A Bayesian approach to visual inference

Susan VanderPlas, Dianne Cook, Christian Röttger, Heike Hofmann*
Department of Statistics and Statistical Laboratory, Iowa State University

August 12, 2020

Contents

1 Introduction

Graphics provide the opportunity to understand statistical data at an intuitive level: we can gain more information about the relationship between two variables by considering a simple scatter plot than we might obtain from an entire day of generating numerical summaries. Graphics leverage the bandwidth of the visual system for implicit data analysis, and because this analysis is implicit, we often assume graphics are not decisive in the same way that an hypothesis test is decisive: generally, graphics do not come with a significance threshold, and in many cases, we do not explicitly construct the hypothesis we might be testing before viewing the chart.

*The authors gratefully acknowledge funding from the National Science Foundation Grant # DMS 1007697. All data collection has been conducted with approval from the Institutional Review Board IRB 10-347

Visual inference allows for us to test graphics as visual statistics: charts are, after all, a quantity computed from values in a sample of data. In order to test whether a chart shows a visually significant result, we can use the same philosophy used by randomization tests: construct a sampling method consistent with the null hypothesis, generate many copies of the test statistic (in this case, the plot), and see where the real statistic falls in the distribution of artificially generated quantites (Buja et al., 2009). An assembly of several null plots with a target (or data) plot is called a *lineup*, after the criminal procedure of the same name. Typically, lineups are composed of 19 "null" plots (generated under the null hypothesis) and one data plot containing the real data.

Of course, with numerical statistics, there is a natural ordering to computed numerical quantities; with plots, we must run each statistic through another process (evaluation by the visual system, or a facsimile thereof[1]) in order to evaluate significance. During this evaluation process, the user selects one or more plots from the lineup which are "different" in some way (though some experiments may specify the particular feature under examination).

Typically, graphical tests utilize a service like Amazon Mechanical Turk to acquire multiple evaluations of the same lineup; informally, if multiple individuals select the plot generated from the data rather than the null plots, the visual statistic is likely to be significant (Majumder et al., 2013).

While visual inference was initially developed to mimic frequentist hypothesis tests, using lineups of 20 plots with one data plot, so that the probability of selecting the data plot randomly is $p = 0.05$, visual inference itself does not demand use of frequentist techniques. In this paper, we develop one possible framework for visual inference, using a Bayesian framework with a Dirichlet-multinomial distribution to model the probabilities of selecting each panel and a Multinomial distribution to model the observed participant selections. This framework for analysis of visual inference data has been in use as part of the `vinference` package (Hofmann and Röttger, ????) for several years (Loy and Hofmann, 2015; Loy et al., 2016; VanderPlas and Hofmann, 2017), but has not been formally described in any publication. Here, we provide the mathematical foundation for the multinomial-dirichlet model used in the analysis of visual inference experiments, exploring the implications of the model and proposing a modification which better describes the perceptual process of lineup evaluation.

1.1 Lineup Evaluations

Buja et al. (2009) introduced two types of lineups: Rorshach lineups, which contain only null plots, and standard one-target lineups, which contain m panels, $m-1$ of which are null plots, and one which shows the real data. VanderPlas and Hofmann (2017) introduced another type of lineup: the two-target lineup. This lineup contains target plots generated from two competing data generating models, with null plots that are generated from a mixture of the two target plot distributions. Two-target lineups can be used to test whether two competing effects have different visual salience. All of the lineup experiments the authors are aware of at this time are one or two-target lineup experiments.

In addition to different types of lineups, there are different types of lineup experiments.

[1] Cite Giora's Deep Learning work, `http://giorasimchoni.com/deep_visual_inference.html`

Scenario 1 K different lineups are shown to K independent individuals. In this scenario, both the data and the null plots in each generated lineup are distinct from those in every other lineup. This scenario is only practical when using purely simulated data (for both the data and null plots) or data large enough to allow for subsampling to generate K different data plots. Under this scenario, we can consider the number of target plot selections t out of K total evaluations, where each lineup evaluation is a bernoulli trial; the total number of data plot evaluations can then be modeled as a Binomial distribution with selection probability $1/m$ (for a single target lineup) (Majumder et al., 2013).

Scenario 2 K different sets of null plots are shown to K independent individuals; the same data plot is used in each lineup. Alternately, L sets of lineups are shown to $K > L$ individuals. In Scenario 2, there are dependencies introduced by reuse of the data or the lineups, providing an intermediate case between the two extremes of Scenario 1 and Scenario 3.

Scenario 3 The same lineup is shown to K independent individuals. This scenario is the most common scenario in the lineup experiments which have been completed to date (Hofmann et al., 2012; Roy Chowdhury et al., 2012; Majumder et al., 2013; Loy and Hofmann, 2015; VanderPlas and Hofmann, 2017). In this scenario, lineup evaluations by independent viewers are not independent because the viewers are evaluating the same combination of 19 null plots and one data plot. Any peculiar features which arise in a null plot may cause participants to select that null plot over the data plot; it is likely that where one individual makes this choice, others might as well. Thus, in Scenario 3, which is the most common lineup experiment scenario, it is not reasonable to assume that all panels are equally likely to be selected under the null hypothesis.

Scenario 3 is the primary motivation for the model which we will develop in the next section: while we cannot consider the panels in each lineup as equally likely to be selected, we can model the individual selection probabilities using a hierarchical model.

2 Lineup Model Specification

We will begin with a generic m-panel lineup, with selection probabilities $\theta_i, i = 1, ..., m$ where $\sum_{i=1}^{m} \theta_i = 1$, that is, the participant will select one (and only one) panel from the lineup as the most different. Our lineup has been evaluated by K individuals, with $c_i, i = 1, ..., m$ the selection count for each panel, and $K = \sum_{i=1}^{m} c_i$.

