# Kernel Probability Estimation For Binomial And Multinomial Data

**Greg Jensen**[1]

[1]**Columbia University**

## ABSTRACT

Kernel-based smoothers have enjoyed considerable success in the estimation of both probability densities and event frequencies. Existing procedures can be modified to yield a similar kernel-based estimator of instantaneous probability over the course of a binomial or multinomial time series. The resulting nonparametric estimate can be described in terms of one bandwidth per outcome alternative, facilitating both the understanding and reporting of results relative to more sophisticated methods for binomial outcome estimation. Also described is a method for sample size estimation, which in turn can be used to obtain credible intervals for the resulting estimate given mild assumptions. One application of this analysis is to model response accuracy in tasks with heterogeneous trial types. An example is presented from a study of transitive inference, showing how kernel probability estimates provide a method for inferring response accuracy during the first trial following training. This estimation procedure is also effective in describing the multinomial responses typical in the study of choice and decision making. An example is presented showing how the procedure may be used to describe changing distributions of choices over time when eight response alternatives are simultaneously available.

Keywords: proportions, time series analysis, kernel estimation, nonparametric methods, binomial data, multinomial data

Rates underlying empirical data often do not correspond nicely with the forms of parametric functions. Just as frequencies often fail to be distributed in a precisely Gaussian manner, so too do rates often fail to change in strictly linear or logistic ways. Biological processes (whether they be the firing of individual neurons or the decision-making of organisms) are particularly irregular in this respect. Consequently, there are considerable benefits to nonparametric methods for estimating rates, as these permit inferences to be made about data without imposing inappropriate assumptions. In light of this need, a vast literature of nonparametric methods for such estimation has arisen, much of which concerns kernel-based procedures that convert discrete clusters of observations into smoothed estimates (Rosenblatt, 1956; Parzen, 1962).

Kernel-based smoothing replaces each observation with a density function (or ‘kernel’), the most common of which are Gaussian. This replacement has the effect of smearing each observation across the measurement scale. The resulting collection of density functions are summed, yielding the smoothed estimate.

In the simplest case, the estimate depends on a single parameter: The kernel’s *bandwidth*, which governs its dispersion. The standard deviation, for example, is the bandwidth for Gaussian distributions. Remarkably, so long as each kernel is symmetric and unimodal, the shape of the kernel matters much less to the overall accuracy of the estimate than does the bandwidth. Consequently, a great deal of effort has been devoted to bandwidth selection (Hall and Marron, 1991; Jones et al., 1996).

It is important to distinguish between *kernel density estimation* and *kernel rate estimation*, as this distinction has an impact on the bandwidth selection procedure. Density estimation aims to estimate a

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1156v1 | CC-BY 4.0 Open Access | rec: 2 Jun 2015, publ: 2 Jun 2015

probability density whose integral is 1.0, whereas rate estimation aims to make instantaneous estimates, at any point in a time series, of the frequency of events per unit time. Rate estimation is of particular importance to neuroscience, where it is used to obtain the firing rates of individual neurons (Dayan and Abbott, 2001; Shimazaki and Shinomoto, 2010). Such procedures can also be used to estimate the rates of other events, such as button presses, heart beats, or frequency of base pairs over a length of DNA.

## PROBABILITY ESTIMATION IN TIMES SERIES

Binomial data (e.g. correct/incorrect responses) reflect a different kind of frequency than those typically examined by rate estimation procedures. Given a times series of yes/no responses, one might reasonably wish to estimate the *probability* of a correct response at each time point, rather than instantaneous estimates of the frequency of responses. Importantly, unlike frequencies in time, probabilities should *not* vary as a function of overall response density: If the probability is 0.5, then the estimate should approximate 0.5 regardless of whether few responses or many are expected to occur at that moment.

This problem is especially pressing when behavior is not only intermittent, but also occurs at different times in different sessions. For example, in a 'transitive inference' procedure, subjects are shown pairs of stimuli from an ordered set in a randomized order. In one session, a subject might see the pair $AB$ on the first trial and the pair $BC$ on the second trial, whereas $AB$ might not be seen until the fourth trial in a subsequent session. Kernel smoothing, in principle, should allow a pooling of information from $AB$ trials over time, even if a subject sees that pair during a different trial in every session.

A very powerful approach for solving this kind of problem is that of *generalized additive models* (or "GAMs," Hastie and Tibshirani, 1986). These procedures combine local regression techniques (similar to LOESS regression, Cleveland, 1979) with generalized linear models (Nelder and Wedderburn, 1972), which in turn permits localized smoothing using logistic regression. This yields a continuous estimation of a probability.

However, GAMs present analysts with several drawbacks. One difficulty is that GAMs are sophisticated models, and well beyond the ability of many researchers to implement. Another is that much of the machinery intrinsic to fitting GAMs relies on numerical approximation, which is implemented differently by different software packages. Furthermore, many implementations of GAM estimation are prone to overfitting, a tendency whose correction requires further procedures (Wood, 2008). Thus, although GAMs may currently represent the gold standard for semiparametric rate estimation in binomial data, there is nevertheless a need for a rate estimation procedure whose workings can be understood by a broad audience.

## KERNEL PROBABILITY ESTIMATION

In order to make an estimation of the probability of success or failure at any time point, all that is required is a kernel rate estimate of successes and a separate kernel rate estimate of failures. This procedure is depicted in Figure 1.

