Robust visual estimation as source separation

Mordechai Z. Juni
Department of Psychology, New York University, New York, NY, USA

Manish Singh
Department of Psychology and Center for Cognitive Science, Rutgers University, New Brunswick, NJ, USA

Laurence T. Maloney
Department of Psychology, New York University, New York, NY, USA, & Center for Neural Science, New York University, New York, NY, USA

We developed a method analogous to classification images that allowed us to measure the influence that each dot in a dot cluster had on observers’ estimates of the center of the cluster. In Experiment 1, we investigated whether observers employ a robust estimator when estimating the centers of dot clusters that were drawn from a single distribution. Most observers’ fitted influences did not differ significantly from that predicted by a center-of-gravity (COG) estimator. Such an estimator is not robust. In Experiments 2 and 3, we considered an alternative approach to the problem of robust estimation, based on source separation, that makes use of the visual system’s ability to segment visual data. The observers’ task was to estimate the center of one distribution when viewing complex dot clusters that were drawn from a mixture of two distributions. We compared human performance to that of an ideal observer that separated the cluster into two sources through a maximum likelihood algorithm and based its estimates of location using the dots it assigned to just one of the two sources. The results suggest that robust methods employed by the visual system are closely tied to mechanisms of perceptual segmentation.

Keywords: center of gravity, source separation, perceptual segmentation, scission, robust statistics


Introduction

Visual tasks are often modeled as statistical estimation tasks. In Figure 1A, for example, the viewer is confronted with the scatter plot of a sample of dots \( p_i = (p_i^x, p_i^y) \), \( i = 1, \ldots, n \) drawn from a circularly symmetric distribution centered on the point \( \mu = (\mu_x, \mu_y) \). The challenge for the viewer is to estimate the location of the invisible center of the distribution given only the visible dots.

Each dot is an independent piece of information about the unknown center and the resulting estimation task is an example of a cue combination problem. Cue combination models typically make strong assumptions about the distributions involved. They are typically focused on estimators that are unbiased and that minimize the expected variance of the estimate (see Landy, Maloney, Johnston, & Young, 1995). If, for example, we know that the distribution is bivariate Gaussian, then the unbiased estimation rule with minimum variance is the center-of-gravity (COG) estimator, \( \hat{\mu} = (\bar{\mu}_x, \bar{\mu}_y) \), where \( \bar{\mu}_x \) is the average (mean) of the dots’ x-coordinates and \( \bar{\mu}_y \) is the average (mean) of the dots’ y-coordinates.

There are previous reports that COG is employed in visual localization tasks. Whitaker and Walker (1988) found that comparison of the horizontal offset of two dot clusters depended on COG. Morgan, Hole, and Glennerster (1990) reported that the perceived distance between two target dots located within separate dot clusters is biased toward the distance between the two clusters’ respective COGs. Morgan et al. concluded that the COG acts as a reference point for each cluster as a whole. McGowan, Kowler, Sharma, and Chubb (1998) reported that saccades to dot clusters land near the COG.

The potential pitfall of using COG is that its variance can inflate markedly if the distribution from which the sample is drawn is not Gaussian. First, consider the Gaussian case like the 100-dot cluster in Figure 1A. If instead of COG we choose to estimate the center of the cluster by taking the median values of the horizontal and vertical coordinates of the dots, the resulting horizontal and vertical estimates of the center of the cluster would have about 56% higher variance. The relative efficiency of the median observer relative to the COG observer is the ratio of the variance of the latter to the variance of the former (see Appendix A for discussion of the relative efficiency of estimators). The median has 36% lower efficiency than COG. For Gaussian clusters, the COG estimator has the highest possible efficiency (the lowest possible variance) of any estimator.

However, now consider the 100-dot cluster in Figure 1B. Its horizontal and vertical coordinates are independent samples from the Student’s t-distribution with three degrees of freedom, denoted $t_3$. The resulting cluster is superficially similar to the Gaussian in Figure 1A, but now the estimates of the median observer have about 59% higher efficiency than the estimates of the COG observer. The mean and median have exchanged roles. Part of the reason why the mean is less efficient than the median when estimating the center of $t_3$ is that the distribution tends to generate dots very far from the center of the distribution ("outliers"), which greatly affect the variance of the mean of the sample but not that of the median.

The statistician’s choice seems clear. If he knows that the sample is taken from a Gaussian, he should use the mean of the sample as an estimate of the center of the population. If he knows that the sample is taken from $t_3$, he should avoid using the mean and instead use the median or some other estimator with a higher efficiency than the mean. However, we are rarely in the position of knowing for certain that the data we confront were drawn from a specific distributional family such as the Gaussian. In 1960, the statistician John Tukey published a seminal article pointing out the remarkable consequences of our ignorance.

He showed that "…nearly imperceptible non-normalities may make conventional relative efficiencies of estimates of scale and location entirely useless…. If contamination is a real possibility (and when is it not?), neither [sample] mean nor [sample] variance is likely to be a wisely chosen basis for making estimates …” (Tukey, 1960, p. 474). Tukey’s article triggered considerable interest (see Huber, 2001) in developing what came to be known as robust estimators and the mathematical theory of robustness.

Robustness in statistics is often presented as the process of down-weighting the influence of distant “outliers” in the data. To that end, robust methods were developed to arrive at nearly optimal estimates both in the presence and in the absence of outliers (Hampel, 1974; Huber, 1981). We refer to this definition of robustness as robustness in the statistician’s sense. We use the terms outlier robust and outlier robustness in discussing it. To our knowledge, outlier robustness was introduced into the vision literature by Landy et al. (1995).

However, there is a second way to think about robustness that also has its origins in Tukey’s (1960) work. We can imagine that the stimulus is composed of dots drawn from multiple sources. We wish to ignore all but one source and estimate parameters belonging to that one source alone. With this interpretation, source-robust estimation is linked to perceptual segmentation.

Cohen, Singh, and Maloney (2008), for example, presented observers with stimuli that consisted of a large, elliptical cluster and a small, circular cluster that was placed, at least partially, beyond the perimeter of the ellipse. The observers were asked to rotate a probe pattern to match the perceived orientation of the cluster, and they were not explicitly informed of the complex nature of the stimuli. The study found that with increased lateral separation between the main, elliptical cluster and the secondary, small cluster, observers’ orientation settings approached the principal axis of the main, elliptical cluster, away from the principal axis of the entire cluster. This indicates that with greater separation, the small secondary cluster lends itself more and more to be segmented away from the main cluster into its own separate part and thus to be treated as an independent source. This down-weighting mechanism relies on detecting that there is more than one source generating the stimulus. It is a form of source robustness.

Outlier robustness does not use segmentation cues to find suspect dots, but rather down-weights the most extreme dots indiscriminately, solely because they fall at the extremes of the data set. Such a process ensures that extreme outliers will not have undue effects on one’s estimate of the parameter in question. This, however, is different from source-robust estimation in which the influence of dots is reduced to the extent that they are assigned to a second, distinct source that is not relevant to the estimation task.