A natural data model for this data is the Multinomial distribution, which has parameters $K, \boldsymbol{\theta}$, where K describes the number of distinct outcomes and $\boldsymbol{\theta} = \theta_1, ..., \theta_m$ describes the probabilities of each outcome. We will fix K, as that is controlled by the experimental design, and model $\boldsymbol{\theta}$.

$$f(\boldsymbol{c}|K, \boldsymbol{\theta}) = \frac{K!}{c_1! \cdots c_m!} \prod_{i=1}^{m} \theta_i^{c_i} \tag{1}$$

We model panel selection probabilities θ_i using a Dirichlet distribution with concentration hyperparameter $\boldsymbol{\alpha}$, which happens to be conjugate to the multinomial distribution. As the position of the panels within the lineup are random, we use a symmetric Dirichlet distribution, with $\alpha_i = \alpha, i = 1, ..., m$, that is, the concentration hyperparameter is constant. This allows us to vary the lineup difficulty through the hyperparameter α without having to specify which plot i is the target plot.

The pdf of the symmetric Dirichlet distribution is

$$f(\boldsymbol{\theta}|\alpha) = \frac{(\Gamma(\alpha))^m}{\Gamma(m\alpha)} \prod_{i=1}^{m} \theta_i^{\alpha-1} \tag{2}$$

Using the conjugate relationship between the Dirichlet and Multinomial distributions, we then get the posterior distribution as the Dirichlet($\boldsymbol{c} + \boldsymbol{\alpha}$) distribution.

$$\begin{aligned}
(\alpha_1, ..., \alpha_m) = \boldsymbol{\alpha} &= \text{concentration hyperparameter} \\
(c_1, ..., c_m) = \boldsymbol{c} &= \text{observed plot selections}, \sum_{i=1}^{m} c_i = K \\
p(\boldsymbol{\theta}) &\sim Dirichlet(\boldsymbol{\alpha}) \\
f(\boldsymbol{c}|\boldsymbol{\theta}) &\sim Multinomial(\boldsymbol{\theta}, K) \\
p(\boldsymbol{\theta}|\boldsymbol{c}, \boldsymbol{\alpha}) &\sim Dirichlet(\boldsymbol{c} + \boldsymbol{\alpha})
\end{aligned} \tag{3}$$

where $Multinomial(\boldsymbol{\theta}, K)$ is defined as in Equation 1 and $Dirichlet(\boldsymbol{\alpha})$ is defined as in Equation 2.

Typically, when evaluating lineups, we compare the number of target plot identifications with the aggregate number of null plot identifications (see Majumder et al. (2013)). This is equivalent to the marginal distribution of c_t, where $t \in 1, ..., m$ is the index of the target panel in the lineup; that is, in Majumder et al. (2013), a binomial distribution, and in this formulation, a beta-binomial distribution.

$$\begin{aligned}
\alpha &= \text{concentration hyperparameter} \\
c_t &= \text{target plot selections}, \\
K &= \text{total evaluations} \\
p(\theta_t) &\sim Beta(\alpha, (m-1)\alpha) \\
f(c_t|\theta_t) &\sim Binomial(\theta_t, K) \\
p(\theta_t|\boldsymbol{c_t}, \alpha) &\sim Beta(c_t + \alpha, K - c_t + (m-1)\alpha)
\end{aligned} \tag{4}$$

While this model was originally developed under a Bayesian framework, it is philosophically agnostic: it would be equally reasonable to think of this as an overdispersed multinomial model.

Examining the parameters of either the full or marginal model specifications in Equation 3 and Equation 4, it is evident that α provides the equivalent of pseudo-observations for each plot; that is, the effect of α is equivalent to adding α identifications to each panel

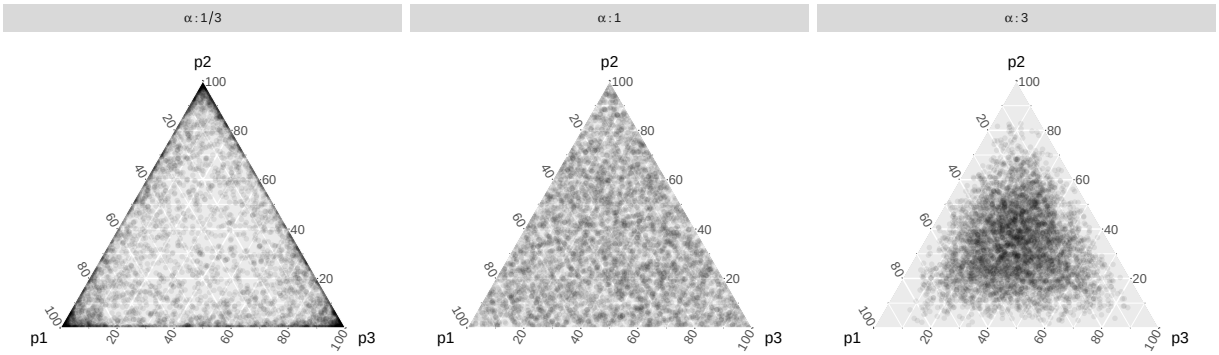

Figure 1: Dirichlet distributed samples on the 2-dimensional simplex.

in the lineup. When α is small, these pseudo-observations have relatively little influence, but when α is large, the pseudo-observations can quickly dwarf any information provided by the data. This is particularly true for the marginal Beta-Binomial model, where the equivalent of $(m-1)\alpha$ pseudo-observations are added. In most lineup studies, a plot might be evaluated between 10 and 30 times; with a $m = 20$ lineup, even $\alpha = 1$ can easily dominate the participant selection data.