Let $T$ represent the full span of all times $t$ observed during the experiment, whereas $R$ represents the set of responses $r_i$. Correct trials will be denoted with a subscript plus ($R_+$), whereas incorrect trials will be denoted with a subscript minus ($R_-$). Let $N$ represent the total number of responses observed, whereas $n$ represents the total number of sessions. Given this notation, the kernel rate estimates for correct

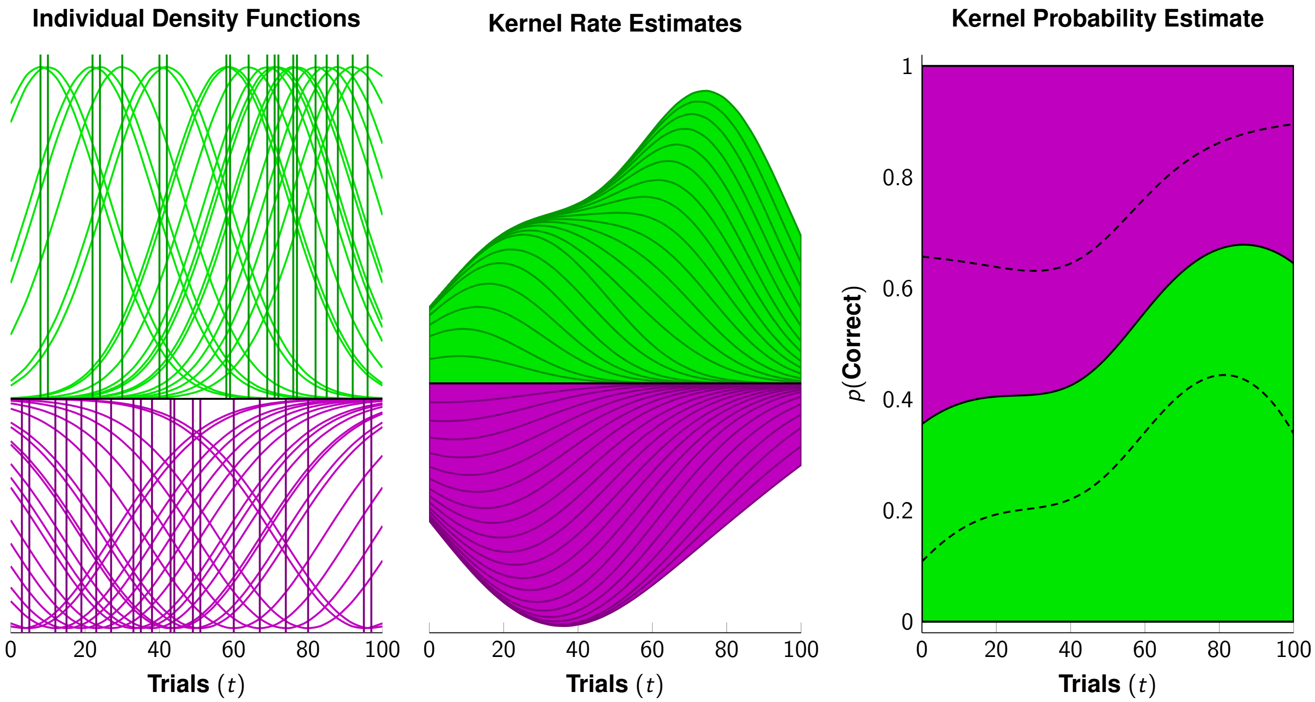

**Figure 1.** Visualization of the kernel probability estimation procedure. **(Left)** Correct responses (in green) and incorrect responses (in purple) are treated as independent time series, in which each is assigned a Gaussian distribution with an optimized bandwidth. This replaces the observed response times (vertical lines) with Gaussian density functions. **(Middle)** The Gaussian kernels are summed to yield kernel estimates of the instantaneous rate of each event over time, per Equation 1. **(Right)** The relative rate of events is calculated for each time point, per Equation 2. In addition, the 95% credible interval for the estimate is calculated using Equations 3 and 4 and shown as dashed lines.

outcomes ($K_+$) and incorrect outcomes ($K_-$) are as follows:

$$\begin{aligned} K_+(t) &= \sum_{r \in R_+} \mathcal{N}(t|r, \omega_+) \\ K_-(t) &= \sum_{r \in R_-} \mathcal{N}(t|r, \omega_-) \end{aligned} \tag{1}$$

Here, $\mathcal{N}(t|r,\omega)$ denotes the density of a normal distribution at time $t$ with a mean of $r$ and a standard deviation of $\omega$. $\omega_+$ and $\omega_-$ represent the optimal bandwidths for the rate estimates of correct and incorrect trials, respectively. An efficient procedure for optimal bandwidth selection is described by Shimazaki and Shinomoto (2010). The Gaussian implementation of this method, with a small correction for continuity, is detailed in the appendix. Ordinarily, $K_+$ and $K_-$ would be scaled by a factor of $\frac{1}{n}$ to average across multiple sessions. In this case, the scaling factor is omitted because it cancels out in the next operation.

The kernel probability estimate ($KPE$) is obtained by computing the relative proportions of $K_+$ and $K_-$:

$$KPE(t) = \frac{K_+(t)}{K_+(t) + K_-(t)} \tag{2}$$

Thus, although the estimated rates of $K_+$ and $K_-$ may rise and fall (because of uneven sampling of data, or because of biased estimates near the edges of the observed interval), their *relative* rates may nevertheless be compared throughout.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1156v1 | CC-BY 4.0 Open Access | rec: 2 Jun 2015, publ: 2 Jun 2015

It is also important to estimate a ‘credible interval’ for the estimated proportion parameter. This can be accomplished using the Jeffreys interval (Brown et al., 2001), which performs well under both frequentist and Bayesian interpretations of uncertainty.