In Experiment 1, we asked observers to estimate the centers of dot clusters drawn from single distributions and tested whether the visual system systematically assigns lower weights to dots further from the cluster’s center. We do so for two kinds of distributions, Gaussian and $t_3$. For the latter distribution in particular, the COG would be a poor choice of estimator as we saw above. Even for the
former, we might expect to see down-weighting of extreme dots since the visual system has no way of knowing that the dots are drawn from an “uncontaminated” bivariate Gaussian. To anticipate our results, observers’ behavior in Experiment 1 was consistent with the use of COG for both distributions. Whatever the merits of Tukey’s argument, the visual system does not use a robust rule in estimating the centers of dot clusters that are perceived as arising from a single source.

In Experiment 2, we addressed Tukey’s (1960) source separation problem by examining how observers estimate the center of dots drawn from one distribution in the presence of dots drawn from a second distribution. In Experiment 3, we report an analogue of the task in Cohen et al. (2008), using a localization task rather than an orientation task.

Before we report our experimental results, we consider one last question. Why should the visual system choose rules of combination that minimize variance? Minimum variance is a standard cost function commonly employed in both the statistics literature and also in the visual cue combination literature (Landy et al., 1995; Oruç, Maloney, & Landy, 2003). However, why should an observer in our tasks minimize variance rather than some other criterion more closely aligned with the demands of everyday tasks? If the observer uses his visual estimate as the aim point in rapidly reaching to touch a small circular disk or in throwing a ball at a basket, should he not adopt a visual rule of combination that reflects the task and maximizes the probability of success?

We conjecture that minimizing variance is precisely the criterion that is aligned with the demands of many everyday tasks. If the observer combines a large number of dots in a dot cluster by a weighted average of their spatial locations, then it is plausible that the resulting visual estimate will be close to Gaussian in form. We also know that motor error in many tasks is close to Gaussian and there are theoretical reasons why it should be (Harris & Wolpert, 1998). Combined visual and motor error is typically Gaussian in form (Trommershäuser, Maloney, & Landy, 2008) as it must be if it is simply the sum of independent Gaussian visual and motor errors.

If the goal of the observer in rapidly reaching to touch a circular target is to maximize his probability of success, then he achieves this goal by placing as much of the probability density of his visual motor error within the target area. If his visuomotor error is Gaussian, he should (a) aim for the center of the target and (b) minimize his visuomotor variance by minimizing his visual and motor variances separately. In making this argument, we ignore the effect of visual feedback on movements. However, so long as visual feedback does not alter the Gaussian form of visuomotor error, it is still minimized by minimizing the variance of the initial visual estimate. So long as visual error is Gaussian in form, then minimizing its variance automatically optimizes performance in a wide class of everyday tasks.

Experiment 1

To investigate whether observers use estimators that are outlier robust, we had observers estimate the centers of dot clusters drawn from isotropic, bivariate distributions. Each observer participated in two different conditions. The only difference between the two conditions was the underlying distribution from which the dots were drawn. In one condition, the dots were drawn from a Gaussian, and in the other condition, they were drawn from a $t_3$-distribution with three degrees of freedom, denoted $t_3$. We predicted that even if observers use a non-robust approach when estimating some property of the Gaussian cluster, they would nevertheless adopt a robust approach when estimating some property of $t_3$, because it has longer tails than the Gaussian. The need for robust methods to protect one’s estimate from outliers is more acute when there are very extreme dots present in the data.

Methods

Observers

Seven observers at New York University participated in the experiment. All were naive as to the purpose of the experiment and were paid for their participation.

Apparatus

Stimuli were displayed on a 34 cm by 27 cm LCD monitor (Dell 1708FP) with a 60-MHz refresh rate. Observers viewed the stimuli from a distance of 144 cm, and we used a chin rest to maintain the viewing distance. Thus, the screen subtended 13.5 by 10.7 degrees of visual angle. We restricted the display region to an invisible square (subtending 10.3° by 10.3°) to ensure that each stimulus was isotropic and not horizontally elongated. The experiment was programmed and run using MATLAB and the Psychtoolbox libraries (Brainard, 1997; Pelli, 1997). Responses were recorded using a keyboard.

Stimuli

Each Gaussian stimulus consisted of a single isotropic, bivariate Gaussian cluster ($\sigma = 0.89^\circ$) containing 100 dots, and each dot had a diameter of 2.4 minutes of arc. The $x$ and $y$ coordinates of the $t_3$ stimuli were independent, both drawn from a $t_3$-distribution with 3 degrees of freedom. We scaled the $t_3$ stimuli so that, on average, 90% of the dots appeared within the same region as 90% of the dots of the Gaussian stimuli. This ensured that the two types of stimuli would look roughly similar. Figure 1 shows an example of the equated stimuli.

An important concern was the possibility that dots drawn would fall beyond the edges of the display region. For draws from the Gaussian distribution, dots were
unlikely to fall outside the display region over the course of the experiment. However, draws from the $t_3$ distribution would often include a dot that was not displayed. In all of our analyses, we used only the dots that were visible to the observer.

**Design**

We used the method of constant stimuli. On each trial, we shifted the cluster horizontally to one of seven locations with respect to two collinear, vertical reference lines placed above and below the stimulus. The seven horizontal locations were: $-0.378^\circ$, $-0.252^\circ$, $-0.126^\circ$, $0^\circ$, $0.126^\circ$, $0.252^\circ$, $0.378^\circ$. In the $0^\circ$ condition, the mean of the underlying distribution (be it Gaussian or $t_3$) was aligned with the reference lines. Figure 2 shows the effect of shifting the cluster to the left and to the right of the reference lines.

There were two stimulus types and seven shift locations resulting in 14 conditions. Each condition was repeated 200 times resulting in 2800 total trials. We divided the experimental session into four blocks of 700 trials each. In two blocks, we showed observers the Gaussian clusters, and in the other two blocks, we showed them the $t_3$ clusters. Some observers were shown the Gaussian blocks first, while other observers were shown the $t_3$ blocks first. The observers were not informed of these details but rather told that the experiment was divided into four blocks to allow for easy breaks. The entire experiment, which comprised a short practice block followed by four experimental blocks, lasted approximately 1 h and 20 min.

**Procedure**

The task was two-alternative forced choice. On each trial, observers responded whether the cluster’s center was to the left or to the right of the reference lines. The instructions were given as follows: “Press the ‘z’ key with your left hand if you think that the center of the cluster is to the left of the reference lines, and press the ‘?’ key with your right hand if you think that the center of the cluster is to the right of the reference lines.”

The reference lines, which were located at the horizontal center of the screen, were displayed continuously throughout the experiment (except during the breaks between blocks). On each trial, we displayed the stimulus for 250 ms at one of the seven shift locations. Observers were allowed to respond as soon as the stimulus appeared, but the stimulus remained on the screen for 250 ms even when the observers responded while the stimulus was still on the screen. The next trial started 750 ms after the observer’s response (or 750 ms after the stimulus disappeared, if the observer responded while the stimulus was still on the screen). Figure 3 shows a schematic of the trial sequence.

**Analysis**

**Generalized linear model (GLM)**

We used a Generalized Linear Model (McCullagh & Nelder, 1989) to estimate the influence of the dots on observers’ PSE as a function of the dots’ horizontal eccentricity from the population mean. These estimates could be compared to the influences predicted by a center-of-gravity (COG) model. If observers use an estimator that is outlier robust, then, with increasing dot eccentricity, we would expect to see a decreasing influence of the dot on PSE compared to the influence predicted by the COG model. Figure 4 compares a hypothetical robust model to the simple COG model. We used the GLM to measure the influences of the dots grouped into bins. The GLM analysis treated every single
A dot within a bin as if it had the same influence on PSE regardless of the dot’s precise location within the bin. While our model is limited in resolution, by fitting enough bins we will have a large enough influence pattern to test whether observers are employing a robust strategy. Since we apply the same binning procedure in computing the influence assigned by the COG ideal observer, any effect of binning would affect data and prediction identically and therefore a comparison of actual performance and ideal would not be affected by binning.

