Statistical Modeling and Analysis of SST-Pyr Synapses

Taisuke Yasuda 10/01/2016

1 Results

1.1 Summary

Our studies show that the observed data is best explained by a distribution of postsynaptic amplitudes over contacts with a high variance. We first find that the simple model obtained by assuming that the release probabilities, which we refer to as p, and response amplitudes, which we refer to as q, of each contact are the same does not sufficiently explain the data obtained from the experiment. We then find that when we take the response amplitudes across contacts to have a uniform distribution with large width parameter, we recover a model that explains the data well, as measured by QQ plots.

1.2 Assuming constant p and constant q at each contact

The first attempt at inferring the underlying biology from the observed data was to build a statistical model for a single trial (i.e. a single spike in a train of 10) assuming that the release probabilities and postsynaptic response amplitudes were all the same for every one of the N contacts that contribute to the observed summed amplitude. We constructed a model that depended on four parameters: the number of contacts N, a constant variance σ^2 at each contact, a constant release probability p at each contact, and a constant response amplitude q at each contact. Although these assumptions are not biologically reasonable for every synapse, they help us assess whether the frequency at which the extremely large amplitudes occur is simply a result of the structure of the statistical model or not.

1.2.1 Visualization of the model via histograms

We can start a first investigation of this model by visualizing simulations of the model with histograms. In this section, along with an initial assessment of the data and the statistical model, we develop methodologies that we will use later to assess a different model. In order to simulate the model, we estimate parameters using the data that we collected. When we assume that the release probabilities are all constant for each contact, we can estimate the release probability using the failure rate p_f, by

$$p = 1 - \sqrt[N]{p_f}.$$

Given this release probability and the sample mean $\hat{\mu}$ of the observed data, we estimate the response amplitude by

$$q = \frac{\hat{\mu}}{Np}.$$

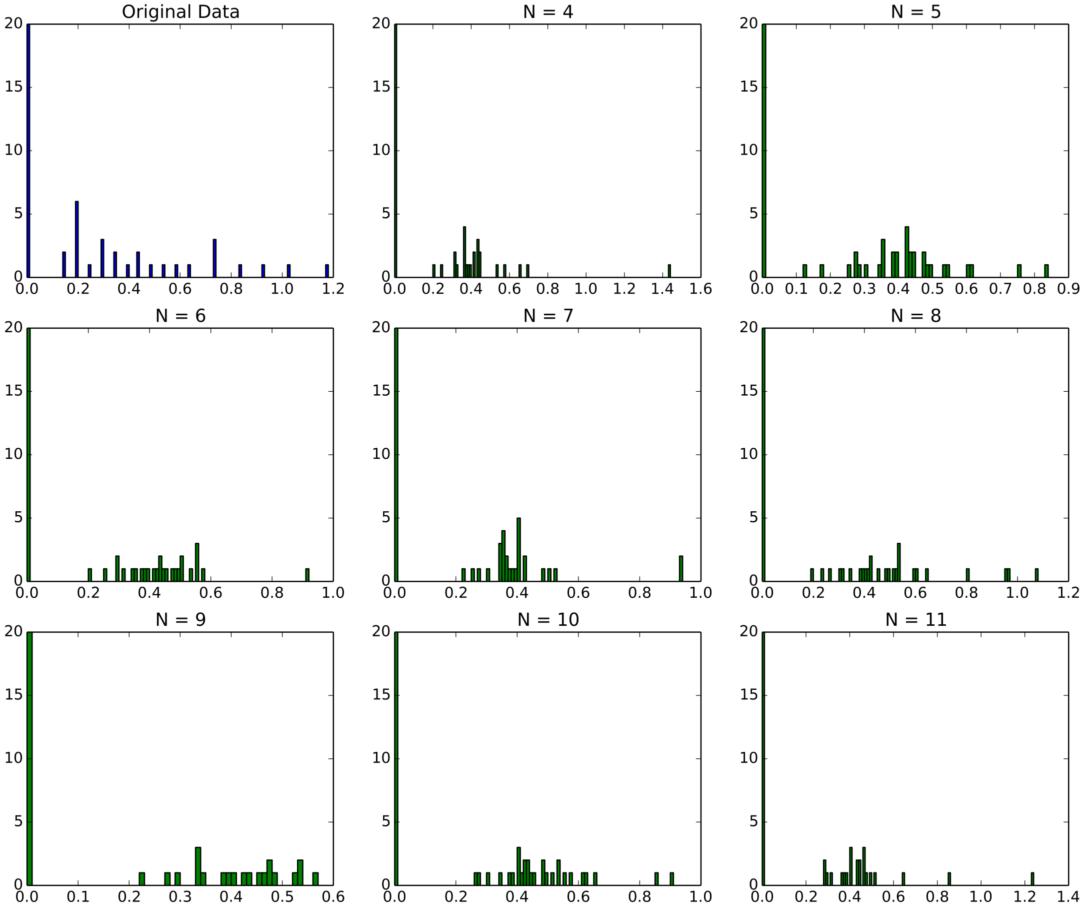
Finally, we estimate the parameter variance using the sample variance s^2 of the observed data by

$$\sigma^2 = \frac{1}{N}s^2.$$

The justifications for these estimations are shown in the methods section. Note that we do not make an estimation on N, yet all of our other estimations rely on knowledge of N. We attempted an estimation of N via clustering and fitting a mixture of Gaussians model, but the results were inconclusive. Thus, we simulate data for all N in the range $N = 4$ to $N = 11$, which is a reasonable range to assume given the range of possible N seen from reconstructions of the cells.