The estimate of the credible interval depends on the approximate number of observations contributed to each estimate (i.e. “On how many points does the estimate at time $t$ depend?”). This can be accomplished by converting the normal distributions that act as kernels into unscaled Gaussian functions (whose mode has a value of 1.0 regardless of bandwidth). Since this requires only the cancelation of the common normalizing factor across the density function, we may ‘count’ the contribution to each time point as follows:

$$\mathcal{C}\left(t\right) = \sqrt{2\pi}\left(\omega_{+} \cdot K_{+}\left(t\right) + \omega_{-} \cdot K_{-}\left(t\right)\right) \tag{3}$$

Using $K_{+}$ and $K_{-}$ to estimates of local probability and $\mathcal{C}\left(t\right)$ to estimate the sample size, the credible interval is specified using the beta distribution:

$$\begin{aligned}
\left[CI_t^{-}, CI_t^{+}|\alpha\right] &= Beta_{inv}\left(\left[\frac{\alpha}{2}, 1-\frac{\alpha}{2}\right]|A+\frac{1}{2}, B+\frac{1}{2}\right) \\
&\text{where} \\
A &= KPE(t)\cdot\mathcal{C}\left(t\right) \\
B &= \left(1-KPE(t)\right)\cdot\mathcal{C}\left(t\right) \\
I\left(p|A,B\right) &= \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}\int_0^p x^{A-1}(1-x)^{B-1}dx \\
Beta_{inv}\left(y|A,B\right) &= \left\{y : I\left(p|A,B\right) = y\right\}
\end{aligned} \tag{4}$$

This credible interval should be interpreted with some caution, as will be evident in the examples below. Its bounds depend on the assumption that the underlying probability does not change faster than can be captured by the optimized bandwidth. Optimized variable bandwidths (as described by Shimazaki and Shinomoto, 2010) provide a satisfactory solution to this problem, but yield a more complex model.

| Scenario | Span | Function |
|---|---|---|
| Big Step | $[0:149]$ | $p = 0.1$ |
| | $[150:300]$ | $p = 0.9$ |
| Step And Ramp | $[0:39, 91:179, 281:300]$ | $p = 0.2$ |
| | $[40:90]$ | $p = 0.7$ |
| | $[180:280]$ | $p = 0.8 - \frac{3t-540}{500}$ |
| Sine Wave | $[0:300]$ | $p = 0.5 + \frac{1}{3}\sin\left(\frac{t}{6\pi}\right)$ |
| Logarithmic Learning | $[0:300]$ | $p = \left(1+\exp\left(-0.015t\right)\right)^{-1}$ |

**Table 1.** Simulation scenarios. Each function specifies a canonical probability $p$ over a span of trials $t$, given in brackets. These functions are depicted by the dashed lines in Figure 2.

In order to demonstrate the efficacy and shortcomings of the kernel probability estimate, four scenarios were used to generate simulated data. These scenarios are described in Table 1, and the estimates based on the simulated data are depicted in Figure 2. In each case, estimates were based on 30 simulations, over an interval of 301 trials. For the simulations in the bottom row of Figure 2, every trial was simulated for every session. However, for the top and middle rows, only 32 and 96 trials were used, selected without replacement for each simulation from the 301 possible. This demonstrates the efficacy of the estimation procedure when sampling is sparse over the observed interval.