As shown in the four examples in Figure 5, the number of dots in each bin changes naturally across trials. The GLM analysis uses the natural sampling variation of the stimuli from trial to trial to estimate the influences of the dots on PSE. It is similar in spirit to classification image approaches (Ahumada, 2002; Ahumada & Lovell, 1971; see Knoblauch & Maloney, 2008). The information that the GLM uses from each trial is the following: the observer’s response (Right or Left), the horizontally shifted location of the stimulus $s$, and the number of dots in each bin, which we call the bin counts $B_1, B_2...B_n$. We denote influence per dot in each bin by $\gamma_1, \gamma_2...\gamma_n$. The GLM links an observer’s response profile to a Gaussian psychometric function ($\Phi$) using the following equation, where $t$ denotes the trial number:

$$P[\text{Right} \mid t] = \Phi_{\mu, \sigma}(s^t + \gamma_1 B_1^t + \gamma_2 B_2^t + ... + \gamma_n B_n^t).$$  

(1)

Figure 4. Influence. The solid blue line is the influence that a single dot has on the center-of-gravity (COG) estimator as a function of the dot’s horizontal displacement from the center of the cluster. The absolute magnitude of the influence of dots increases without bounds with increasing distance from the center of the cluster, indicative of the large influence that outliers have on the COG estimator. The dashed red line is the influence that a single dot has on a hypothetical robust estimator. Characteristically, the influence of dots is bounded in absolute magnitude and decreases to zero with increasing distance from the center.
The fitting routine returns estimates for the following parameters: $\mu$, which is the observer’s overall bias; $\sigma$, which is the slope of the psychometric function; and $\gamma_1$ through $\gamma_n$, which are the influences per dot on observer’s PSE for each bin.

**Left–right symmetry**

Preliminary analyses on pilot data led us to surmise that the observers’ influence functions were odd-symmetric about 0. If so, we could get more accurate estimates of influence by forcing the GLM to return odd-symmetric influence functions. Accordingly, we first tested each observer’s data using a nested hypothesis test (Mood, Graybill, & Boes, 1974, p. 440 ff.), to see if we could reject the null hypothesis that the influence functions were odd-symmetric. We did this for different choices of number of bins and we were not able to reject the null hypothesis in any of them (lowest $p = 0.217$). Based on this conclusion, we used GLM to fit influence functions that were odd-symmetric about 0. We report only the parameters to the right of 0; those to the left are identical in absolute value and opposite in sign.

**Results**

We fitted the Gaussian data and the $t_5$ data separately and thus there were 1400 trials per data set (for each

![Figure 5. Random variation in bin count. We used the natural variation in the number of dots in each bin across different stimuli to estimate the influence of any one dot in that bin on observers’ PSE. The number of dots in bin $B_2$, for example, varies from twelve dots in stimulus (A), to eight dots in stimulus (B), to sixteen dots in stimulus (C), to eighteen dots in stimulus (D).](image-url)
We used a maximum likelihood procedure to find the best fits of the parameters, and we calculated the 95% confidence intervals of each parameter using bootstrap methods (Efron & Tibshirani, 1993).

The results of interest are the fits for the binned influences per dot on PSE for each of the two stimulus types. Figure 6 shows the fits for each observer for each of the two stimulus types with bootstrapped 95% CI. The blue dashed diagonals connect the fits for the binned influences per dot on PSE for the simple center-of-gravity (COG) model. The figure also shows the mean influence of each bin across all observers (with 95% CI), for each of the two stimulus types.

In the results presented here, we divided the stimulus space into eight bins (four to the right of 0 and four to the left of 0), which, as noted above, means that we only fit and display four influence-per-dot parameters because of the symmetry assumption. As for the length of each bin,
remember that we scaled the \( t_3 \) stimuli so that, on average, 90% of the dots appeared within the same region as 90% of the dots of the Gaussian stimuli. In the results presented here, we decided to divide the region that covers 90% of the dots into six bins and used two more bins, one on each side, to cover the remaining 10% of the dots. Thus, the length of each of the inner bins (labeled 1, 2, and 3), for both the Gaussian and the \( t_3 \) analyses, was 0.49°. There were six of these, three to the right of 0 (which are shown in the graph) and three to the left of 0 (which are not shown in the graph because they are defined symmetrically), and so the horizontal region covered by these bins was 2.9°. The length of the outermost bin (labeled 4) was 3.7°, extending from the end of bin 3 until the end of the display region, with the outermost bin on the left being defined symmetrically.

No difference was evident between observers who saw the Gaussian stimuli first (O1, O3, O4, O6, O7) and observers who saw the \( t_3 \) stimuli first (O2, O5). We ran two other observers who saw the \( t_3 \) stimuli first, but their precision (measured as the slope of the psychometric function) was much lower than the precision of the observers reported here. The fits for the influences-per-dot parameters for these two observers were similar to those of the seven observers whose results are shown, but with markedly larger confidence intervals. Hence, we omit them from the graphical summary of data.

Most observers’ fits for the influence-per-dot parameters were not significantly different from that predicted by the COG model for either of the two stimulus types (\( p > 0.05 \) based on 95% bootstrapped CI). Only one observer (O2) had significantly less influence per dot for the outermost bin than that predicted by COG for both stimulus types (\( p < 0.05 \) based on 95% bootstrapped CI). One other observer (O6) had less influence per dot for the outermost bin than that predicted by COG only for the \( t_3 \) condition and not for the Gaussian condition.

An interesting result that we did not expect was the over-weighting of the outermost bin for the Gaussian stimuli by two observers (O5 and O7). One possibility is that these observers might be basing their estimate in part on the convex hull of the cluster (the smallest convex region containing all the dots of the cluster). Such a strategy could yield very high influences for the outermost bins when fitting the data with our GLM. (It has previously been shown that saccades to highly non-uniform dot clusters tend to land at the COG of the shape implied by the dot cluster, rather than the COG of the dots themselves; see Melcher & Kowler, 1999.) These observers were evidently not using an outlier-robust method in estimating the center.

A second interesting result that we did not expect was the down-weighting of the innermost dots (bin 1) for many of the observers. We suspect that this is due to the fact that the cluster is very dense in the innermost bins. Hence, a decrease in influence per dot could occur due to crowding (McGowan et al., 1998).

**Stimulus discriminability**

When asked after the experiment, all observers reported that they did notice that the stimuli in Experiment 1 changed between the second and third blocks (which is when the stimuli switched from Gaussian to \( t_3 \) or vice versa). Despite these reports, we considered the possibility that observers used the same rule of combination for the Gaussian and \( t_3 \) distributions because they could not readily discriminate samples from the two distributions.

We ran four naive observers in a control experiment. We told them that they would see a series of samples of dot clusters of two different types and that they should attempt to identify the two types and classify them by pressing one of two keys on the computer keyboard. No further instructions or training were given.