Using the parameters, we simulate the model and plot the results in a histogram. Figure 1 shows an example of a simulation that captures the shape of the observed data fairly well for some values of N, at least on first sight. However, goodness of fit is difficult to assess just with the human eye. Therefore, we use QQ plots to assess the goodness of fit of these models.

Figure 1: Simulations assuming constant parameters



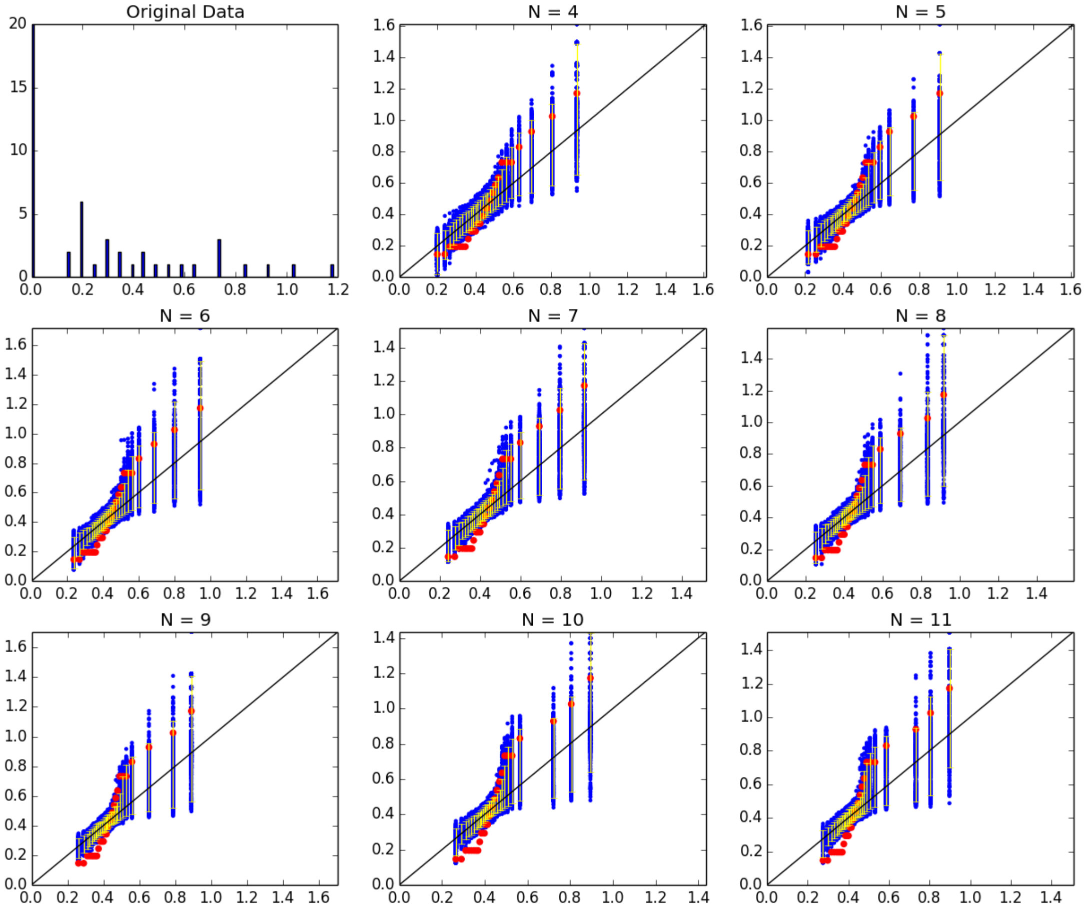
1.2.2 Visualization via QQ plots

QQ plots allow us to directly compare the observed data with our proposed model by sampling from the proposed model many times and comparing the quantiles of this sampling with the quantiles of the observed data. Because the simulated percentage of zeros matches the percentage of zeros of the observed data by construction, we will just analyze the nonzero values (Figure 2).

In this figure, we show the QQ plot for the same cell as Figure 1, with the same range of values for N. In the QQ plot, the blue points indicate the simulated values from the constant parameter model, the yellow points indicate the 95% confidence band, and the red points indicate the observed data, all plotted against the quantiles of the simulated values. With this analysis, we see that though the largest events of the observed data are towards the high end of the simulated values, they are still within the 95% confidence band. However, the lower values are outside of the 95% confidence band, so we cannot say that this model truly explains the data well: the problem is to explain the high amplitude events in relation to the rest of the data.

Upon inspecting the QQ plots for other cells, we see that high amplitude events are actually explained fairly well. However, in general, this model seems to sacrifice the goodness of fit of the events at lower amplitudes, consistently overestimating the value of the quantile. Intuitively, we should have expected this result: because we set all the contacts to have the same amplitude, the probability of the occurrence of these particular amplitudes and their multiples becomes higher.