In addition to the pseudo-observation interpretation, α provides information about the number of panels in a lineup which are likely to attract participant interest. It is useful to detour slightly from the discussion of visual inference to explore the impact and interpretation(s) of α in the context of statistical lineups.

2.1 Dirichlet Hyperparameter

When $\alpha = 1$, the symmetric Dirichlet distribution is uniform on the $m-1$ dimensional simplex. When $\alpha < 1$, the mass of the distribution is along the edges of the simplex, where most values of θ_i will be close to 0. When $\alpha > 1$, the mass of the distribution is in the center of the simplex, with most of the θ_i having similar values. Figure 1 shows ternary plots (Hamilton and Ferry, 2018) of values simulated from a 3-dimensional Dirichlet distribution which illustrate the effect of α on the sampled $\boldsymbol{\theta}$.

While graphical illustrations of the 20-dimensional dirichlet distribution are more difficult, we can use prior predictive simulations to assess the meaning of α as it relates to how many panels in a lineup attract participant attention. Figure 3 shows simulated c_i counts (sorted for visual clarity) for several values of α; it is evident that for $\alpha < 0.05$ only one panel of the lineup receives significant attention, while for $\alpha > .25$, participant attention is divided among several interesting panels of the lineup.

The data model used in Majumder et al. (2013) assumes $\theta_t = 1/m$, that is, θ_t is fixed and equal to the selection probability of every other panel in the lineup. This assumption, which would correspond to infinite α, does not match our experience when evaluating a lineup, nor the accumulated experimental evidence (assembled under Scenario 3, as discussed in subsection 1.1) which shows that even null plots do not show equal selection probabilities for each panel. When examining a lineup, we are generally drawn immediately to 1-4

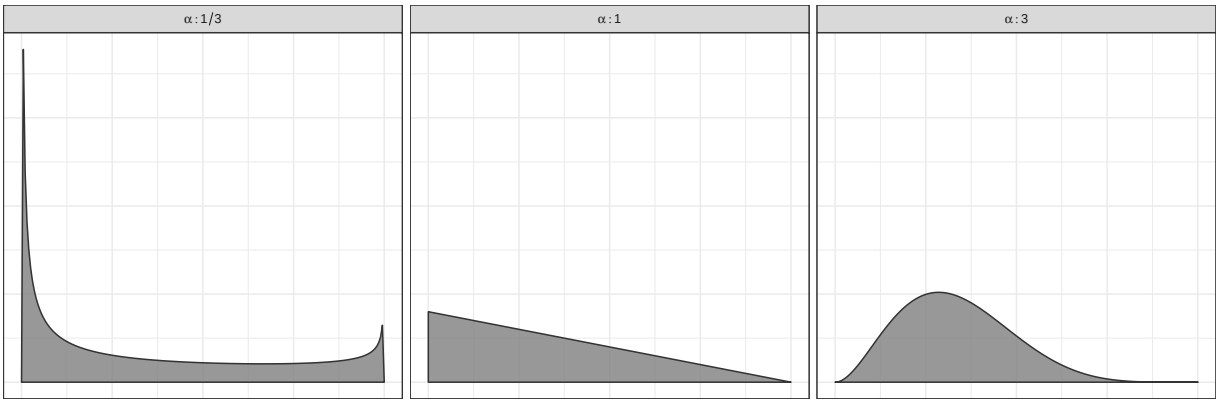

Figure 2: Marginal Beta$(\alpha, 2\alpha)$ densities corresponding to the above Dirichlet densities.

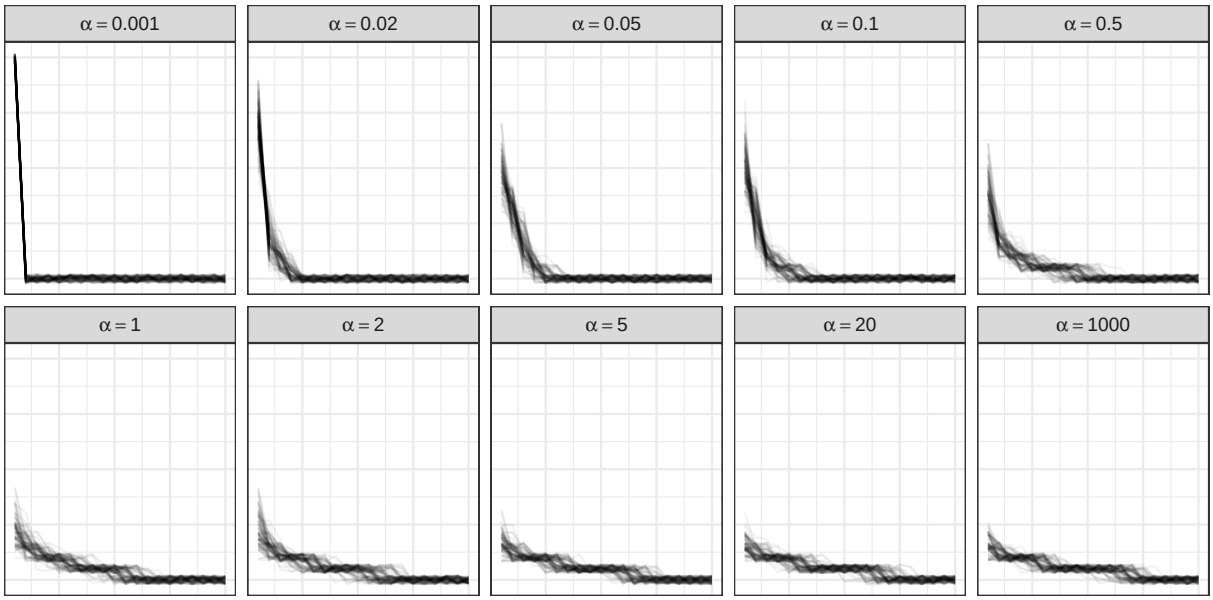

Figure 3: Prior predictive distribution of number of participant selections of panels c_i (sorted by frequency) for different values of α, with fixed $K = 20$ plot evaluations. Low values of α have fewer plots with any participant selections, while higher values of α have more plots with participant selections.