We presented 210 examples (105 Gaussian samples and 105 \( t_3 \) samples, interleaved at random). The distributions were presented just as they were in Experiment 1. We performed a signal detection analysis (Green & Swets, 1966/1974) on the resulting data, arbitrarily declaring that onekeypress response counted as a “Gaussian” response, the other a “\( t_3 \)” response.

We estimated the sensitivity parameter \( d' \) separately for each observer based on each observer’s responses. The resulting \( d' \) could be positive or negative, depending on whether the observer chose the same assignment of distributions to keys as we did. We report the absolute values of the estimated \( d' \) values. The observers’ \( d' \) values ranged between 1.5 and 2.2 in absolute magnitude, indicating that observers could readily differentiate samples from the two distributions used in Experiment 1 even when they were interleaved.

**Discussion**

This experiment investigated whether the visual system employs estimators that are outlier robust in estimating the center of simple, uncontaminated dot clusters drawn from a single distribution. We found that only one observer deviated from the minimum variance COG rule for the Gaussian, and only that observer and a second observer showed any evidence of robustness for the \( t_3 \) distribution. We also found that there was no significant difference between the Gaussian condition and the \( t_3 \) condition across observers. Regardless of whether or not the stimuli had extreme dots, most observers’ fits for the binned influences per dot on PSE were not significantly different from that predicted by the simple COG model.

One possible explanation for observers’ lack of robustness when dealing with the \( t_3 \) distribution in Experiment 1 is that observers cannot weight a cluster’s dots differentially. In Experiment 2, we explicitly asked observers to assign differential weights to the dots, and the
results indicate that they are fully capable of doing so. Hence, we can rule out the possibility that observers’ lack of robustness in Experiment 1 was due to some sort of inability to assign differential weights to the dots in a cluster.

In the Introduction section, we broached the idea that robustness could be construed as source separation. So far we have only found observers consistently employing robust methods when the stimulus has evident spatial part structure (Cohen & Singh, 2006; Cohen et al., 2008; Denisova, Singh, & Kowler, 2006). The results of Experiment 1 do not conflict with this hypothesis since the stimuli had no evident spatial part structure. Observers had no reason to expect that the visible dots were generated by more than one process. However, what if we showed them how a complex stimulus was drawn from more than one source? Would they then employ a robust method based on source separation even in the absence of evident spatial part structure?

**Experiment 2**

Experiment 1 revealed that most observers do not employ visual estimators that are outlier robust when estimating the center of simple, uncontaminated dot clusters drawn from a single distribution. In Experiment 2, we tested whether observers employ a source-robust rule when estimating the center of a cluster that is drawn from a mixture of two distributions: a bivariate Gaussian distribution and a bivariate Uniform distribution. The observer’s task was to estimate the center of the dots drawn from the Gaussian source and ignore the dots from the Uniform source.

Each stimulus had 140 dots total. One hundred of them were drawn from a Gaussian distribution, identical to the one we used in Experiment 1, and the remaining 40 were drawn from a Uniform distribution that covered the entire display. Observers were shown during training how the stimuli were drawn from two different sources, and they were tasked with estimating the center of the Gaussian cluster.

We predicted that observers would employ a source-robust rule in down-weighting the influence of dots as a function of their increased eccentricity. The motivation for this down-weighting is that with increased eccentricity there is less of a chance that the dots arose from the Gaussian distribution. For the analysis, we used the same GLM and fitting procedure to estimate the binned influences per dot on PSE at different eccentricities. We compare human performance to that of a maximum likelihood model of an ideal source-robust estimator described below.

**Methods**

**Observers**

Eight paid observers at New York University participated in the experiment. All were naive as to the purpose of the experiment and had not participated in the first experiment.

**Apparatus**

The apparatus was the same as in the first experiment. However, the display region included the entire screen and not just a square region centered on the screen as in Experiment 1.

**Stimulus**

Each 140-dot complex stimulus consisted of a 100-dot isotropic, bivariate Gaussian cluster ($\sigma = 0.89^\circ$) and a 40-dot noise cluster that was uniformly distributed across the screen (13.5° by 10.7°). Every dot in the display had a diameter of 2.4 minutes of arc just as in the first experiment. Figure 7 shows a schematic of the stimulus.

**Design**

We used the method of constant stimuli, randomly shifting the 100-dot Gaussian cluster to one of seven locations with respect to two collinear, vertical reference lines placed at the horizontal center of the screen. The seven locations were: $-0.378^\circ$, $-0.252^\circ$, $-0.126^\circ$, $0^\circ$, $0.126^\circ$, $0.252^\circ$, $0.378^\circ$. In the $0^\circ$ condition, the mean of the underlying Gaussian distribution was aligned with the reference lines.

The experimental session was divided into five blocks to allow for easy breaks. There was only one stimulus type (the $t_3$ distribution was not included in this experiment) and seven shift locations resulting in 7 conditions. Each condition was repeated 200 times resulting in 1400 total trials. The entire experiment lasted approximately 45 min.

**Training**

At the start of the experiment, observers learned how each stimulus was drawn from two different sources. The 100-dot Gaussian cluster was colored red and the 40 noise dots were colored green. The 100-dot Gaussian cluster appeared randomly at one of the seven shift locations. During training, the stimulus remained on the screen until the observer responded. At the start of the first practice trial, observers were told that there were two clusters on the screen: “the stimulus you are seeing is a combination of two different clusters which we will call ‘cluster-1,’ colored in red, and ‘cluster-2,’ colored in green” (see
The observers were then asked to respond whether the center of cluster 1, the red one, was to the left ("z" key) or to the right ("?" key) of the reference lines. Observers did this for 10 to 15 trials while the experimenter asked if they had any trouble doing the task. All reported that it was straightforward.

The experimenter then showed a typical stimulus to the observer and switched the color of the noise dots back and forth between green and red to emphasize that the stimulus was comprised of the same two kinds of clusters (cluster 1 and cluster 2) that they were seeing before, but that now, and during the actual experiment, both clusters would have the same color (see Figure 8B). The observers were told that the task remained the same: “respond whether the center of cluster 1 is to the left or to the right of the reference lines.” After a few trials, all observers...

**Figure 8.** Experiment 2—Training and experimental task. (A) During training, observers judged the location of the center of the red (Gaussian) cluster. They were told to ignore the green (noise) dots. (B) In the actual experiment, observers continued to judge the location of the center of the Gaussian cluster, but now the noise dots were also red. They were instructed to ignore the noise dots.
reported that they understood the task, that it was difficult, but that they did not have any trouble making a response. The observers were then allowed to practice for a few more trials. Finally, observers were shown the stimuli again under color-coding for a few trials and they were told to pay attention to the two distributions. They were again informed that during the actual experiment all the dots would be colored in red.

**Procedure**

This was the same as in the first experiment, except that observers viewed a complex stimulus rather than a simple stimulus and thus the instructions, given during training (as described above), were slightly different. A 2AFC task was employed whereby observers responded whether the center of the 100-dot Gaussian cluster was to the left or to the right of the reference lines. The stimulus also contained 40 noise dots (uniformly distributed throughout the display) that the observers were supposed to ignore. The reference lines, which were located at the horizontal center of the screen, were displayed continuously throughout the experiment (except during the breaks between blocks). On each trial, we placed the 100-dot Gaussian cluster at one of the seven shift locations with respect to the reference lines, and the entire 140-dot stimulus remained on the screen for 250 ms. Observers were allowed to respond as soon as the stimulus appeared; but the stimulus remained on the screen for 250 ms even when the observers responded while the stimulus was still on the screen. The next trial started 750 ms after the observer’s response (or 750 ms after the stimulus disappeared, if the observer responded while the stimulus was still on the screen).