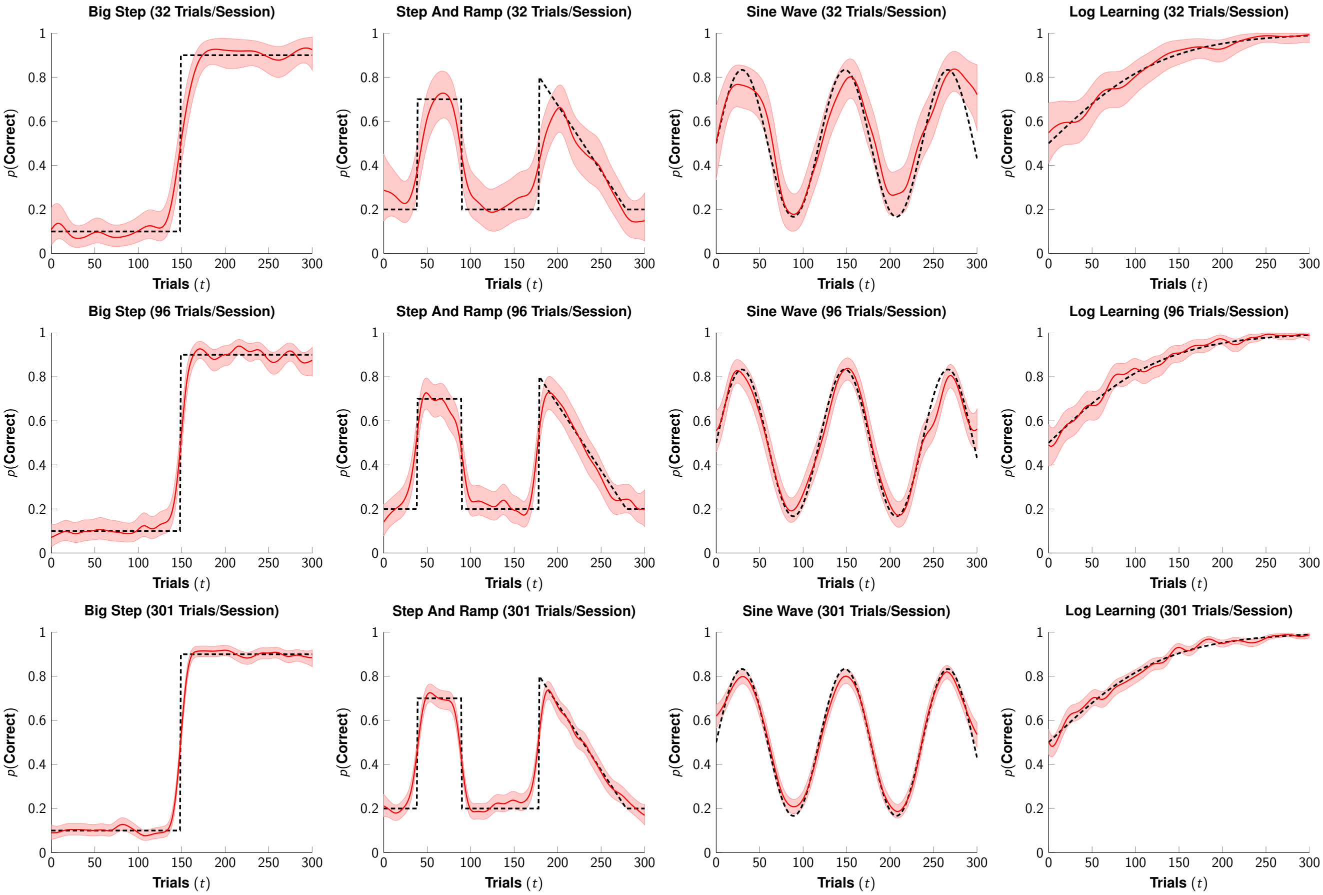

**Figure 2.** Simulation demonstration for the four scenarios (one per column) described in Table 1. Each $KPE$ is based on 30 independently simulated sessions. Canonical functions are plotted as dashed lines, while kernel fits are plotted as solid red lines. Each 95% credible interval is depicted by a shaded overlay. The fits in the top row are based on subsets of only 32 trials per session, sampled randomly from a possible 301. The fits in the middle row are based on subsets of 96 trials per session. The fits in the bottom row are based on the full 301 trials for each session.

Over most ranges of data, the estimated probability closely approximates the canonical function underlying the simulation. However, certain features are not captured. Sharp discontinuities (such as the edges of a step function) cannot be rendered precisely, as a result of the static bandwidth used to compute $K_+$ and $K_-$. In principle, however, the method for calculating $KPE$s could be implemented for other kernel-based smoothers that are more sensitive to highly localized features.

An important consideration for the credible intervals is the exchangeability of the sessions. It would be appropriate, for example, to implement the above credible interval for multiple sessions performed by the same subject under similar conditions. It would not, however, be an appropriate method for combining sessions performed by many subjects. Each subject's estimate at time $t$ might display dramatically different uncertainty than that of another subject.

In order to obtain a credible interval across subjects (taking differing uncertainty into account), it is necessary to convolve the beta distributions for each individual's estimates. Let the density function of the beta distribution associated with subject $s$ (out of a total of $S$ subjects) be defined as follows:

$$g_s\left(x\right) = Beta\left(x|A_s, B_s\right) = \frac{\Gamma(A_s+B_s)}{\Gamma(A_s)\Gamma(B_s)}x^{A_s-1}(1-x)^{B_s-1} \text{ where } \begin{array}{l} A_s = KPE(t)\mathcal{C}\left(t\right) \\ B_s = (1-KPE(t))\mathcal{C}\left(t\right) \end{array} \tag{5}$$

The sampling distribution for a sum of two of independent random variables is the convolution of their individual uncertainties (here denoted by the $*$ operator):

$$(g_1 * g_2)(x) = \int g_1(x - \tau)\, g_2(\tau)\, d\tau \tag{6}$$

This process is associative, so the sampling distribution for the sum across all subjects can be accomplished by further convolution, $(((g_1 * g_2) * g_3) * ... * g_S)(x)$. Note, however, that this distribution has no closed-form general solution, and can be computationally expensive to approximate numerically. If credible intervals are desired for the entire time series rather than for a single critical time-point, then determining the sampling distribution by simulating means of random samples drawn from the individual subject beta distributions is likely to be a more efficient approach.

Note, also, that the approach implied by Equation 6 does not treat subjects as random effects, and makes no effort to represent the population distribution of possible subjects. A credible interval based on subject-level convolution yields the uncertainty for the sample mean only, and as such should be treated as descriptive. Population-level inferences based on $KPEs$ are most tractably undertaken using bootstrapping, sampling randomly from the subject pool, and then randomly from the sampled subjects' respective uncertainties.

## RESPONSE ACCURACY IN A TRANSITIVE INFERENCE TASK

One of the most common binomial measures in behavior analysis is response accuracy. When task data consist of strings of correct/incorrect responses, estimating the proportion correct at any given moment is a problem of general interest.