Figure 2: QQ plots for constant parameter model



1.3 Assuming a uniform distribution on q

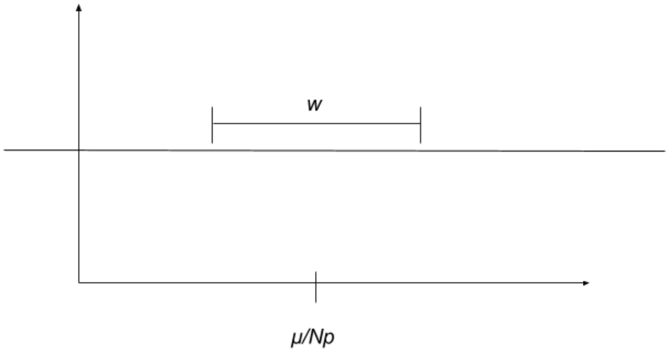
Apparently, setting everything as constant does not explain the data sufficiently, as shown by the QQ plot analysis. The next simplest assumption we could make is to assume some other distribution on p or q while keeping the other constant. By perturbing the parameters of these assumed distributions and seeing how the resulting distribution of $\boldsymbol{A}$ changes, we can possible gain more understanding on the relationship between the parameters and the resulting distribution. Thus, we assume a uniform distribution on q, where we assume that the q_js are drawn from a range $[a, b]$ with equal probability. To simplify this exploration, we will simply draw N equidistant q_js from an assumed range. We should still satisfy the constraints given by the failure rate p_f and sample mean $\hat{\mu}$, so we derive expressions for q_j that comply to these constraints. Because we kept p_j all constant, we can simply take the choice from before, where $p_j = p = 1 - \sqrt[N]{p_f}$ for all $1 \leq j \leq N$. To make a suitable choice for q_j, we require that the mean of the uniform distribution match the mean of the observed data and choose the distribution to have a width w, which is a parameter we control.

To make a suitable choice for q_j, first note that:

$$\bar{\boldsymbol{A}} = \sum_{j=1}^{N} p_j q_j = p \sum_{j=1}^{N} q_j \iff \frac{1}{N} \sum_{j=1}^{N} q_j = \frac{\bar{\boldsymbol{A}}}{Np}$$

Then, if we decide the q_js to span a width $0 \leq w < 2\frac{\bar{\boldsymbol{A}}}{Np}$, then we can choose $q_j = \frac{w}{N-1} j + \frac{\bar{\boldsymbol{A}}}{Np} - \frac{w}{2}$ for $j \in$

Figure 3: Choosing the q_j from a distribution with mean $\hat{\mu}/Np$ and width w.



$\{0, 1, ..., N-1\}$.

Now, we will take a random trial of a random cell with a random N to see how imposing this structure on q_j will affect the resulting QQ plot for various values of w (Figure 3).

In Figure 4, we actually see that as we take the width of the uniform distribution w to be larger and larger, we bring the observed distribution closer and closer to the simulated distribution! In fact, for large enough values of w, this puts the observed data within the 95% confidence intervals of the simulated data. This effect was seen in general across all trials and guesses for Ns for this particular cell. The observation that a particular effect of changing the qs on the distribution would hold across trials is not surprising, as we expect the qs of a cell to stay the same across trials. However, because this effect holds across our guesses for Ns of a cell, we conclude that this is a fundamental characteristic of the data. When we try these simulations on more cells, this effect holds for them as well, sometimes achieving particularly excellent results (Figure 4).

The existence of these excellent fits suggest that this is an optimal method of explaining the data – we expect that fitting the data even better is just a matter of tuning the parameters initialized from the equidistant q_js with width w. In Figure 5, we show what the simulations of these fits look like on the same cell as before.

2 Methods

2.1 Underlying statistical model of one trial

We first build a statistical model based on the underlying biology of the observed measurements:

$$\boldsymbol{A} = \sum_{j=1}^{N} \boldsymbol{A}_j, \boldsymbol{A}_j = \begin{cases} N(q_j, \sigma_j) & \text{if } \boldsymbol{S}_j = 1 \\ 0 & \text{if } \boldsymbol{S}_j = 0 \end{cases}, \boldsymbol{S}_j = \text{Bern}(p_j)$$