**Analysis**

We used the same GLM as in Experiment 1 to get the fits for the binned influences per dot on PSE. The bin count for each trial is the total number of dots in a bin irrespective of whether the dot belongs to the 100-dot Gaussian cluster or to the 40 noise dots. Importantly, this means that the expected ratio of Gaussian dots to noise dots within a bin decreases with increased bin eccentricity. We compared the fits returned from the observers’ data to the fits predicted by three different models: the *ignorant model* that makes its response based on the COG of all 140 dots, the *omniscient model* that makes its response based on the COG of just the 100-dot Gaussian cluster, and the maximum likelihood model developed below.

The influence for the ignorant model (plotted in Figure 9 as a blue dashed diagonal) increases linearly with increased eccentricity (just as in Experiment 1). However, the initial increase in influence for the omniscient model (plotted in Figure 9 as a black dashed curve) tapers off with increased eccentricity and eventually goes to zero for very high eccentricities. This is exactly what we expect because with increased bin eccentricity there is a decrease in the ratio of Gaussian dots to noise dots, and with very high eccentricities, there are unlikely to be any Gaussian dots within the bin. Hence, as the probability of Gaussian dots in the bin decreases, the influence per dot for a source-robust estimator should decrease as well.

If observers treated all dots equally, then their fits should not be significantly different from the fits predicted by the ignorant model. If, on the other hand, they were perfectly able to discriminate the two sources, as if the stimuli were presented with the color-coding from the training session, then their fits should not be significantly different from the fits predicted by the omniscient model. If they employ a source-robust rule by down-weighting
dots as a function of eccentricity simply because of the decreased chances of the dot belonging to the Gaussian cluster, then their fits should lie between the two ideal models.

**ML model**

We derived a maximum likelihood (ML) estimator for the visual estimation task (Duda, Hart, & Stork, 2000, Chap. 2). The stimulus consists of 140 dots $p_1, \ldots, p_n$, where $p_j = (p_{jx}, p_{jy})$ are the horizontal and vertical coordinates of a dot. Exactly 100 of these dots are drawn from an isotropic Gaussian distribution $\phi(p; \mu, \sigma^2)$ with unknown mean $\mu = (\mu_x, \mu_y)$ and known variance $\sigma^2$, and the remaining 40 dots are drawn from a Uniform distribution whose probability density function is $U(x, y) = C$, a constant. Let $\pi: \{1, \ldots, 100\} \rightarrow \{1, \ldots, 100\}$ be a 1–1 function that maps the integers from 1 to 100 to integers in the range 1 to 140. This function only serves to select the 100 dots that will be assigned to the Gaussian in computing likelihood. The log-likelihood for any choice of free parameters $\mu, \pi$ is

$$
\lambda(\mu, \pi) = \sum_{i=1}^{100} \log \phi(p_{\pi(i)}; \mu, \sigma^2) + C,
$$

where $C$ is the constant derived from the probability density function of the 40 dots that are assigned to the Uniform instead of the Gaussian. We wish to maximize Equation 2 by choice of the free parameters $\mu, \pi$. The resulting estimates $\hat{\mu}, \hat{\pi}$ are the maximum likelihood estimates. The estimate of concern to us is the horizontal coordinate of $\hat{\mu}$, denoted $\hat{\mu}_x$.

The maximization is made easier by the following observation. For any choice of $\mu$, the expression in Equation 2 is maximized when the function $\pi$ assigns to the Gaussian the 100 dots that are nearest to $\mu$. Thus, we need only search on $\mu$, computing Equation 2 for the 100 nearest dots.

We simulated the performance of this ML observer in Experiment 2 (using 2000 trials for each of the seven shift locations) and computed the resulting influence function. It is plotted in Figure 9 as a red dashed curve and labeled “ML, $n = 108$” because it assigns 108 dots to the Gaussian.

**Results and discussion**

Figure 9 shows observers’ mean influence per dot on PSE for each of the nine bins, together with the 95% CI of each mean. The figure also shows the predictions of the ignorant, omniscient, and ML models. The inner bins (labeled 1, 2...8) were the same length as the inner bins in the first experiment; i.e., 0.49° each. There were 16 of these (eight on each side) and so the horizontal region covered by these bins was 7.8°. Just as in Experiment 1, we were unable to reject our null hypothesis that the influence functions were odd-symmetric (lowest $p = 0.287$). Hence, we report only the binned influences to the right of 0; those to the left are identical in absolute value and opposite in sign. The outermost bin (labeled 9) was 2.85° long, extending from the end of bin 8 until the edge of the screen, with the outermost bin on the left being defined symmetrically. Remember that there were dots present throughout the screen, and thus it is easier to see the influence pattern if we fit more eccentricities than the small number of eccentricities that we fit in Experiment 1.

As shown in Figure 9, the mean of the observers’ fits lies between the ignorant and the omniscient predictions. Observers did assign a reduced influence to points in the outermost bins, suggesting that observers do employ a source-robust rule.

It is apparent from Figure 9, however, that the ML model developed above—which assigns 100 dots to the Gaussian cluster, and the remaining 40 dots to the uniform noise—does not capture the estimated influences exhibited by the observers.

A natural possibility we considered is that the restriction of assigning exactly 100 dots to the Gaussian cluster was too severe. After all, the observers likely have only a rough estimate of the number of dots in the Gaussian cluster. Therefore, we generalized the ML model in Equation 2 to

$$
\lambda(\mu, \pi; n) = \sum_{i=1}^{n} \log \phi(p_{\pi(i)}; \mu, \sigma^2) + C,
$$

where $n$ is the number of dots assigned to the Gaussian cluster (with the remaining $140 - n$ dots assigned to the uniform noise). For each value of $n$, ranging from 94 to 116, we repeated the simulation of the ML model’s performance and compared it against the mean influence function exhibited by the observers. We found that the ML model with $n = 108$ best fitted the influence function derived from observers’ responses. The simulated performance of this ML model is plotted in Figure 9 as a red solid curve and labeled “ML, $n = 108$” because it assigns 108 dots to the Gaussian.

This analysis indicates that the ML model of an ideal source-robust estimator does a good job of capturing observers’ performance; observers, however, seem to assign more dots to the Gaussian cluster than it actually contains.

**Experiment 3**

In Experiment 2, we found that the visual system is source robust when estimating the center of a Gaussian dot cluster presented together with noise dots: it down-weights
the influence of dots likely to have come from a second distribution. In Experiment 3, we presented observers with two-part stimuli that were similar to those used by Cohen et al. (2008). We asked observers to estimate the center of the same 100-dot Gaussian cluster that we used in Experiments 1 and 2, and we contaminated each stimulus by adding a small 15-dot cluster at one of 19 offsets from the center of the main cluster. Observers were shown during training how the stimuli were drawn from two different sources, and they were tasked with estimating the center of the Gaussian cluster. One crucial difference between this experiment and Cohen et al.’s experiment is that in this experiment the small cluster could be placed at or near the center of the larger cluster. In their experiment, however, the small cluster always fell near or outside the apparent boundary of the larger cluster. We return to this point in the discussion.