Within this domain, a specific variety of estimate is consistently problematic: Estimated response accuracy on the first trial. For example, Jensen et al. (2015) presented rhesus monkeys (*Macaca mulatta*) with pairs of stimuli from the ordered list $ABCDEFG$. During an initial training phase, subjects were only presented with adjacent pairs (e.g. $AB$, $BC$, etc.). After approximately 200 trials, subjects were then presented with all possible pairs (e.g. $BD$, $CF$, etc.). If subjects were able to infer the ordering from the adjacent-pair training, then they are said to have performed a *transitive inference* (i.e. if $B > C$ and $C > D$, then $B > D$). Additionally, it is often observed that stimuli whose list positions are more widely separated yield higher accuracy, a so-called *symbolic distance effect*.

Testing for transitive inference is trickier than it appears, however, because any post-transition feedback potentially confounds the interpretation that it was the prior training (and not the new feedback) that explains the effect. It is thus desirable to have a procedure that permits a principled estimate of performance at the very first trial after training.

Figure 3 (top) shows the performance of one monkey on adjacent pairs (in red) and non-adjacent pairs (green, blue, and violet, representing distances of two, three, and four respectively), based on 51 sessions. Trial number is centered at transfer. 'Terminal pairs' (i.e. those that include the terminal items $A$ and $G$) are not included in this estimate.

If a subject performs above chance on non-adjacent, non-terminal items following adjacent pair training, they may be said to have performed a transitive inference. This can be seen clearly in Figure 3 (bottom), which shows the cross section of the sampling distributions for stimuli of all four types. Although adjacent pairs are only marginally significant ($p < .05$), the non-adjacent pairs are unambiguously above chance. Furthermore, the symbolic distance effect is clearly manifest: Distance 4 pairs yield the highest accuracy, followed by distance 3 pairs, and then by distance 2 pairs.

The clear advantage of using the $KPE$ to estimate performance, relative to existing methods, is that it permits inferences to be made efficiently (given the data available) and nonparametrically. An incidental

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1156v1 | CC-BY 4.0 Open Access | rec: 2 Jun 2015, publ: 2 Jun 2015