panels, and the remaining evaluation is to decide between those panels; we also know that typically the same panels are selected across multiple individuals. To account for this issue, recent analyses of lineups calculate visual p-values using the mass function

$$P(C \geq x) = \sum_{x=C}^{K} \binom{K}{x} \frac{1}{B(\alpha, (m-1)\alpha)} \cdot B(x+\alpha, K-x+(m-1)\alpha) \tag{5}$$

where C is the number of data panel detections and K is the number of independent evaluations of the lineup (Hofmann and Röttger, ????; VanderPlas and Hofmann, 2017; Loy and Hofmann, 2015). A derivation of this mass function is provided in the appendix. A similar method is found in the `vinference` package, which calculates visual p-values by simulating draws of θ from a uniform distribution (corresponding to the assumption that $\alpha = 1$, as shown in Figure 1) but the more general solution is useful to consider, as we may not actually believe θ is uniformly distributed over the $(m-1)$ simplex.

2.2 Hyperparameter Selection

We know from experience as well as cognitive principles that it is unreasonable to assume that the selection probability of every null plot is precisely equal: null plots are randomly generated, and occasionally, the randomly generated plot will have an interesting feature (that may or may not be present in the target plot). When that occurs, the interesting null plot will be selected more frequently than the other nulls, despite being generated by the same distribution. The ability to identify stimuli as being different from one another is a fundamental part of cognition; the abstractions that allow us to use the terms 'same' and 'different' are fundamental to human intelligence (Ming and Stewart, 2017). As a result, when presented with a lineup, we will typically gravitate towards one or two panels which are different from our mental representation of a generalized panel on some measure, though not always the measure that's under investigation. Note that even though the interesting null plot effect is found in all scenarios proposed in subsection 1.1, it is not a systematic issue in Scenario 1; in Scenario 3, the null plots are not re-generated with each lineup evaluation, so an interesting null plot selected by one individual may also effect the evaluation of by other individuals. That is, in Scenario 3, the interesting null plots affect all evaluations, and this effect can significantly affect the experiment results. However, in Scenario 3, we can also estimate the size of this effect: by examining repeated evaluations of a lineup, we can leverage that replication to estimate the distinctiveness of the set of null plots in the lineup. While this effect may exist in Scenario 1, we cannot estimate the effect because there are not repeated evaluations of the same set of lineups. Thus, in Scenario 1, the best estimate we can make for the generic θ is $1/m$, which is equivalent to the Multinomial model without the Dirichlet hyperparameter and with $\theta = 1/m$.

In the theoretical Multinomial-Dirichlet model proposed in this paper, the number of panels which could be expected to be visually different in a generalized lineup is a function of α, the hyperparameter in Equation 2 and Equation 3. In practice, we would create a null plot generating model first, and set α according to the perceived difficulty of lineup evaluation using the null generating model in question. A difficult lineup, with many