Methods

Observers

Five paid observers at New York University participated in the experiment. All were naive as to the purpose of the experiment and had not participated in either of the first two experiments.

Apparatus

The apparatus was the same as in Experiment 2.

Stimulus

Each 115-dot complex stimulus was comprised of a 100-dot isotropic, bivariate Gaussian cluster ($\sigma = 0.89^\circ$) and a small 15-dot contamination cluster uniformly distributed across a small square region ($0.89^\circ$ by $0.89^\circ$) that was placed at one of several offsets from the center of the main cluster. Every dot in the display had a diameter of 2.4 minutes of arc just as in the first two experiments.

Design

Because the focus of this experiment was to test the effect of varying degrees of overlap between the stimulus’ two clusters, we express the location of the small cluster as its distance in standard deviations away from the center of the main cluster. The main cluster (i.e., the 100-dot Gaussian) was identical to the one we used in the first two experiments ($\sigma = 0.89^\circ$). The length of the small contamination cluster (which was drawn from a squared uniform distribution of $0.89^\circ$ by $0.89^\circ$) was equal to the size of the standard deviation of the main cluster. Hence, by simply stating how many standard deviations (SDs) away the small cluster’s center was from the main cluster’s center, one could quickly surmise how much overlap in SD units there was between the two clusters. For example, placing the small cluster 1.5 SDs away meant that the 15-dot cluster extended from 1 SD to 2 SDs on the main cluster. When the small cluster was placed 0.5 SD away, it fell in a dense region of the main cluster. On the other hand, when the small cluster was 4 SDs away, the two clusters were readily separable.

As shown in Figure 10, the small cluster was placed at one of 19 eccentricities from the center of the main cluster. Nine of these eccentricities were to the left of the main cluster, nine to the right, and one at the center of the main cluster ($SD = 0$). The eccentricities to the left were the same as the eccentricities to the right, and for the analysis and results, we collapsed the data after testing and failing to reject our null hypothesis that the influence of the small cluster on observers’ PSE was odd-symmetric (lowest observer’s $p = 0.57$). For example, we do not distinguish between offsetting the small cluster 2 SDs to the left and offsetting it 2 SDs to the right. Hence, for simplicity, we describe the eccentricities of only 10 offsets keeping in mind that each one (save for $SD = 0$) corresponds to both an offset to the left and an offset to the right of the main cluster. The 10 eccentricities (in SD units) were: 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5.

We used the method of constant stimuli by shifting each stimulus type to one of 13 locations with respect to two collinear, vertical reference lines that were placed above and below the stimulus. The 13 locations were: $-0.942^\circ$, $-0.785^\circ$, $-0.628^\circ$, $-0.471^\circ$, $-0.314^\circ$, $-0.157^\circ$, 0°, 0.157°, 0.314°, 0.471°, 0.628°, 0.785°, 0.942°. In the 0° condition, the mean of the underlying Gaussian distribution was aligned with the reference lines.

The experimental session was divided into 10 blocks to allow for easy breaks. There were 10 stimulus types (i.e., the 10 offsets of the small cluster with respect to the center of the main cluster) and 13 shift locations (i.e., the 13 shifts of the entire stimulus with respect to the reference lines) resulting in 130 conditions. Each condition was repeated 20 times resulting in 2600 total trials. The entire experiment lasted approximately 1 h and 20 min.

Training

Training was done in the same way as the training in Experiment 2, except that for this experiment the observers were tasked with ignoring the small color-coded contamination cluster. Figure 11 shows an example of a stimulus with and without color-coding.

Procedure

This was the same as in the second experiment. A 2AFC task was employed whereby observers responded whether the center of the 100-dot Gaussian cluster was to the left or to the right of the reference lines. The stimulus also contained a 15-dot contamination cluster that the observers
were supposed to ignore. The reference lines, which were located at the horizontal center of the screen, were displayed continuously throughout the experiment (except during the breaks between blocks). On each trial, we placed the 100-dot Gaussian cluster at one of the 13 shift locations with respect to the reference lines, and the 15-dot contamination cluster at one of 10 eccentricities from the center of the main cluster (see Design section for details). The entire 115-dot stimulus remained on the screen for 250 ms. Observers were allowed to respond as soon as the

Figure 10. **Experiment 3**—Construction of the complex stimuli. Each stimulus consisted of a large Gaussian cluster presented together with a small contamination cluster. The large circle marks the approximate area of the large 100-dot Gaussian cluster. The horizontally aligned black squares mark the centers of the 19 locations where the small 15-dot contamination cluster could be placed. On each trial, the contamination cluster was placed at only one of the 19 eccentricities. In this figure, the contamination cluster is offset 2.5 SDs away from the center of the large cluster. The dashed square surrounding the contamination cluster marks the boundaries of the small Uniform distribution from which the contamination cluster was drawn at that offset.

Figure 11. **Experiment 3**—Training and experimental task. (A) During training, observers judged the location of the center of the large, red (Gaussian) cluster. They were told to ignore the small, green (contamination) cluster. (B) In the actual experiment, observers continued to judge the location of the center of the large Gaussian cluster, but now the contamination dots were also red. They were instructed to ignore the contamination dots.
stimulus appeared, but the stimulus remained on the screen for 250 ms even when the observers responded while the stimulus was still on the screen. The next trial started 750 ms after the observer’s response (or 750 ms after the stimulus disappeared, if the observer responded while the stimulus was still on the screen).

Analysis

Each of the 10 stimulus types was analyzed separately to compute each observer’s PSE at each of the 10 eccentricities. To that end, we recorded the percentage of trials that the observer responded “Right” for each of the 13 shift locations and fit this data to a Gaussian psychometric function using a maximum likelihood procedure.

We used each observer’s PSE at $SD = 0$ as a measure of the observer’s overall bias. We then compared the observer’s PSEs at the other nine eccentricities (i.e., $SD \geq 0.5$) to the observer’s overall bias to measure the influence of the small cluster on observer’s PSE at each eccentricity. Figure 12 shows the mean influence of the small square region for any choice of $\alpha^x$ and $\alpha^y$, with $115 - n(\alpha^x)$ being the number of dots that lie outside the square region. The ML estimator assigns exactly 100 dots to the Gaussian cluster and exactly 15 dots to the contamination cluster. Furthermore, it is not allowed to assign to the contamination cluster any dots that lie outside the square region. Hence, we rule out from the solution space all locations $\alpha^x$ where $n(\alpha^x) < 15$ (by assigning a likelihood of 0, i.e., a log-likelihood of $-\infty$).

For all locations $\alpha^x$ where $n(\alpha^x) = 15$, the ML estimator assigns the 15 dots that lie inside the square region to the contamination cluster and the 100 dots that lie outside the square region to the Gaussian cluster. For all locations $\alpha^x$ where $n(\alpha^x) > 15$, the ML estimator must assign some of the dots ($n(\alpha^x) - 15$ to be precise) that lie inside the square region to the Gaussian cluster, because there are less than 100 dots lying outside the square region. In the ML model derived for Experiment 2, we noted that the likelihood is maximized by assigning to the Gaussian cluster the dots that are nearest to $\mu^x$. So, from all the dots lying inside the square region, the ML estimator assigns to the Gaussian the $n(\alpha^x) - 15$ dots that are nearest to $\mu^x$ so that, together with the dots lying outside the square region, there are exactly 100 dots assigned to the Gaussian. This assignment maximizes likelihood for any given choice of $\mu^x$ and $\alpha^x$.