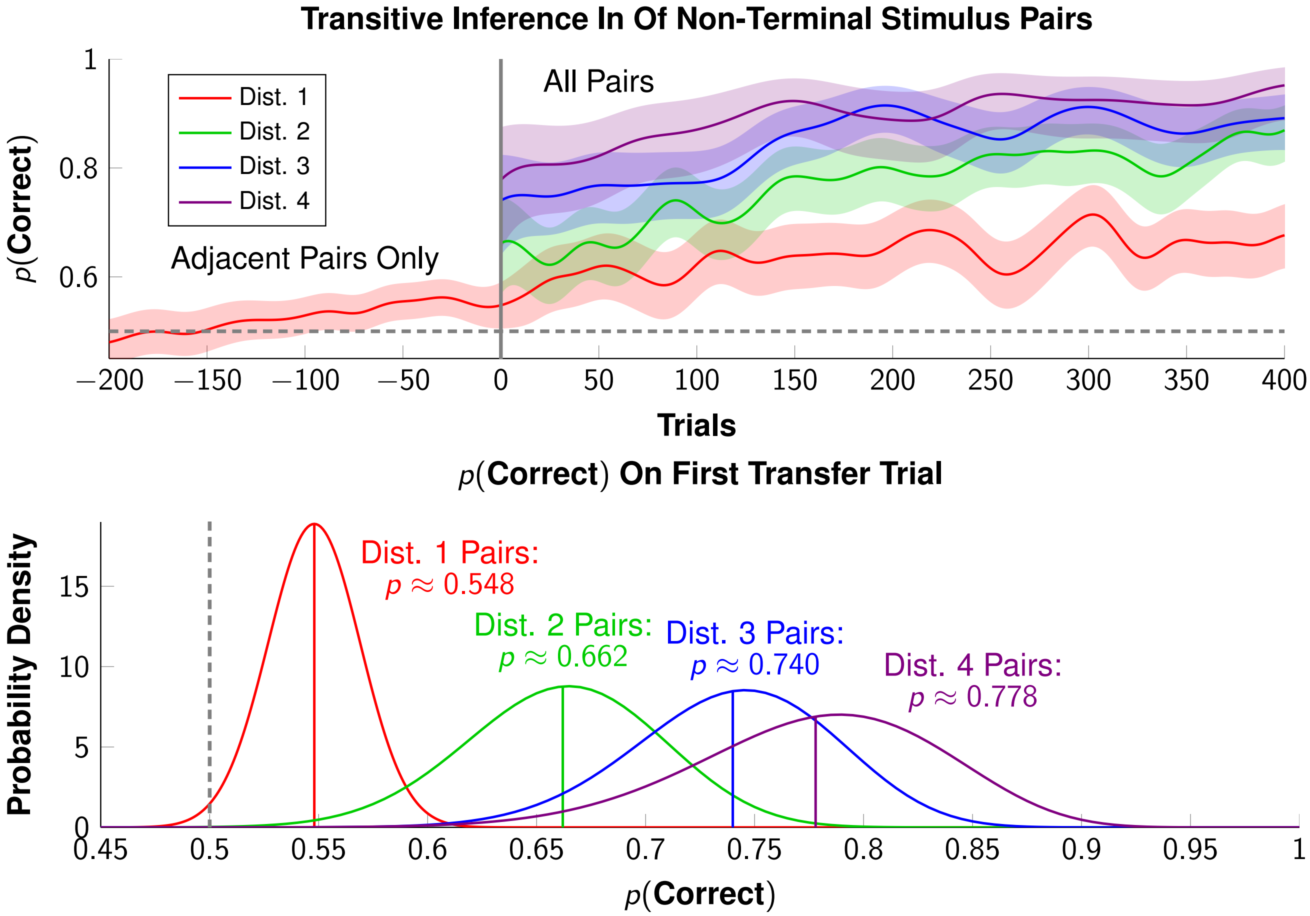

**Figure 3.** Response accuracy by a rhesus monkey performing a transitive inference task over 51 sessions(Jensen et al., 2015). **(Top)** Kernel probability estimate of response accuracy to non-terminal stimulus pairs, grouped by symbolic distance (red = distance one; green = distance two; blue = distance three; violet = distance four). Each estimate's 95% credible interval is indicated by the corresponding shaded region. Trial number is centered on the transition from adjacent-pair training to all-pair testing, also indicated by the solid gray line. Chance performance is indicated by the dashed gray line. **(Bottom)** Sampling distribution for each response accuracy at trial 0, showing estimated performance on the first trial following adjacent-pair-only training.

benefit to these features is the precise description of single subjects. The example presented in Figure 3 showcases fine-grained performance in a single subject, rather than relying on group data to provide an estimate. This in turn facilitates the consideration of individual differences, which are essential in distinguishing effective models of individual behavior from 'learning curves' that only resemble groups of subjects (Gallistel et al., 2004).

## MULTINOMIAL EXTENSION AND CHOICE BEHAVIOR

In time series with three or more outcomes, kernel probability estimates may be calculated in the same manner as described above by replacing the outcomes [Success, Failure] with a set of categories [A, B, C, ... ]. Rate estimates for each category $K_{\mathrm{A}}$, $K_{\mathrm{B}}$, etc. are estimated using the procedure described by Equation 1. The rate estimates are then set relative to the sum across categories to yield the $KPE$, and a perimeter bounding the credible region for the estimates may correspondingly be computed, based on the Dirichlet distribution (Chafaï and Concordet, 2009). Because the marginal credible intervals for each

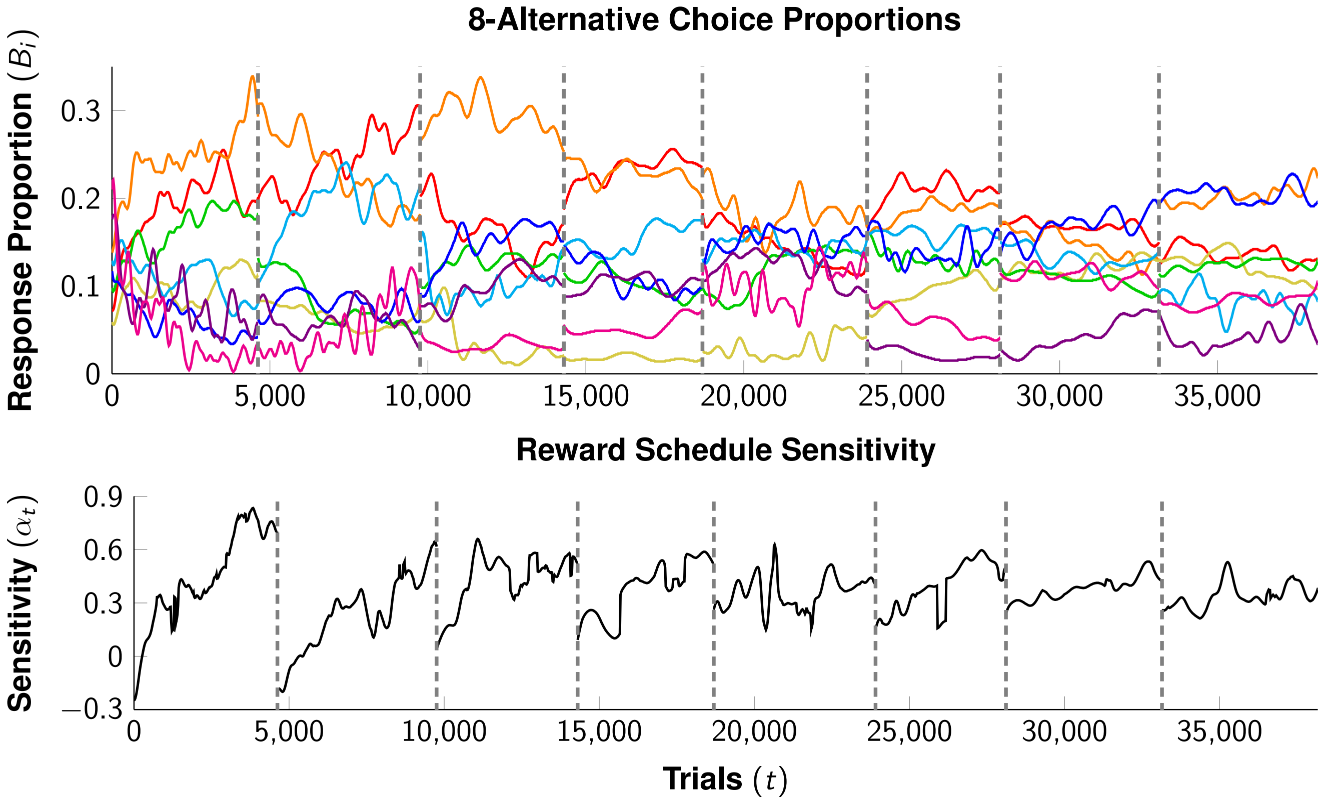

**Figure 4.** Transition in preference during a 8-item choice procedure performed by a rat (Jensen, 2014a). **(Top)** Each color corresponds to one of eight choice alternatives (see Table 2), and the black dashed line marks the boundary between phases. Observed shifts in behavior are in response to changes in the reward schedule from one session to the next.

proportion in a Dirichlet distribution are governed by the beta distribution, Equation 4 may be used when plotting intervals for each alternative separately.