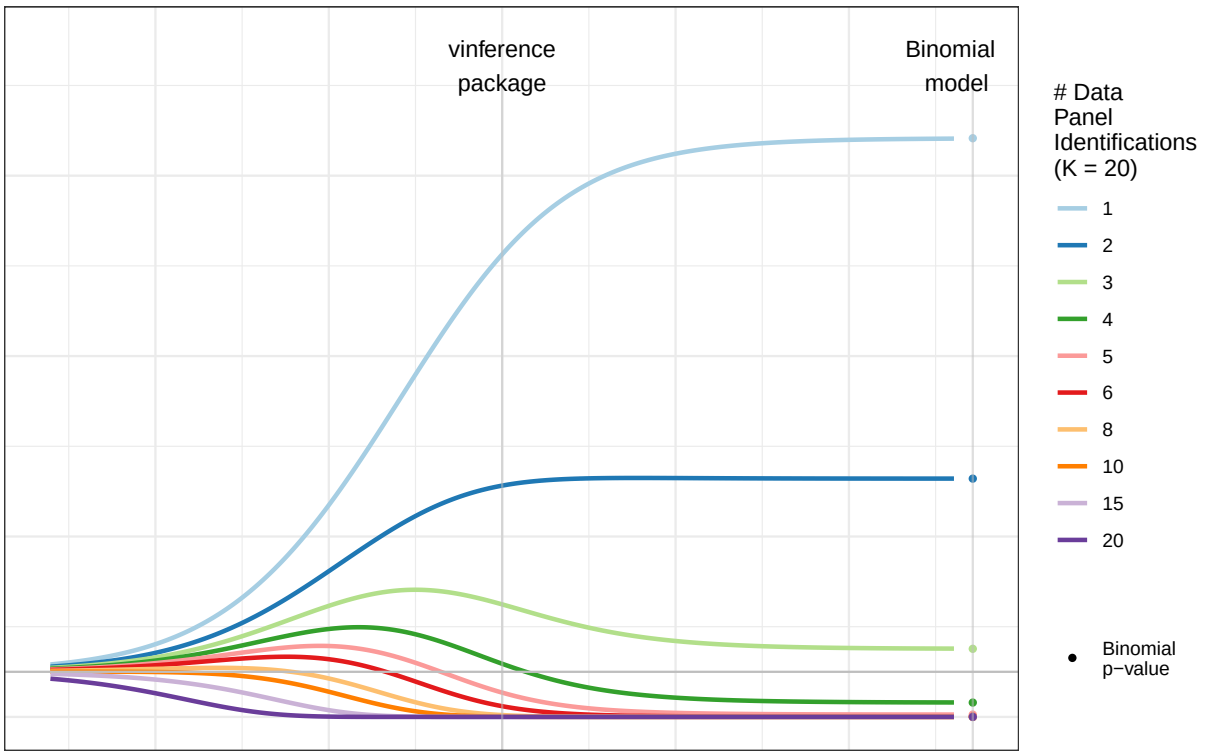

Figure 4: Sensitivity of visual p-value to selection of α under the beta-binomial model. Corresponding values for the binomial model are shown on the right side of the plot; as $\alpha \to \infty$, the beta-binomial p-values converge to the binomial model p-value.

potentially interesting panels, would have a higher α value than an easy lineup with no null panels which were visually salient relative to the data panel.

α plays a similar role in modulating the visual p-value calculated from Equation 5: lower α values produce a higher visual p-value estimate, and higher α values produce a lower visual p-value estimate.

Clearly, the choice of α is critical. From a practical perspective, the number of plots which are visually salient and thus likely to be selected by participants is a factor of the lineup design (zero, one, or two targets), null plot generation method, and possibly the form of the plot (aesthetics, geometric representations, scales). While the prior predictive distributions in Figure 3 are illustrative, in practice, we do not usually have a good instinct for what a reasonable value of α would be for a particular lineup generation method. In the next section, we present a method for estimating α from Rorshach or standard lineups and discuss possible uses for this method in assessing null plot generation and lineup difficulty.

3 Estimating α

3.1 Derivation of MLE for α

In order to estimate α from the null plots of a lineup, let m_0 be the number of null plots in an m- panel lineup. If the lineup is a Rorshach lineup, then $m = m_0$. We can estimate α using a set of such lineups, $j = 1, ..., n$, where the lineups in the set were generated under the same distribution. The likelihood function is then

$$\begin{aligned}\mathscr{L}(\alpha|\theta) &= \prod_{j=1}^{n}\left(\frac{1}{B(\alpha)}\right)^{m_0}\prod_{i=1}^{m_0}\theta_{ij}^{\alpha-1} \\ &= \left(\frac{\Gamma(\alpha m_0)}{(\Gamma(\alpha))^{m_0}}\right)^{n}\prod_{ij}\theta_{ij}^{\alpha-1} \qquad (6)\end{aligned}$$

and the derivative of the log-likelihood function can be calculated as

$$\begin{aligned}\ln\mathscr{L}(\alpha|\theta) &= n\ln\Gamma(\alpha m_0) - nm_0\ln\Gamma(\alpha) + \sum_{ij}(\alpha-1)\ln\theta_{ij} \\ \frac{d}{d\alpha}\ln\mathscr{L}(\alpha|\theta) &= nm_0\psi(\alpha m_0) - nm_0\psi(\alpha) + \sum_{ij}\ln\theta_{ij} \qquad (7)\end{aligned}$$

where $\psi(x)$ is the digamma function, $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$. Setting this to zero, we find that the MLE of α can be obtained empirically from the sum of the log probabilities θ_{ij}.

$$\begin{aligned}0 &= nm_0\psi(\alpha m_0) - nm_0\psi(\alpha) + \sum_{ij}\ln\theta_{ij} \\ nm_0\psi(\alpha) - nm_0\psi(\alpha m_0) &= \sum_{ij}\ln\theta_{ij} \\ \psi(\alpha) - \psi(\alpha m_0) &= \frac{1}{nm_0}\sum_{ij}\ln\theta_{ij} \qquad (8)\end{aligned}$$

The second derivative of the log likelihood function uses the trigamma function, $\psi_1(x) = \frac{d^2}{dx^2}\ln\Gamma(x)$. $\psi_1(x)$ can also be written as the series $\psi_1(x) = \sum_{z=0}^{\infty}\frac{1}{(z+x)^2}$. Thus,

$$\begin{aligned}\frac{d^2}{d\alpha^2}\ln\mathscr{L}(\alpha|\theta) &= nm_0^2\psi_1(\alpha m_0) - nm_0\psi_1(\alpha) \qquad (9) \\ &= nm_0^2\left(\sum_{x=0}^{\infty}\frac{1}{(\alpha m_0+x)^2} - \sum_{x=0}^{\infty}\frac{1}{m_0(\alpha+x)^2}\right) \\ &= nm_0^2\sum_{x=0}^{\infty}\left(\frac{1}{(\alpha m_0+x)^2} - \frac{1}{m_0(\alpha+x)^2}\right) \\ &= nm_0^2\sum_{x=0}^{\infty}\frac{m_0(\alpha+x)^2 - (m_0\alpha+x)^2}{m_0(\alpha m_0+1)^2(\alpha+x)^2} \leq 0 \text{ for } m_0 \geq 1, \alpha > 0 \\ &\quad \text{as } m_0(\alpha+x)^2 \leq (\alpha m_0+x)^2 \text{ for } m_0 \geq 1, \alpha > 0\end{aligned}$$

The empirical solution to Equation 8 is thus a local maximum, and the MLE of α is the numerical solution to the equation Equation 8.

