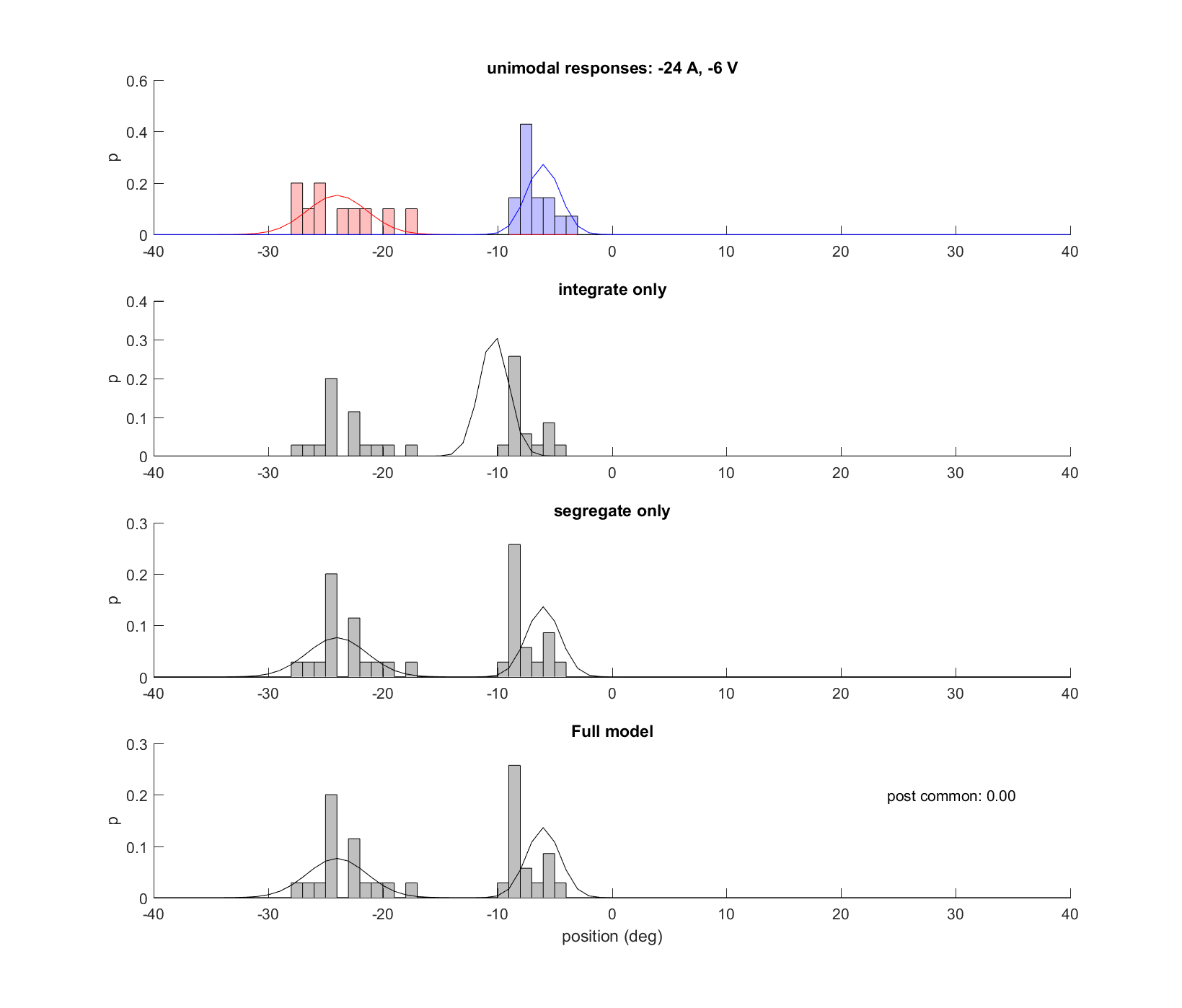
**S2: Evaluation of model fits vs actual data**

In order to work on the model fits, I created some plots which will show the distributions obtained through model fitting (averaged across the K-folds as described in methods), as well as the histograms of the actual saccade distributions. Here I have included many examples, but the main takeaway is that the model does a much much better job fitting the human data than the monkey data. I think the main reason for this is that the auditory distribution fits are really bad on the monkeys for some reason, so we end up with an auditory variance that is way higher than what you would expect given the data.