Figure 4 (top) presents an example of such a multinomial fit, based on data previously reported by Jensen (2014a). In this example, one rat made 38164 responses over the course of eight experimental phases. During each phase, eight response levers were simultaneously available. On every trial, all eight of these levers secretly had a probability of 'setting up' a food reward to be collected, and the subject was required to forage among these eight options to find this hidden food. Transitions between phases are denoted in Figure 4 by dashed lines. Furthermore, each lever is plotted in a different color, with a corresponding column in Table 2.

In 'concurrent choice' procedures such as these, a central topic of interest is the relationship between the scheduled probability of reward associated with a lever $(R_i)$ and the corresponding proportion of responses devoted to that lever $(B_i)$. One of the more robust models of this relationship is the *generalized matching law* (Baum, 1974), whose multinomial form as defined by Jensen (2014b) is as follows:

$$\frac{B_i}{\prod_{j \in S} B_j} = \frac{\kappa_i}{\prod_{j \in S} \kappa_j} \left( \frac{R_i}{\prod_{j \in S} R_j} \right)^{\alpha} \tag{7}$$

Here, the parameter $\kappa_i$ denotes the 'bias' toward a given alternative $i$, while $\alpha$ denotes the 'sensitivity' toward the reward probabilities overall. For schedules of this type, rewards are collected most efficiently when every $\kappa_i = 1.0$ and $\alpha = 1.0$, a state of affairs called 'matching.' It is routinely observed, however, that many species 'undermatch' (i.e. $\alpha < 1.0$).

| Phase | Lever 1 | Lever 2 | Lever 3 | Lever 4 | Lever 5 | Lever 6 | Lever 7 | Lever 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0422 | 0.0357 | 0.0617 | 0.0552 | 0.0097 | 0.0065 | 0.0227 | 0.0162 |
| 2 | 0.0552 | 0.0162 | 0.0357 | 0.0065 | 0.0617 | 0.0227 | 0.0097 | 0.0442 |
| 3 | 0.0162 | 0.0617 | 0.0097 | 0.0357 | 0.0227 | 0.0552 | 0.0422 | 0.0065 |
| 4 | 0.0617 | 0.0422 | 0.0065 | 0.0162 | 0.0552 | 0.0097 | 0.0357 | 0.0227 |
| 5 | 0.0065 | 0.0097 | 0.0162 | 0.0227 | 0.0357 | 0.0422 | 0.0552 | 0.0617 |
| 6 | 0.0357 | 0.0227 | 0.0552 | 0.0617 | 0.0422 | 0.0162 | 0.0065 | 0.0097 |
| 7 | 0.0227 | 0.0065 | 0.0422 | 0.0097 | 0.0162 | 0.0357 | 0.0617 | 0.0552 |
| 8 | 0.0097 | 0.0552 | 0.0224 | 0.0422 | 0.0065 | 0.0617 | 0.0162 | 0.0357 |
| $\kappa_i$ | 1.7196 | 2.0007 | 0.5504 | 1.0990 | 1.2325 | 1.1908 | 0.5850 | 0.5597 |

**Table 2.** Programmed probabilities of reward per trial per phase Figure 4. Also included are the $\kappa_i$ parameters estimated from the data in Figure 4.

Without a means of obtaining instantaneous estimates of response proportions, a traditional matching analysis takes very large sets of data and distills these into a small number of proportions. For example, given eight response alternatives and eight phases, the 38164 responses underlying Figure 4 would be compressed into only 64 proportions, yielding a single set of parameters by regression analysis. This process discards all temporal information, and assumes that all parameters are static over time.

A more granular method for estimating sensitivity is discussed by Jensen (2014a), in which data are considered as a time series and sensitivity is shown to change over time. If $\kappa_i$ is held constant and an estimate of response proportions can be obtained for time $t$, then $\alpha_t$ can be identified using an approach called compositional analysis (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011). However, this requires having instantaneous estimates of $B_i$, and the procedure used by Jensen to obtain those estimates was not straightforward.

Figure 4 (top) provides the estimates of $B_i$ using $KPE$s, and these can then be used to estimate $\alpha_t$. Figure 4 (bottom) shows how $\alpha_t$ changes over time for this subject. During the early stages, the distributions of responses become relatively extreme, resulting in sensitivities at or above 0.6. As additional sessions accumulate, the amount of adjustment made by the subject lessens, until the last few sessions show only slight shifts in responding over time. The subject began the experiment at approximately 10 weeks of age, and the experiment took approximately 16 weeks to perform (about $\frac{1}{6}$ the natural lifetime of a laboratory rat). Consequently, aging is a likely culprit for the gradual reduction in response sensitivity. Effects such as these can only be detected if the analyst has a reliable method for obtaining instantaneous estimates of proportions.

Multinomial choice is not merely of interest in laboratory contexts. Many behavioral dynamics unfold spontaneously outside the lab, and these too are often subject to averaging over long intervals. For example, Romero et al. (2011) performed an analysis on post-conflict affiliative behavior in chimpanzees, collected over an eight year period. To take on these data, a generalized linear mixed model (GLMM) was used, collapsing across time. Kernel probability estimates could provide an inroad for identifying temporal trends that the reported GLMM would not have been able to detect.