3.2 Pooling null evaluations

The MLE for α in Equation 8 allows for the combination of evaluations of multiple sets of null plots (either from Rorshach lineups or lineups with one or more target panels), but in order for this combination of data to be meaningful, the following conditions should hold:

1. Null plots should be generated by the same model
2. Only plots with the same aesthetics should be pooled for α estimation
3. Selection method should be the same (e.g. single or multiple target plots)

When estimating $\hat{\alpha}$ from null plot selections in a standard lineup (e.g. a lineup containing at least one target plot), there is the possibility that no null plots are selected. A more reliable and systematic way to estimate $\hat{\alpha}$ would be to include a Rorshach lineup for each set of parameters used to generate null plots. These Rorshach lineups may be integrated into the testing procedure, or may be part of a pilot study used to assess the null plot generating model before it is used in lineups with data targets.

Intuitively, α is related to the proportion of null plots which would have some visually interesting feature: if α is low, that is, $\alpha << 1$, θs generated by the model would tend to be close to zero, with one or two larger panel selection probabilities (that is, one or two of the panels would be significantly more noticable). If α is high, $\alpha >> 1$, θs generated by the model would be closer to $1/m_0$, that is, each panel would be approximately equally likely to be selected. We know that the data generating model is likely to affect α, and from VanderPlas and Hofmann (2017) we know that the aesthetics can also significantly effect the selection of panels in a lineup.

In order to illustrate the variability in $\hat{\alpha}$ estimates across different lineup studies and to get a sense of the range of reasonable $\hat{\alpha}$ values, we estimated α for several previous single-target lineup studies of various sizes and designs. In several of these studies, multiple panel selections were allowed; these are allocated as partial selections to each set of counts.

Figure 5 shows the estimated αs for each set of data, parameters, and aesthetics used in the studies. In most studies, $\hat{\alpha} \approx 0.07$ (5% - 95% quantiles: 0.063 and 0.085), with very few estimated values over 0.10. There are some exceptions: Study 5 and Study 6 have $\hat{\alpha}$ values between 0.061 and 0.236, which is a much wider range. In these studies, it also appears that plots which allowed participants to select multiple panels tend to have higher $\hat{\alpha}$ values, that is, multiple plot selections might allow for more diffusion of probabilities over multiple null plots. This effect might be heightened by increased lineup difficulty, which would also tend to increase $\hat{\alpha}$ values. One lineup from Study 5 is shown in Figure 6; only one participant identified the data target from among the nulls, suggesting that this is a difficult lineup, but the diffusion of identifications across many null panels indicates that the null data model generates plots with a relatively homogeneous level of visual interest.

The estimated $\hat{\alpha}$ value can then be used either in the specification of the null model α (α_{M_2}) or in the calculation of the visual p-value under the frequentist framework. Currently,

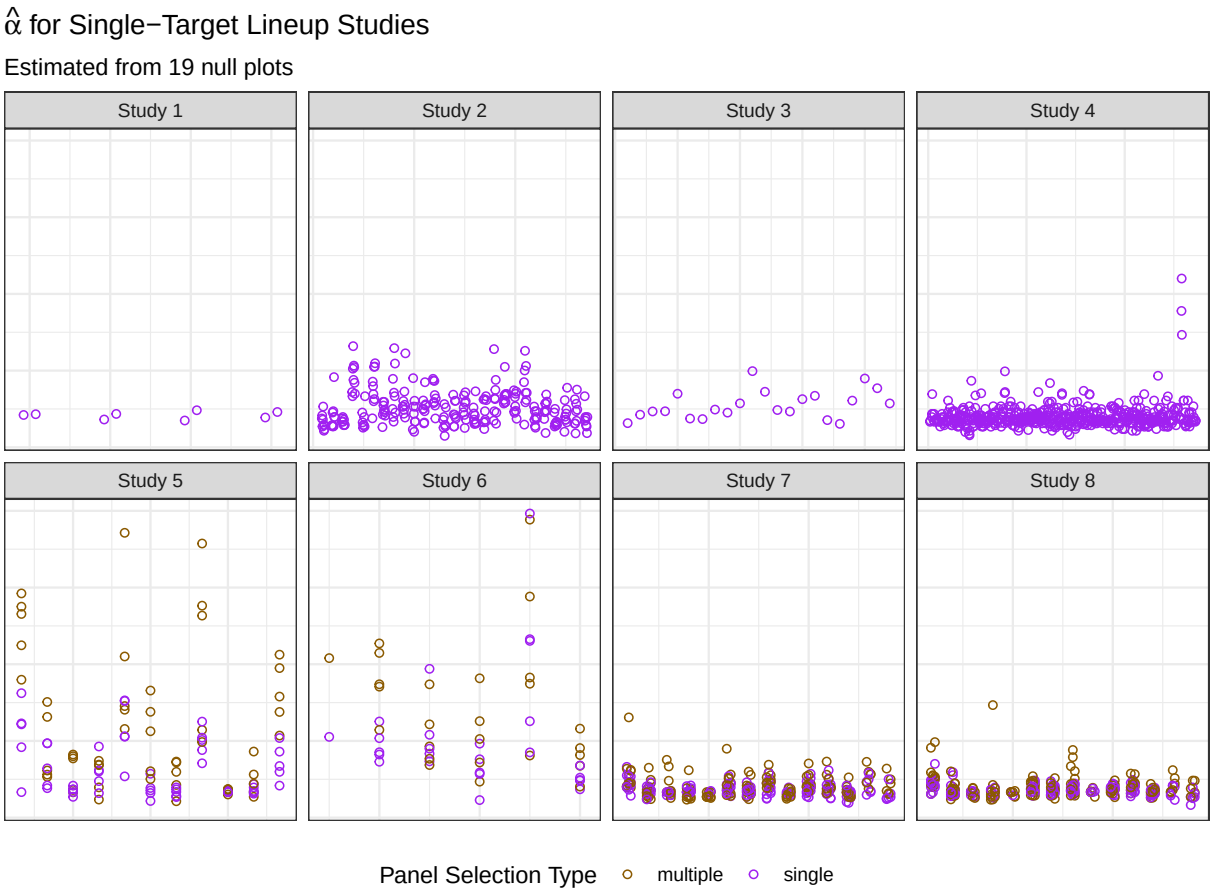

Figure 5: Alpha estimates for several lineup studies.