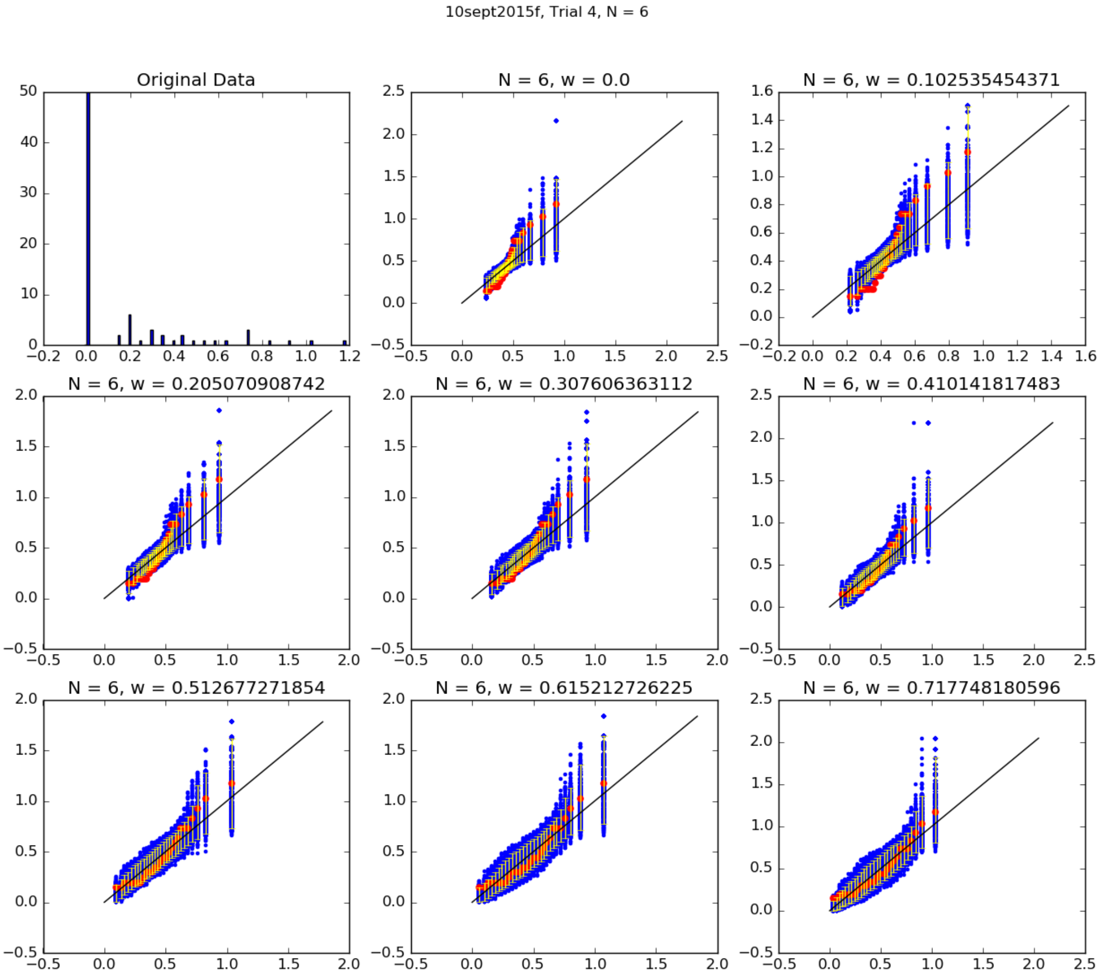
In this model, $\boldsymbol{A}$ represents the overall response amplitude that we observe and is the sum of $\boldsymbol{A}_j$ where j takes on integers from 1 to N. Each $\boldsymbol{A}_j$ represents the response amplitude of an individual synaptic contact and j indexes the synaptic contacts. The response of an individual synaptic contact depends on its release success, so we model release success with a latent Bernoulli variable $\boldsymbol{S}_j$ with parameter p_j representing the release probability of the particular contact. If contact j succeeds in release, e.g. $\boldsymbol{S}_j = 1$, then $\boldsymbol{A}_j$ is normally distributed with mean q_j and standard variation σ_j. If contact j fails to release, e.g. $\boldsymbol{S}_j = 0$, then $\boldsymbol{A}_j = 0$.

2.2 Initial analysis of the model

2.2.1 Trying all constant

A simple way to visualize this statistical model is to assume some number of contacts N, assume some uniform standard deviation σ, mean response q, and release probability p for every one of the N contacts. Under these assumptions, we can make reasonable estimate the parameters p, q, and σ from the data.

Figure 4: QQ plots for uniformly distributed q, varying the distribution width



2.2.2 Estimation of Parameters

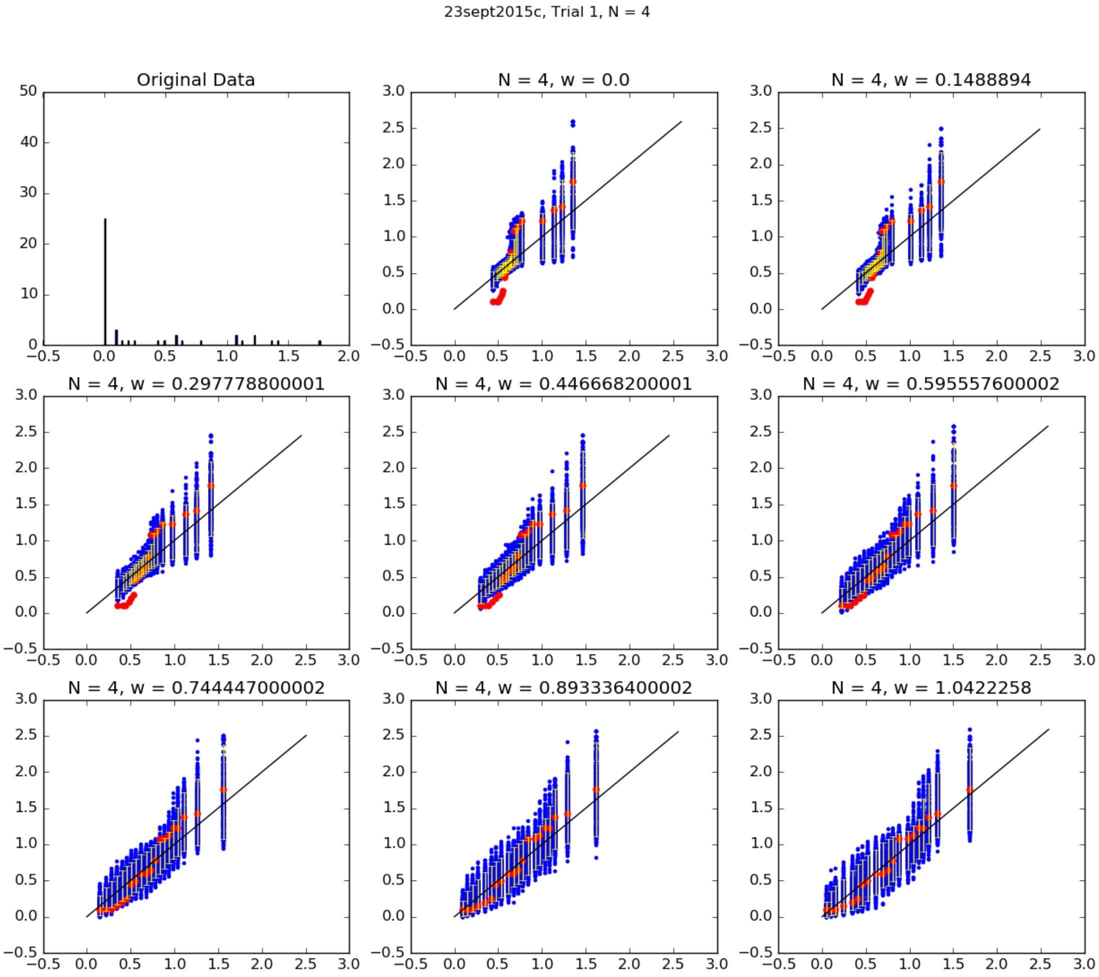
We first estimate p using the observed failure rate p_f. If a cell responds with an amplitude of 0, then this happens exactly when all of its N contacts fail to release. Then, since all the contacts have the same release probability p under our assumptions, we conclude the following value for p:

$$p_f = (1-p)^N \iff \sqrt[N]{p_f} = 1-p \iff p = 1 - \sqrt[N]{p_f}$$

Next, we estimate q using the mean response $\bar{\boldsymbol{A}}$. We can compute the expectation of our observed random variable $\boldsymbol{A}$ as follows:

$$\begin{aligned}\mathbf{E}[\boldsymbol{A}] &= \mathbf{E}[\sum_{j=1}^{N} \boldsymbol{A}_j] = \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j] \\ &= \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j \mid \boldsymbol{S}_j = 1]\mathbf{Pr}\ [\boldsymbol{S}_j = 1] + \mathbf{E}[\boldsymbol{A}_j \mid \boldsymbol{S}_j = 0]\mathbf{Pr}\ [\boldsymbol{S}_j = 0] \\ &= \sum_{j=1}^{N} \mathbf{E}[N(q,\sigma)]p + 0 = \sum_{j=1}^{N} qp = Npq\end{aligned}$$

Figure 5: An example of an excellent fit



Since we have assumed N and estimated p and $\bar{\boldsymbol{A}} \approx \mathbf{E}[\boldsymbol{A}]$, we have an estimate for q. Finally, we make an estimate on σ using the variance of the distribution $\mathbf{Var}(\boldsymbol{A})$. Since we have that each of the contacts are independent of one another, by the linearity of variance of independent random variables, we have that $\mathbf{Var}(\boldsymbol{A}) = \sum_{j=1}^{n} A_j$. In particular, if σ is constant for all the contacts, then $\sigma^2 = \frac{1}{N}\mathbf{Var}(\boldsymbol{A})$

2.3 Assuming distributions on q

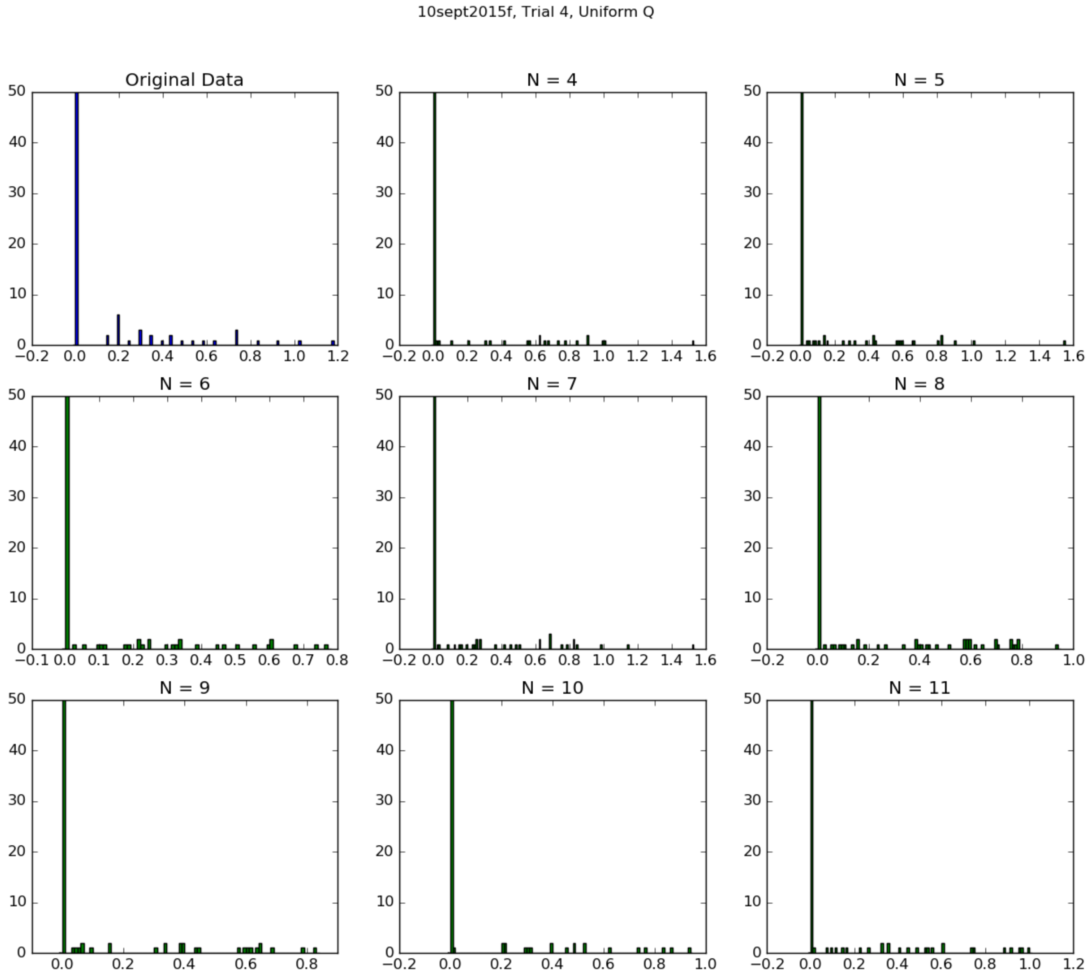
Apparently, setting everything as constant does not explain the data sufficiently, as shown by the QQ plot analysis. The next simplest thing we could do is to assume some other distribution on p or q while keeping the other constant. By perturbing the parameters and see how the resulting distribution of A changes, we can possible gain more understanding on the relationship between the parameters and the resulting distribution.