The categorical behaviors that kernel probability estimation stand to clarify are not limited to those of non-human animals. For example, Flum et al. (2001) report on the change over eleven years of acute appendicitis treatment in Washington State, USA. Despite having a database of tens of thousands of patients, their study resolves the frequency of diagnoses only to the year of the incident. These estimated rates could be resolved to continuous time using a $KPE$.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1156v1 | CC-BY 4.0 Open Access | rec: 2 Jun 2015, publ: 2 Jun 2015

# CONCLUSION

Because binomial and multinomial data are widespread in the biomedical sciences, their appropriate analysis is a high priority. For the most part, data of these types are either not collected as time series (as in many survey methodologies), or are analyzed in a manner that omits temporal information by averaging over large groups of events (as in many decision-making paradigms).

At present, temporal information is ignored because it cannot easily be incorporated into familiar analyses, and this state of affairs will only be remedied using a suite of tools. Sophisticated methods (such as GAMs) permit statistical inference, but their sophistication limits their current use among applied researchers. To quote Brown et al. (2001), "It is generally true in statistical practice that only those methods that are easy to describe, remember and compute are widely used" (p. 115). Kernel probability estimates are intended to accommodate very basic analytic needs: the description and communication of categorical data over time.

# APPENDIX: OPTIMIZED BANDWIDTHS AND IMPRECISE TEMPORAL MEASUREMENT

Shimazaki and Shinomoto (2010) describe an efficient method for identifying an optimum bandwidth for kernel rate estimation. They provide a proof that the single optimal bandwidth $\omega^*$ for the Gaussian kernel minimizes of the following cost function:

$$\Phi\left(\omega|X,n,N\right) = \frac{1}{2\sqrt{\pi}n^2}\left(\frac{N}{\omega} + \frac{2}{\omega}\sum_{i<j}\left[\exp\left(\frac{-\delta_{i,j}}{4\omega^2}\right) - 2\sqrt{2}\exp\left(\frac{-\delta_{i,j}}{2\omega^2}\right)\right]\right) \tag{8}$$

where

$$\delta_{i,j} = \left(X_i - X_j\right)^2$$

Here, $X$ is the set of $N$ distinct times that events occurred over the course of $n$ independent sessions. Their proof relies on two assumptions: (1) that events in different sessions are independent from one another, such that the combined list of events approximates a Poisson process even if individual sessions display temporal dependencies, and (2) that time is measured continuously. Another way to state the second assumption is that $\delta_{i,j} > 0$.

This second assumption is reasonably approximated in electrophysiology (where events are often recorded with millisecond precision), but is much more problematic when considering trial data. The odds of two neurons spiking at the same moment in two different sessions is low, but the odds of two participants making correct responses on the same trial are high. This potentially yields a nontrivial number of instances of $\delta_{i,j} = 0$, which in turn drives $\omega^*$ below its otherwise optimal value.

Let $g$ represent the smallest measurable interval of time in a particular procedure. In trial-based experiments, $g = 1$ since half-trials are not meaningfully defined. In electrophysiology, if intervals are reported in seconds but recorded with millisecond precision, then $g = 0.001$. To correct Equation 8 for simultaneous trials, use the following value for $\delta_{i,j}$:

$$\delta_{i,j} = \left(X_i - X_j\right)^2 + 2\sigma^2, \text{ where } \sigma = \frac{g}{2} \tag{9}$$

In effect, this replaces each $X_i$ with $X_i + \epsilon$, where $\epsilon$ is an error term drawn from $\mathcal{N}\left(0, \frac{g}{2}\right)$. If most $X_i$ differ considerably from one another relative to $g$, then the effect of this error term is trivial. If, however, the data contain many tied events due to overly discrete units of time, this adjustment corrects for continuity.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1156v1 | CC-BY 4.0 Open Access | rec: 2 Jun 2015, publ: 2 Jun 2015

## ACKNOWLEDGMENTS

This work was supported by US National Institute of Health, grant 5R01MH081153-06 awarded to V. P. Ferrera and H. S. Terrace.

## REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22:231–242.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16:101–133.

Chafaï, D. and Concordet, D. (2009). Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association*, 104:1071–1079.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience*. MIT Press.

Flum, D. R., Morris, A., Koepsell, T., and Dellinger, E. P. (2001). Has misdiagnosis of appendicitis decreased over time? a population-based analysis. *Journal of the American Medical Association*, 286:1748–1753.

Gallistel, C. R., Fairhurst, S., and Balsam, P. D. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Science USA*, 101:13124–13131.

Hall, P. and Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, 90:149–173.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–318.

Jensen, G. (2014a). *Beyond Dichotomy: Dynamics of Choice in Compositional Space*. PhD thesis, Columbia University.

Jensen, G. (2014b). Compositions and their application to the analysis of choice. *Journal of the Experimental Analysis of Behavior*, 102:1–25.

Jensen, G., noz, F. M., Alkan, Y., Ferrera, V. P., and Terrace, H. S. (2015). Implicit value updating explains transitive inference performance: The betasort model. *PeerJ PrePrints*, 3:e1178.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, 135:370–384.

Parzen, E. (1962). Estimation of a probability density-function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis*. Wiley.

Romero, T., Castellanos, M. A., and de Waal, F. B. M. (2011). Post-conflict affiliation by chimpanzees with aggressors: Other-oriented versus selfish political strategy. *PLOS ONE*, 6:e22173.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837.

Shimazaki, H. and Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *Journal of Computational Neuroscience*, 29:171–182.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:495–518.