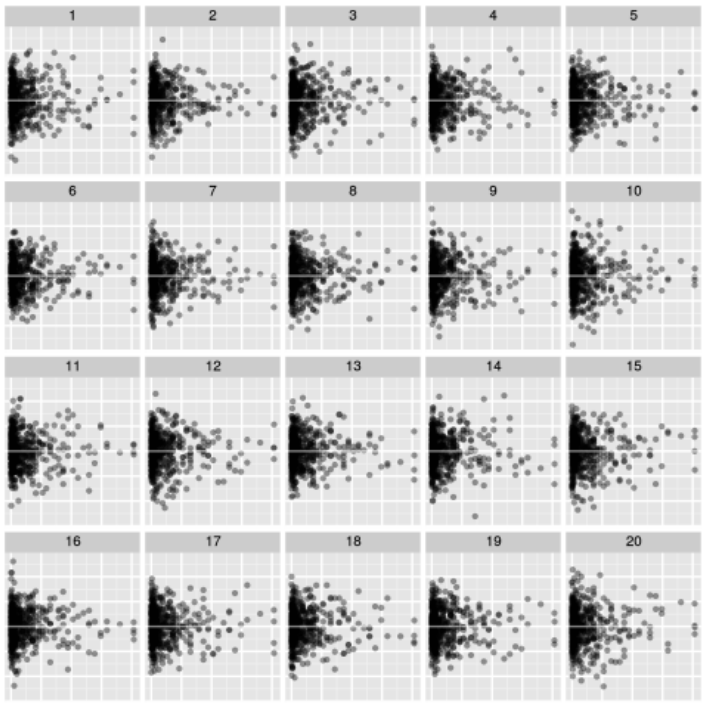

Figure 6: A lineup from Study 5. The target plot is in panel 7+7. Only one participant identified the target plot in 24 lineup evaluations. A total of 15 panels in this lineup were selected, indicating that the $\hat{\alpha}$ associated with this plot should be relatively large; in fact, $\hat{\alpha}$ =0.1624.

in either scenario, the data used in the calculation are the target plot identifications and the sum of all null plot identifications; the individual values of the null plot identifications are not used. These data can be reclaimed to estimate a general α for the overall null model generation scheme, or, in an ideal situation, Rorshach lineups could be used to estimate α directly without any possible contamination effects induced by the presence of target plots.

4 Impact of $\hat{\alpha}$ Estimation

We have already established that the choice of α has a large impact on the visual p-value (Figure 4), but if we use the estimation method described in the last section, how do the results change in practice? Here, we show 3 different lineups - nonsignificant, marginal, and highly significant, and examine the visual p-values computed under the binomial distribution, by simulation using the `vinference` package (with default parameter $\alpha = 1$), and according to the beta-binomial model in Equation 5 with $\alpha = 1$, and finally with $\alpha = \hat{\alpha}$ estimated from null plot selections in the experiment which originally included the lineup in question.

Figure 7 shows a plot which was evaluated a total of 111 times, with 17 target panel selections. The target panel was selected more frequently than any other panel, but a total of 10 panels were selected, and each selected panel was chosen by 10 or more participants. In this case, we would not expect the results to reveal a statistically significant effect, because many panels were selected almost as frequently as the target panel; in fact, the binomial p-value for this model is 0.0000, the simulated vinference p-value is 0.0076, the dirichlet-multinomial p-value with $\alpha = 1$ is 0.0559, and the dirichlet-multinomial p-value with $\hat{\alpha} = 0.0877$ is 0.1061. This lineup represents a case that is not handled well by the binomial model or the multinomial-dirichlet model with noninformative $\alpha = 1$ selection, but is handled reasonably by the multinomial-dirichlet model with an informative choice for α.

Return to the 3 lineups shown in Bayes Factor section, provide estimated alphas and visual p-values for each plot.

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Royal Society Philosophical Transactions A*, 367, 4361–4383.

Hamilton, N. E. and Ferry, M. (2018), "ggtern: Ternary Diagrams Using ggplot2," *Journal of Statistical Software, Code Snippets*, 87, 1–17.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical tests for power comparison of competing designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441–2448.