Let $\pi: \{1, \ldots, 115\} \xrightarrow{1:1} \{1, \ldots, 115\}$ be a 1–1 function that maps the integers from 1 to 115 to the integers in the range 1 to 115. This function only serves to
select the 100 dots that will be assigned to the Gaussian in computing likelihood. The log-likelihood for any choice of free parameters $\alpha^x, \mu^x$ is

$$
\lambda(\alpha^x, \mu^x) = \sum_{i=1}^{100} \log p_{\theta[i]}(x^i, 0, \sigma^2) + C, \quad (4)
$$

where $C$ is the constant derived from the probability density function of the 15 dots that are assigned to the contamination cluster instead of the Gaussian. We wish to maximize Equation 4 by choice of free parameters $\alpha^x, \mu^x$. The resulting estimates $\hat{\alpha}^x, \hat{\mu}^x$ are the maximum likelihood estimates. The estimate of concern to us is $\hat{\mu}^x$.

We simulated the performance of this ML observer in Experiment 3 (using 200 trials for each of the 130 conditions) and computed the resulting influence function. The influence predicted by this ML model is very close to that predicted by the omniscient model and it is plotted in Figure 12 as a red dashed horizontal.

Results and discussion

Figure 12 allows us to compare the results of the ignorant, omniscient, and ML models, to the experimentally measured influence of the small cluster on observers’ PSE as a function of eccentricity. The observers’ data lie between the ignorant and omniscient predictions indicating that observers do partially discount the effect of the small contamination cluster—that is, they exhibit source robustness. When the small cluster is very far away from the main cluster ($SD = 3.5$), its mean influence on observers’ PSE is not significantly different from 0 and the prediction of the omniscient model. This is hardly surprising. We included these conditions to verify that the influence of the small cluster disappears if it is very far away from the main cluster. On the other hand, when the small cluster partially overlaps the main cluster ($SD = 1.5$ to $SD = 3$), its mean influence on observers’ PSE is less than that predicted by the ignorant model but greater than that predicted by the omniscient and ML models. This indicates that, although observers do down-weight the influence of dots that are likely to belong to the small contamination cluster, their ability to separate sources falls short of that of the ML observer. Human observers did not separate sources so well when the small cluster was offset about 1.5 to 2.5 SDs away from the center of the Gaussian.

When the small cluster is near the center of the Gaussian (0.5 SD and 1.0 SD), its mean influence on observers’ PSE is not significantly different from 0 and the prediction of the omniscient model. The dots in the small cluster are evidently not outliers when they are so close to the center of the Gaussian, and an outlier-robust estimator would not reduce their influence. Yet human observers do so. This pattern of performance demonstrates that human observers are not simply detecting outliers and reducing their influence. They behave as we would expect a source-robust estimator to behave.

The ML model predictions for all conditions were indistinguishable from the omniscient model. In effect, the ML model can readily locate the small cluster and remove its effect almost perfectly at every offset. In contrast, human observers have difficulty doing so for offsets from 1.5 to 2.5 SDs. To determine why, we examined examples of stimuli with the small cluster at 0.5 SD, 2.0 SDs, and 3.5 SDs.

When the eccentricity of the small cluster is 3.5 SDs, the ML observer and the human observer can readily segment the small cluster from the large and discount its effect. The ML observer, omniscient observer, and human observer are in good agreement. When the eccentricity is 0.5 SD, the presence of the small cluster is betrayed by a marked increase in density and a concomitant decrease in dot spacing near the small cluster. The human observer can discount the effect of the small cluster almost perfectly.

In the remaining case ($SD = 2.0$), the small cluster falls in a sparser region of the Gaussian. In objective terms, adding the small cluster increases local dot density well above what could be expected from the Gaussian alone. As a consequence, the ML observer has little difficulty in locating the small cluster and reducing its influence almost completely. However, we found that the small cluster was almost invisible when it is offset 2.0 SDs. We could identify whether it fell on the right or left side of the center of the Gaussian, but to the unalerted observer there was little to indicate that it was there at all. We conjecture that the observed deviations between human performance and performance of the ML observer are due to difficulties in detecting the small cluster with intermediate offsets. This masking effect deserves further study given the marked discrepancy between observed human and ML performances.

Conclusions

There are previous studies that suggest that the visual system employs robust estimators in depth cue combination (Girshick & Banks, 2009). Our goal was to evaluate whether the visual system’s estimates of the centers of dot clusters (a form of cue combination where each dot is a cue) are based on detection of outliers (outlier robust) and/or segmentation of dots by source (source robust).

We reported three experiments. In all of them, we used a method analogous to classification images to estimate the influence of each visible dot on the observer’s rule of combination.

In Experiment 1, we sought to determine whether the rule of combination used by human observers systematically
assigned less weight to dots far from the center of the sample, that is, whether the rule of combination is outlier robust. We asked observers to estimate the centers of dot clusters drawn from one of two distributions, a bivariate Gaussian and a $t_3$-distribution with 3 degrees of freedom, denoted $t_3$. An outlier-robust rule would assign less weight to extreme dots generated by either distribution. We found that observers did not give less weight (influence) to potential outliers for either choice of distribution. Their influence estimates were indistinguishable from those of the center-of-gravity (COG) rule.

For the Gaussian, the COG is not outlier robust, but it is the unbiased, minimum variance estimator (see Appendix A). We could argue that the observer, although not outlier robust, has chosen a “good” estimation rule. For the $t_3$, however, observers’ estimates were also not outlier robust, and for this distribution, the COG has markedly higher variance than other unbiased rules of estimation such as the median. The COG is not a “good” estimation rule for the $t_3$, at least in the sense of minimizing variance.

We then considered a different way of framing the problem of robust estimation, namely, as source separation. A key step in perceptual organization is to segment visual data according to source and make visual estimates concerning the properties of each. In Experiment 1, dot clusters were drawn from a single source and observers combined all available information with equal weight into an estimate of the center of the single source. What if instead we made it clear that there were two sources of visual data and asked observers to separate them, returning estimates of the center of just one of the sources?

The kinds of stimuli and task that we used are evidently artificial, but they are standard in studies of localization (see, for example, Battaglia & Schrater, 2007; McGowan et al., 1998; Tassinari, Hudson, & Landy, 2006). Yet, while Experiment 1 (and many previous studies) only measured human ability to combine visual information from multiple sources, Experiments 2 and 3 measured human ability to combine information from multiple sources while suppressing information from other sources (cf. Cohen, Schnitzer, Gersch, Singh, & Kowler, 2007). Given the typical clutter in everyday scenes, the ability to segment sources of visual information in this way is a great advantage to the organism.

In Experiment 2, we presented dot clusters drawn from a mixture of two distributions: a bivariate Gaussian as in Experiment 1 and a large bivariate Uniform distributed uniformly across the entire display. The construction of the stimuli as a mixture of two distributions was explained and demonstrated to observers during training before the experiment. The mixture distribution was inspired by the contamination mixtures used by Tukey (1960).