In doing this analysis, it is actually much easier to perturb q while keeping p constant, since we don't have the constraint of p_f – we can in fact find closed form solutions to the values of q_j we should use for very simple cases.

2.3.1 Uniform Distribution

The simplest distribution that we could assume on q is the uniform distribution, where we assume that the q_is are drawn from a range $[a, b]$ with equal probability. To simplify this exploration, we will simply draw N equidistant q_js

Figure 6: Simulations of a successful fit



from an assumed range. We should still satisfy the constraints given by p_f and $\bar{\boldsymbol{A}}$, so we derive expressions for q_j that comply to these constraints. Because we kept p_j all constant, we can simply take the choice from before, where $p_j = p = 1 - \sqrt[N]{p_f}$ for all $1 \leq j \leq N$. To make a suitable choice for q_j, first note that:

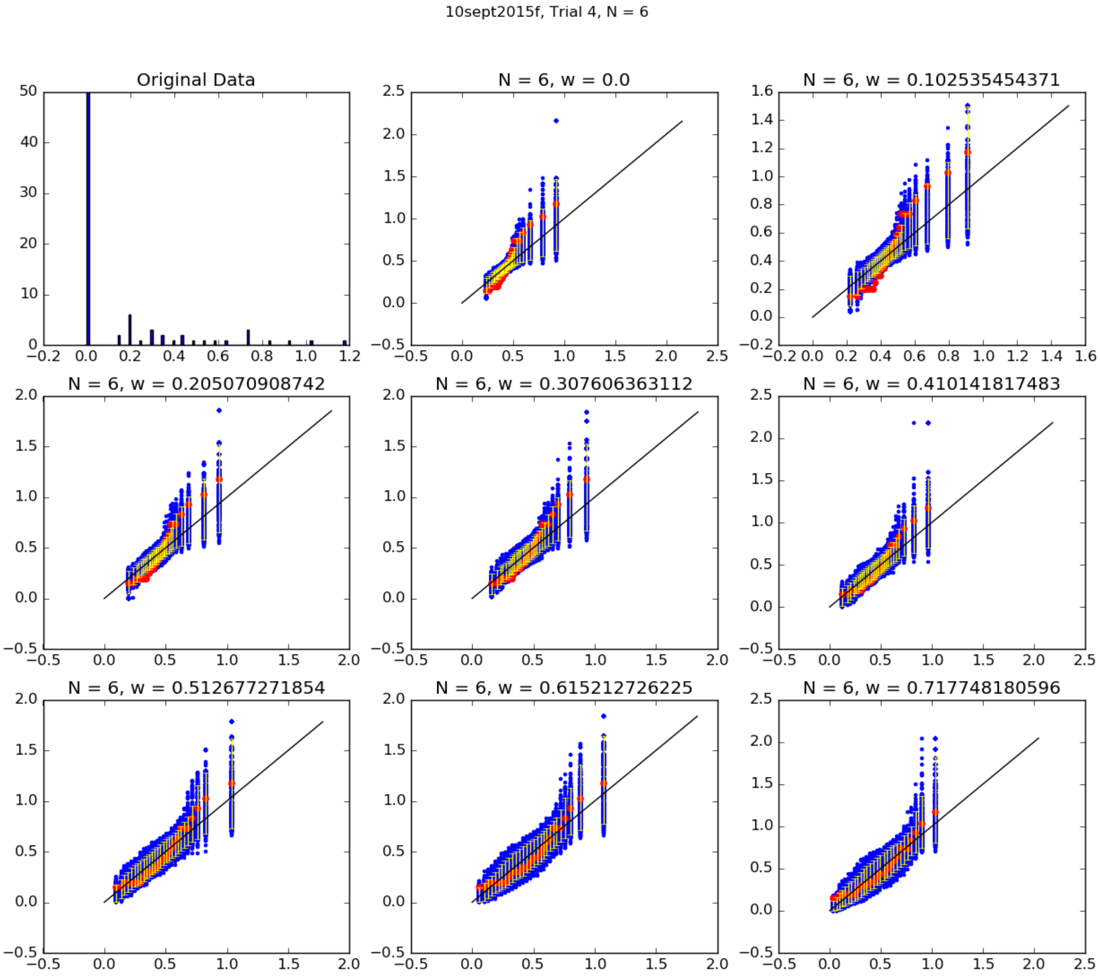
$$\bar{\boldsymbol{A}} = \sum_{j=1}^{N} p_j q_j = p \sum_{j=1}^{N} q_j \iff \frac{1}{N} \sum_{j=1}^{N} q_j = \frac{\bar{\boldsymbol{A}}}{Np}$$