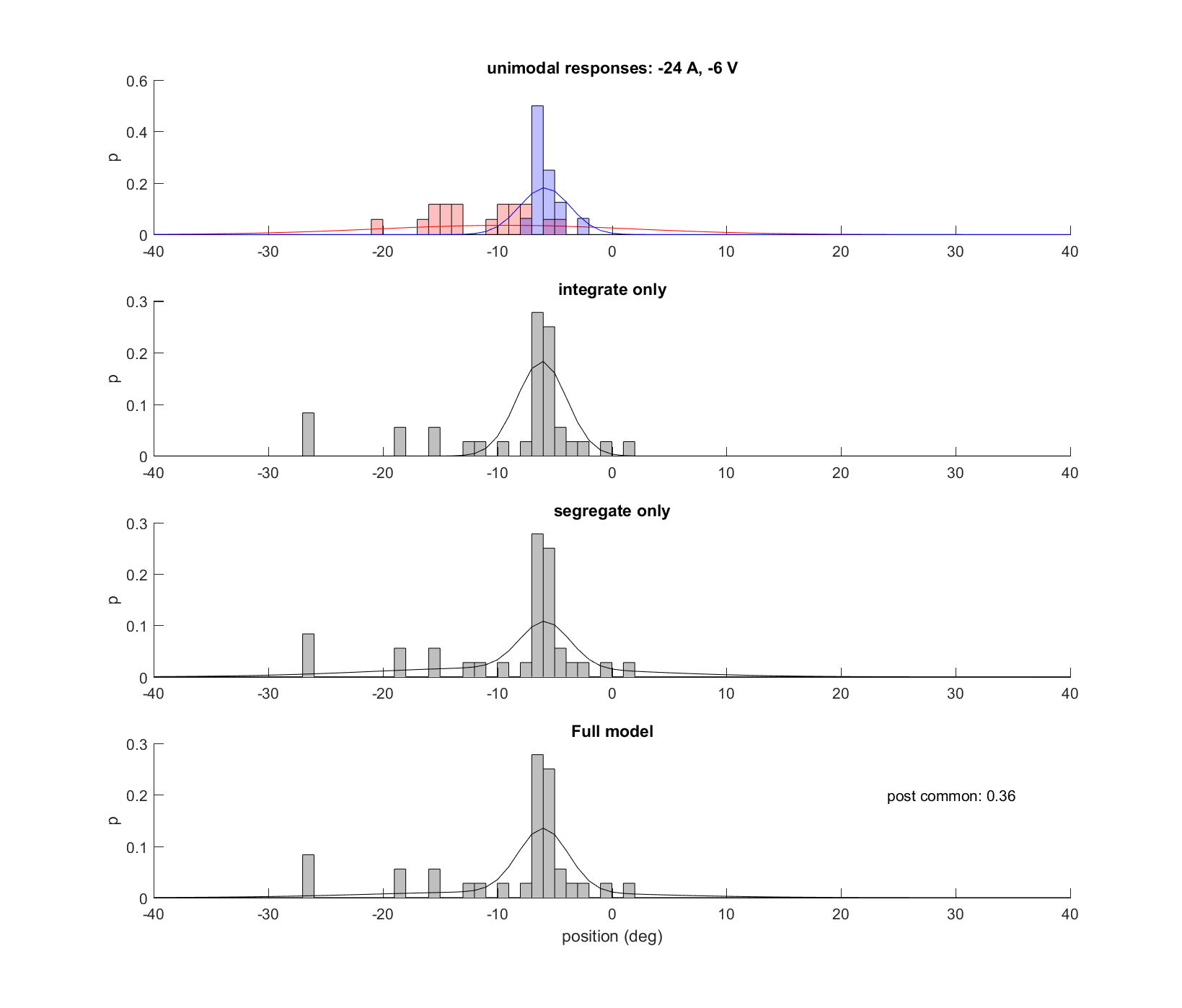
In the plots below, the unisensory trials are plotted in red (auditory) and blue (visual) histograms. The red and blue solid lines are the unimodal distributions *as estimated by the model fitting (i.e. fit on the dual saccade trials, not the unimodal trials)*. The black histograms are the actual saccades for the given trial type, while the black solid lines indicate model predicted distributions for the integrate, segregate, and CI models respectively.