Observers were asked to estimate the center of the Gaussian while ignoring dots from the Uniform. We found that observers now gave less weight to potential outliers that were more likely to be from the Uniform. Observers’ performance in segmenting the distributions matched that of an ideal maximum likelihood (ML) observer. The overall pattern of influence resembled that of a robust estimator, what we might have expected to find in Experiment 1 at least for the $t_3$ distribution but did not find.

In Experiment 3, we presented dot clusters drawn from a mixture of two distributions: a bivariate Gaussian as in Experiment 1 and a small bivariate Uniform placed on the horizontal axis bisecting the Gaussian. The resulting stimulus was a mixture of a large cluster (Gaussian) and a small cluster (Uniform) similar to the complex stimuli of Cohen et al. (2008). The construction of the stimuli as a mixture of two distributions was explained and demonstrated to observers during training before the experiment.

The ideal ML observer for our task could almost perfectly discount the presence of the small cluster at any offset from the large cluster. Observers could match this ideal performance when (i) the small cluster was near the center of the large cluster and (ii) when the small cluster was far from the center of the large cluster. In between, however, observers failed to fully discount the small cluster, suggesting that they may have difficulty detecting or localizing it. If so, this segmentation failure deserves further investigation.

Human ability to discount the influence of the small cluster when it was far from the center of the large cluster is consistent with outlier robustness or source robustness. However, human ability to discount the influence of the small cluster when it was embedded near the center of the large cluster indicates that observers were source robust as these dots were not outliers.

In summary, we find little indication that the visual system uses robust estimators in estimating the centers of dot clusters that are perceived as arising from a single source (Experiment 1). However, in Experiments 2 and 3 we found that observers can segment scenes by source and discount the effect of data from one source on estimates of the properties of a second source. In Experiment 2, we obtained influence functions that looked like those for an outlier-robust estimator. Dots further from the center of the Gaussian cluster received less weight. However, we showed that human performance also mimicked that of a source-robust rule based on an ML model that classified dots by source.

Our results indicate that, at least for estimates of centers of dot clusters, the visual system does not employ outlier-robust estimators. Instead, it segments dots by source with the recognition that contamination of the data corresponding to one source is just data belonging to some other source. We conclude that robust methods employed by the visual system are closely tied to mechanisms of perceptual segmentation. These mechanisms themselves cannot be captured by robust statistics in large part because standard robust methods rely primarily on standardized residuals. The visual system, in contrast, uses more sophisticated
generative models for “objects” in performing segmentation (see Cohen et al., 2008). The robust behavior exhibited by the visual system appears to rely on the prior operation of mechanisms of perceptual segmentation.

We have focused on separation of sources, but we can also consider a further question. Once observers have learned to interpret dot clusters as a superposition of clusters from two sources, can they ignore that interpretation? For example, if we asked them to estimate the centers of the stimuli in Experiment 2 or 3, but now to take into account all the dots, not just those from Source 1, could they do so, or would they continue to down-weight those that plausibly arise from the second source? This question is worthy of future investigation.

## Appendix A

### Statistical terminology

In parametric estimation, the typical goal is to estimate one or more parameters $\theta_1, \ldots, \theta_m$ of a parameterized distribution with probability density function $f(x; \theta_1, \ldots, \theta_m)$ given a sample $X_1, \ldots, X_n$ from that distribution. A familiar distribution is the Gaussian $f(x; \mu, \sigma^2)$ with parameters $\mu$ and $\sigma^2$. Any function of the sample $T(X_1, \ldots, X_n)$ is an estimator, and since it is a function of random variables, any estimator is a random variable itself. A familiar example is the mean

$$T(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n}.$$  \hfill (A1)

### Unbiased estimator

An unbiased estimator of a parameter $\theta_i$ has expected value $E[T(X_1, \ldots, X_n)] = \theta_i$. Intuitively, it gets the right answer “on average.” For the Gaussian distribution, the sample mean is an unbiased estimator of the parameter $\mu$ also called the population mean. Of course, the population mean is also the population median and the population mode, and ultimately, it just marks the center of the distribution. The median of the sample is a second example of an unbiased estimator of the center of the Gaussian.

### Minimum variance unbiased estimator

Among unbiased estimators of a parameter, there may be one whose variance is less than the variance of all other estimators. For samples drawn from a Gaussian distribution, for example, the mean is the minimum variance unbiased estimator of the parameter $\mu$.

## Relative efficiency

The relative efficiency of two estimators is the ratio of the variance of the second to the variance of the first (note the order). If the second has a lower variance, then the efficiency is less than 1. Efficiency is a measure of the loss incurred by using the first estimator in place of the second for a particular sample size from a particular distribution. Tukey (1960) considered the relative efficiency of the median relative to the mean when estimating the center of contaminated Gaussian distributions. Most of the dots in a sample are drawn from a Gaussian with unknown mean $\mu$ and known variance $\sigma^2$, but a small proportion $p$ are drawn from a distribution with the same mean but higher variance. In the following example, we use $16\sigma^2$ (or $4\sigma$).

### The mean is not robust

While the mean is the minimum variance unbiased estimator for the Gaussian case, even small amounts of contamination can inflate its variance to the point that other unbiased estimators such as the median have lower variance. To illustrate this point, we computed the efficiency of the median relative to the mean for a sample of size 100 taken from the contaminated Gaussian just described. Figure A1 shows the relative efficiency of the median as a function of the proportion of contamination in the sample. The relative efficiency of the median is about 0.64 when no contamination is present. That is, the mean has about 36% lower variance than the median. With 10% contamination, however, the mean and median exchange roles, the relative efficiency is about 1.39, and now it is the median that has about 28% lower variance than the mean.

![Figure A1](Image317x183 to 575x299)

Figure A1. Relative efficiency. The efficiency of the median relative to the mean for a contaminated Gaussian is plotted as a function of the proportion of contamination in the sample (see Appendix A for details). When the relative efficiency of the median is 1, that means that the mean and the median have the same variance. When the relative efficiency is less than 1, then the median has a higher variance than the mean. When the relative efficiency is greater than 1, then the median has a lower variance than the mean.
Tukey’s paradox

Tukey (1960) argued that other estimators such as trimmed means would outperform the mean even when the degree of contamination was so slight that its presence or absence could not be detected in the sample. These alternative estimators can outperform the mean (i.e., relative efficiency > 1) when dealing with samples drawn from contaminated Gaussians, while maintaining a relative efficiency that is only slightly less than 1 when dealing with samples drawn from a true Gaussian. They are robust to failures of distributional assumptions. Since we can never be certain that empirical data are drawn from an uncontaminated Gaussian, we can never justify the use of the mean instead of its robust cousins when analyzing data.

Acknowledgments

This research was funded in part by CCF-0541185 (MS) and a Humboldt Research Award (LTM). We thank Dylan Simon for helpful discussions and suggestions.

Commercial relationships: none.
Corresponding author: Mordechai Z. Juni.
Email: mjuni@nyu.edu.
Address: Department of Psychology, New York University, 6 Washington Place, Room 275, New York, NY 10003, USA.

Footnotes

1See Appendix A for explanations of statistical terms used.
2Tukey (1960) only considered large sample, asymptotic behavior. We have verified his conclusions for samples as small as 5 as well.
3Cohen et al. (2008) use the term part-based robustness. In this paper, we consider the more general notion of source robustness.
4Maximizing log-likelihood is equivalent to maximizing likelihood. We maximize log-likelihood because it is convenient computationally.

References