Thus, we just need to choose q_j to be an arithmetic progression with mean $\frac{\bar{\boldsymbol{A}}}{Np}$. If we decide the q_js to span a width $w < 2\frac{\bar{\boldsymbol{A}}}{Np}$, then we we choose $q_j = \frac{w}{N-1} j + \frac{\bar{\boldsymbol{A}}}{Np} - \frac{w}{2}$ for $j \in \{0, 1, ..., N-1\}$. Note that when $j = 0$, $q_j = \frac{\bar{\boldsymbol{A}}}{Np} - \frac{w}{2}$ and when $j = N-1$, $q_j = \frac{\bar{\boldsymbol{A}}}{Np} + \frac{w}{2}$.

Now, we will take a random trial of a random cell with a random N to see how imposing this structure on q_j will affect the resulting QQ plot for various values of w (Figure 3).

In Figure 3, we actually see that as we take the width of the uniform distribution w to be larger and larger, we bring the observed distribution closer and closer to the simulated distribution! In fact, for large enough values of w, this puts the observed data within the 95% confidence intervals of the simulated data. This effect was seen in general across all trials and guesses for Ns for this particular cell. The observation that a particular effect of changing the qs on the distribution would hold across trials is not surprising, as we expect the qs of a cell to stay the same across trials. However, because this effect holds across our guesses for Ns of a cell, we conclude that this is a fundamental

Figure 7: QQ plots for uniformly distributed q, varying the distribution width



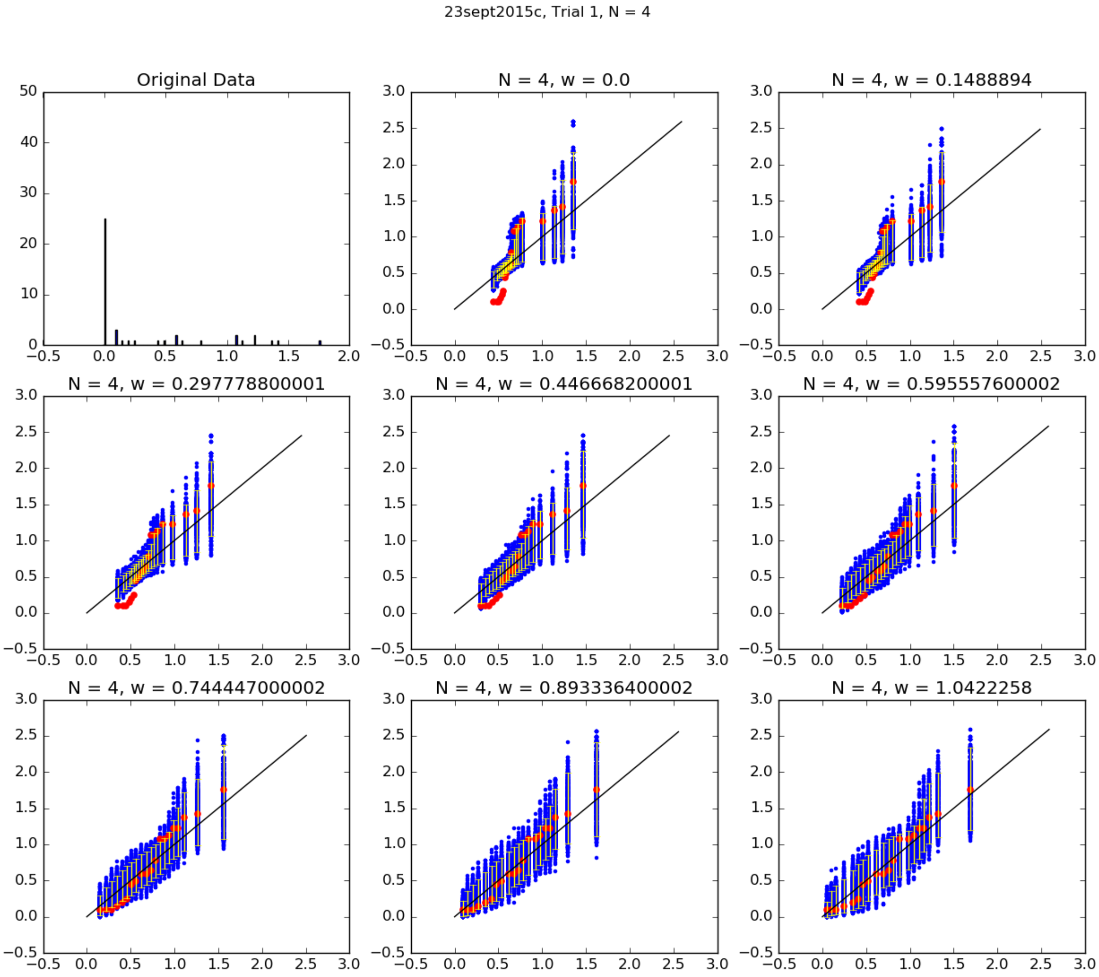
characteristic of the data. When we try these simulations on more cells, this effect holds for them as well, sometimes achieving particularly excellent results (Figure 4).