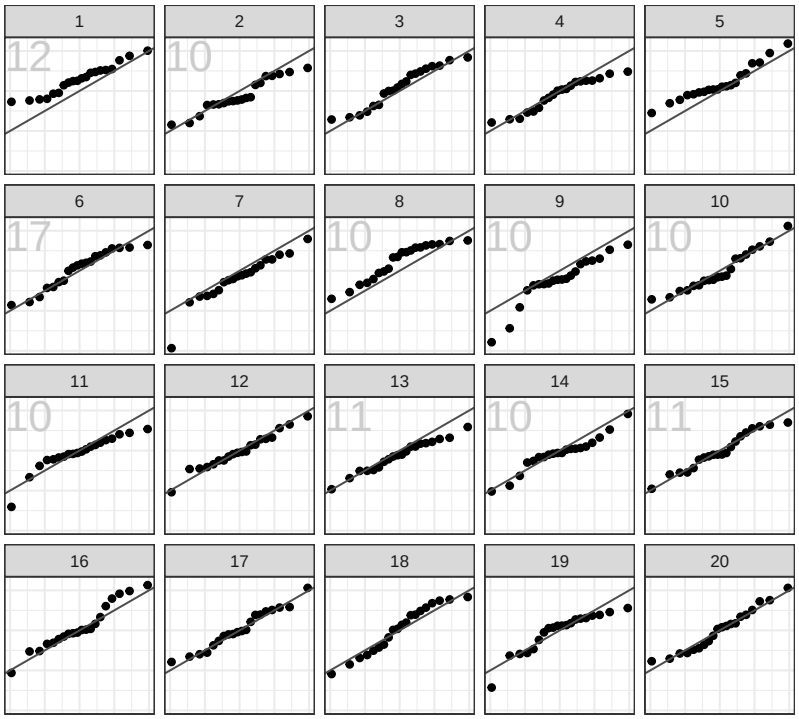

Figure 7: A lineup from Loy and Hofmann (2015) which was evaluated a total of 111 times. For panels with non-zero selections c_i the numbers are shown at the top left of each panel. The binomial p-value for this model is 0.0000, the simulated vinference p-value is 0.0076, the dirichlet-multinomial p-value with $\alpha = 1$ is 0.0559, and the dirichlet-multinomial p-value with $\hat{\alpha} = 0.0877$ is 0.1061

Hofmann, H. and Röttger, C. (????), *vinference: Inference under the lineup protocol*, r package version 0.1.1.

Loy, A., Follett, L., and Hofmann, H. (2016), "Variations of Q–Q Plots: The Power of Our Eyes!" *The American Statistician*, 70, 202–214.

Loy, A. and Hofmann, H. (2015), "Are You Normal? The Problem of Confounded Residual Structures in Hierarchical Linear Models," *Journal of Computational and Graphical Statistics*, 24, 1191–1209.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of visual statistical inference, applied to linear models," *Journal of the American Statistical Association*, 108, 942–956.

Ming, S. and Stewart, I. (2017), "When things are not the same: A review of research into relations of difference," *Journal of Applied Behavior Analysis*, 50, 429–455.

Roy Chowdhury, N., Cook, D., Hofmann, H., and Majumder, M. (2012), "Where's Waldo: Looking Closely at a Lineup," Tech. Rep. 2, Iowa State University, Department of Statistics.

VanderPlas, S. and Hofmann, H. (2017), "Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics," *Journal of Computational and Graphical Statistics*, 26, 231–242.

A Derivation of Visual p-value Distribution

In the marginal case, where we have K evaluations resulting in C target plot evaluations of a m plot lineup, we start with the following:

$$\begin{aligned} f(\theta|\alpha) &= \text{Beta}(\alpha, (m-1)\alpha) \\ &= \frac{\Gamma(m\alpha)}{\Gamma(\alpha)\Gamma((m-1)\alpha)} \cdot \theta^{\alpha-1}(1-\theta)^{(m-1)\alpha-1} \\ &= B\left(\alpha, (m-1)\alpha\right) \cdot \theta^{\alpha-1}(1-\theta)^{(m-1)\alpha-1} \end{aligned}$$

$$\begin{aligned} P(C|K,\theta) &= \text{Binomial}(C, K, \theta) \\ &= \binom{K}{C}\theta^{C}(1-\theta)^{K-C} \end{aligned}$$

By Bayes Theorem, if $A_1 = C + \alpha$ and $A_2 = K - C + (m-1)\alpha$,

$$\begin{aligned} f(\theta|C, K, \alpha) &= \frac{f(\theta|\alpha)P(C|\theta)}{P(C = c)} \\ &= \frac{B\left(\alpha, (m-1)\alpha\right)}{P(C = c)}\theta^{\alpha-1}(1-\theta)^{(m-1)\alpha-1} \cdot \binom{K}{C}\theta^{C}(1-\theta)^{K-C} \\ &= \frac{1}{P(C = c)}B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}\theta^{A_1-1}(1-\theta)^{A_2-1} \end{aligned}$$

As $f(\theta|C, K, \alpha)$ is a probability distribution, it integrates to 1. So we can infer that

$$\begin{aligned} P(C = c) &= \int B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}\theta^{A_1-1}(1-\theta)^{A_2-1}d\theta \\ &= B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}\int \theta^{A_1-1}(1-\theta)^{X_2-1}d\theta \\ &= B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}\frac{B(A_1, A_2)}{B(A_1, A_2)}\int \theta^{A_1-1}(1-\theta)^{A_2-1}d\theta \\ &= \frac{B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}}{B(A_1, A_2)\alpha)}\int B(A_1, A_2)\theta^{A_1-1}(1-\theta)^{A_2-1}d\theta \\ &= \frac{B\left(\alpha, (m-1)\alpha\right)\binom{K}{C}}{B\left(C + \alpha, K - C + (m-1)\alpha\right)} \end{aligned}$$

Thus, the visual p-value for a lineup with C target selections out of K evaluations is

$$P(x \geq C) = \sum_{x=C}^{K} \frac{\binom{K}{x}B(\alpha, (m-1)\alpha)}{B(x + \alpha, K - x + (m-1)\alpha)} \tag{10}$$

A similar derivation holds in the full Dirichlet-Multinomial model.