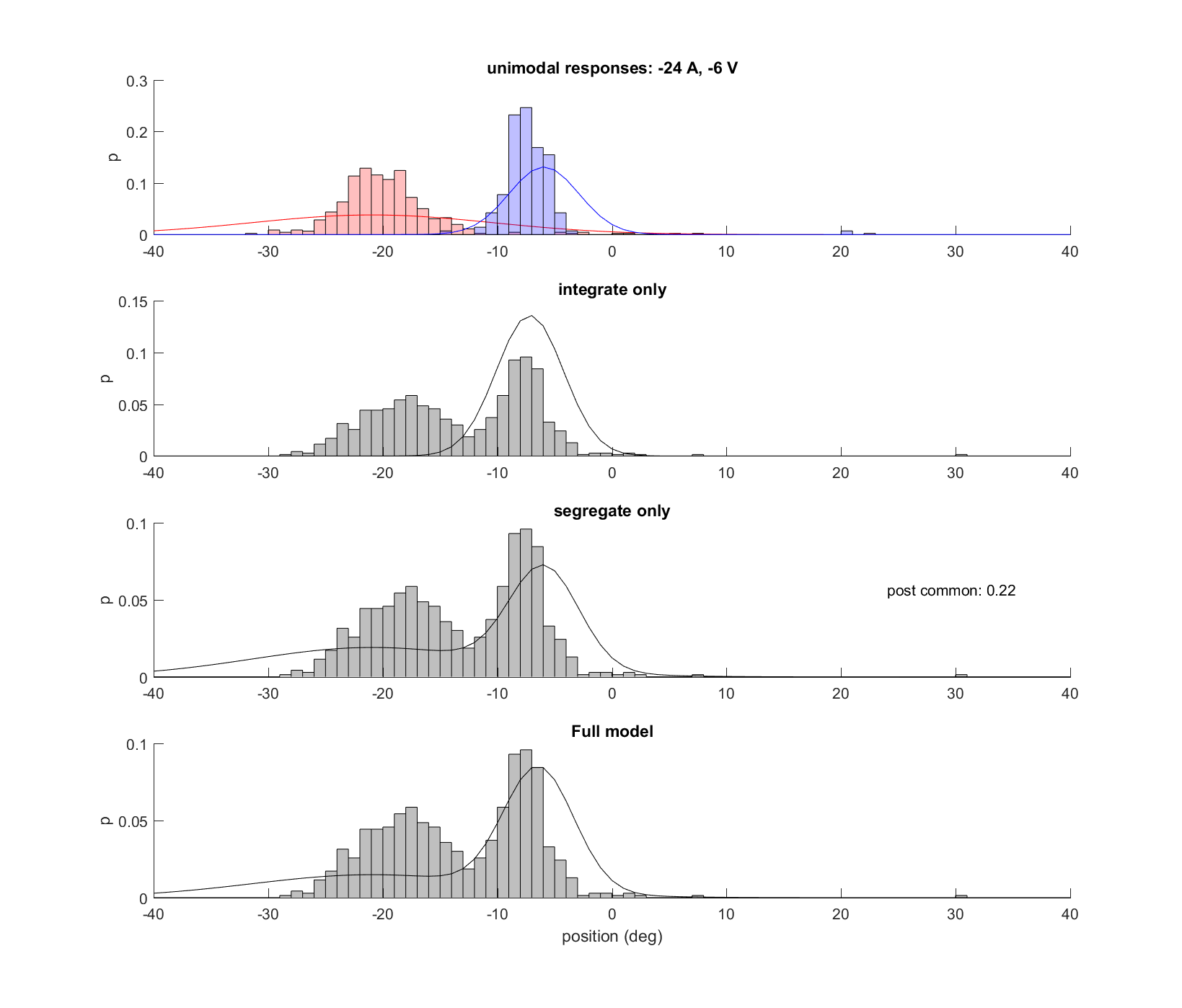
This data has been manipulated slightly by including an eyetracker bias correction step. This step uses a linear regression from the visually guided saccades to correct for any inaccuracies introduced by incorrectly adjusting the eye tracker during recording (eye tracker gain and offset are set manually for each subject during a short initial training phase of 20-50 trials).