The existence of these excellent fits suggest that this is an optimal method of explaining the data – we expect that fitting the data even better is just a matter of tuning the parameters initialized from the equidistant q_js with width w. In Figure 5, we show what the simulations of these fits look like on the same cell as before.

2.4 Statistical model for the entire experiment

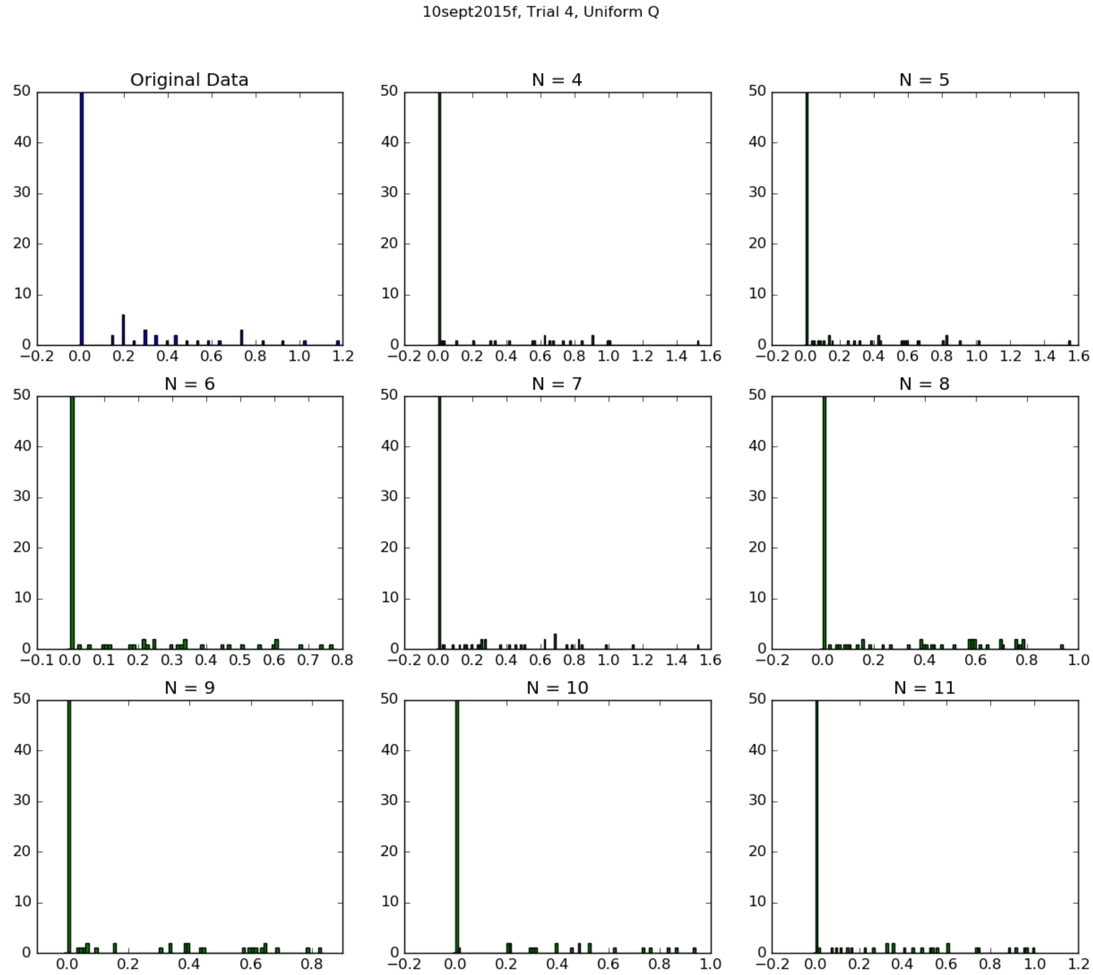
At this point, we are fairly confident that we can explain single trials of events. However, we have further constraints to satisfy – for the trials that we know belong to the same cell, we know that the q_js should be the same. Thus, we should be able to find a single set of q_js for all of the trials for each cell. If we keep our assumption that p is constant for each trial, then we can find this value in exactly the same way as we did for the analysis of single trials. We proceed as follows: we first estimate the mean value of q_j from the whole data set as before, with the formula $\frac{1}{N}\sum_{j=0}^{N} q_j = \frac{\bar{\boldsymbol{A}}}{Np}$. We then find the q_j found by the arithmetic progression and repeat the QQ plot measurement for each trial with their own p values, estimated as before with the formula $1 - \sqrt[N]{p_f}$.

Figure 8: An example of an excellent fit



3 Technical Appendix

Figure 9: Simulations of a successful fit


10sept2015f, Trial 4, Uniform Q
Original Data
N = 4
N = 5
N = 6
N = 7
N = 8
N = 9
N = 10
N = 11