**Human 08 – good aud localizer, good model fits**

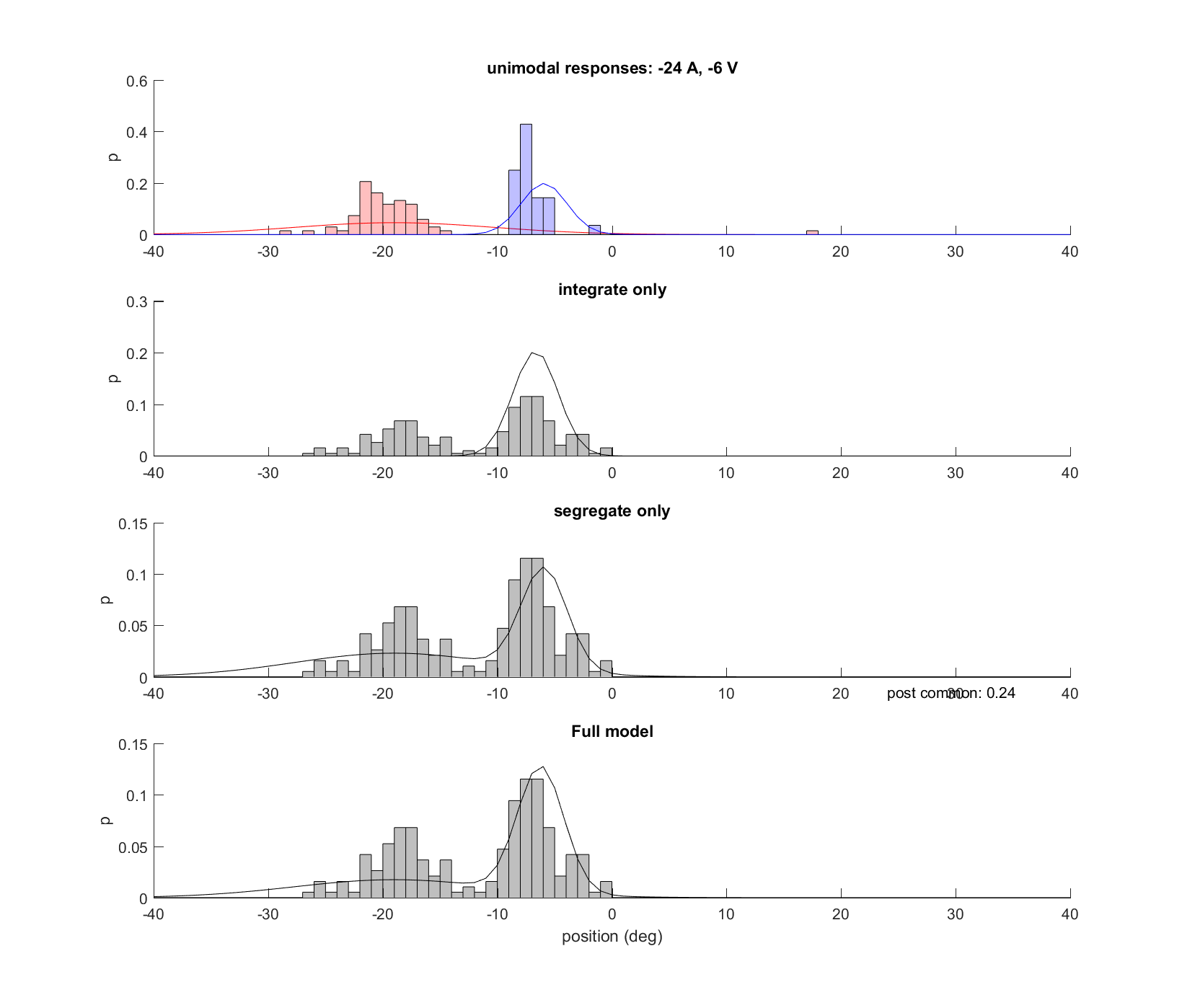


**Human 05 – bad aud localizer, bad model fits**



**Monkey J **

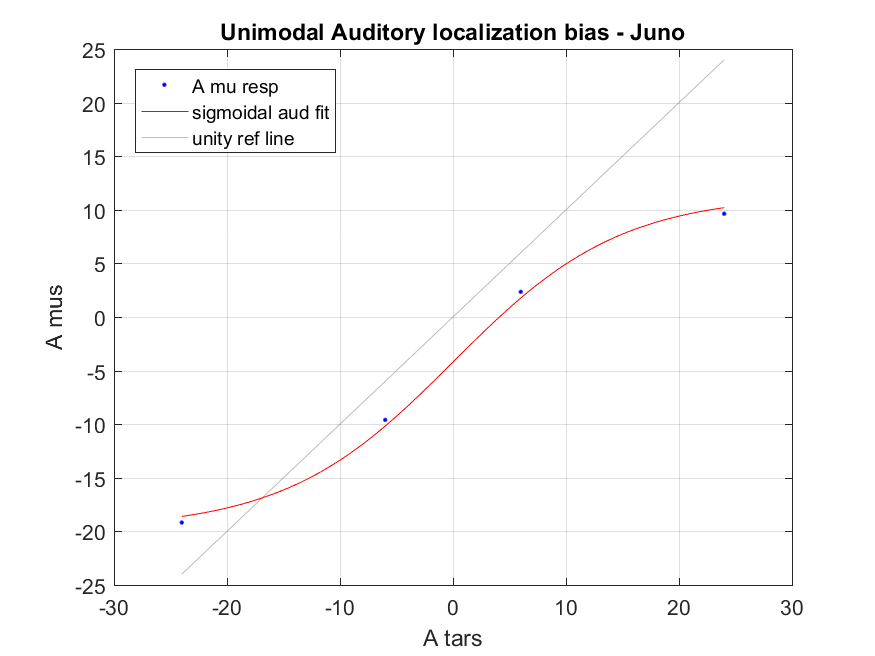
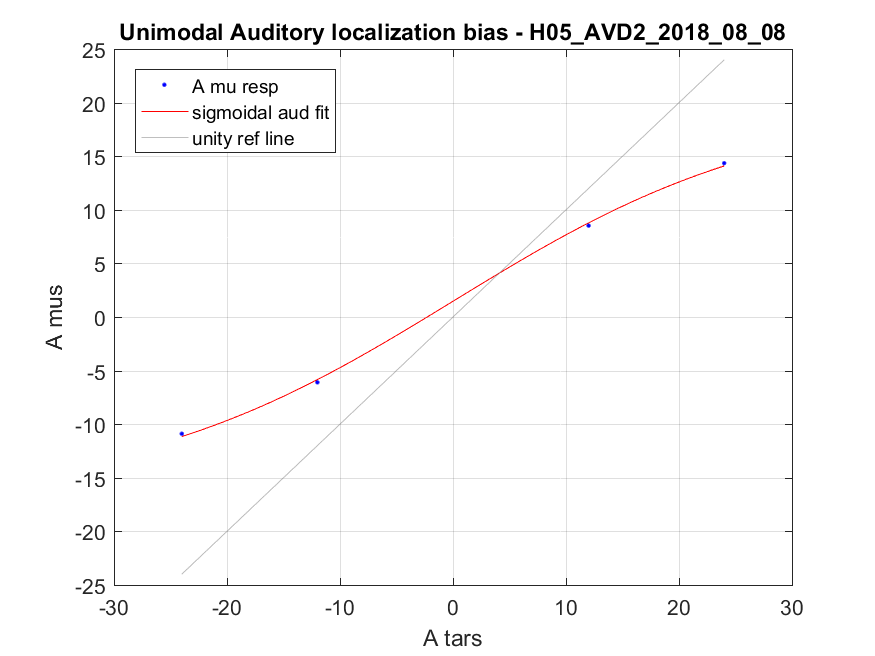
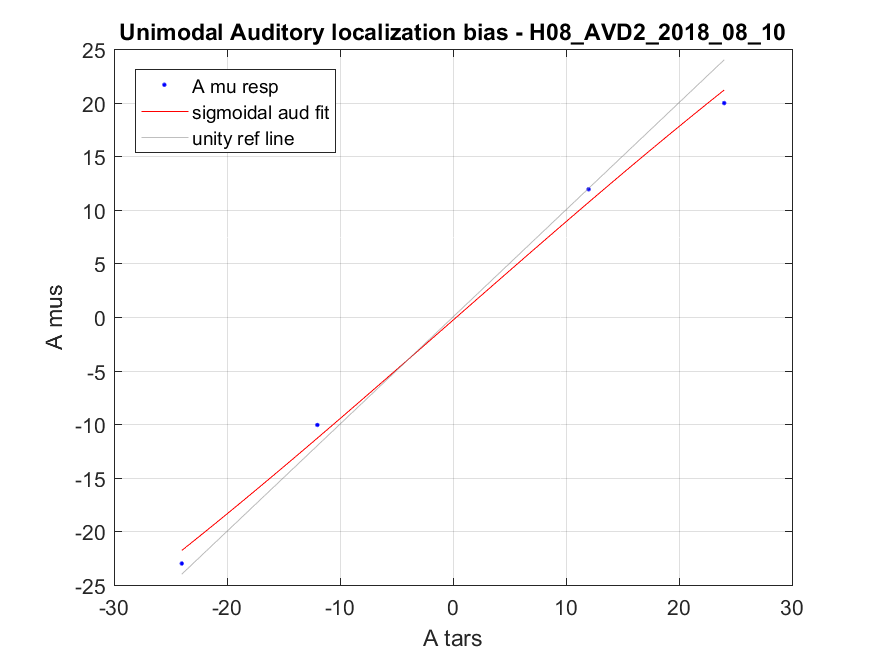
**Monkey Y**

****

**Auditory bias might be part of the problem**

One possible problem with these model fits is the presence of auditory localization bias. The model assumes that subjects are going to report locations that are centered on the actual target location, with the variance around that location being the only free parameter. However this doesn’t seem to be the case for some subjects. Comparing H08 to H05 below makes this very clear, as H05 is very bad at localizing the auditory stimuli even in the unisensory condition. Monkey J likewise has an auditory bias, although hers is both a shift and a compression (versus H05 which appears to be only a compression). It’s possible that these biases results in a very high variance estimate when doing the model fitting, because the model has no ability to account for this kind of systematic bias other than by increasing the variance to unrealistically high levels.

The alternative is that there are some errant saccades which are included in the analysis, and these might have a disproportionate effect on the model fitting because the distribution is assumed to be Gaussian.

****