Location-Spread Displays of Many Point Estimates and Their Associated Measures of Dispersion

M. Peruggia, The Ohio State University
J. Hsu, The Ohio State University

Technical Report No. 730

January, 2004

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247
Location-Spread Displays of Many Point Estimates and Their Associated Measures of Dispersion

Mario Peruggia and Jason Hsu
Department of Statistics
The Ohio State University
Columbus, OH 43210-1247
January 2004

Authors’ Footnote
Mario Peruggia is Associate Professors and Jason Hsu is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210. This material is based upon work supported by the National Science Foundation under Grant No. SES-0214574. We thank Tao Wang for analyzing the Blader et al. (2001) data set.
Abstract

We consider the problem of representing graphically a large number of point estimates with their associated measures of accuracy or dispersion. Because of clutter, traditional graphical summaries become visually ineffective when more than a few pairs of point estimates and associated dispersions have to be represented. The location-spread display introduced in this article overcomes the limitations stemming from visual clutter and can represent effectively very many pairs. The construction of the display is first presented in the context of a multiple comparisons application. The use of the location-spread representation is then illustrated with two additional examples. In the first example, the representation is used to summarize aspects of the posterior distributions of numerous parameter contrasts in a hierarchical Bayes model. In the second example, the representation is used to summarize thousands of gene expression estimates arising from an analysis of microarray data.

Keywords: Hierarchical Bayes models, Mean-mean scatterplot, Microarrays, Midpoint-radius representation, Multiple comparisons.

1 Introduction

In many statistical applications inferential conclusions are summarized numerically by a collection of point estimates. Along with each estimate a measure of accuracy or dispersion is typically reported and often the two are combined to form some type of interval estimate. Perhaps most commonly this circumstance occurs when sample means and standard errors are computed for several groups of observations. The calculated numerical values of the statistics can of course be presented in a table. Ideally, the numbers should be sorted in some meaningful fashion, either on the basis of the values themselves or on the basis of available covariate information.

When there are very many values, even if good care has been used in constructing and ordering the table, extracting information from the numbers and deriving any insightful interpretation can be a daunting task. Quite naturally, one might turn to graphical methods in the hope of obtaining a
synoptic representation of the numbers that is more readily amenable to making comparisons and drawing meaningful conclusions.

Several graphical representations are available. Some of them are general-purpose representations that are essentially equivalent to a table listing. For example, a display often encountered in the medical literature is built by arranging sequentially, along the $x$-axis, vertical bars with heights proportional to the point estimates. The top of each bar is then endowed with a whisker of length proportional to the precision of the point estimate. A slight improvement on this theme has the bars and whiskers replaced by vertical segments. Each segment is centered at a point with $y$-coordinate equal to the corresponding point estimate and extending on either side of its center by an amount proportional to the precision of the point estimate. (An asymmetric interval would be drawn when different amounts of dispersion are assessed on either side of a point estimate of location.)

The major limitation of these types of displays is that the arrangement of the bars and whiskers (or of the intervals) along the $x$-axis is not directly related to either of the quantities of principal interest, namely the point estimates and the associated estimates of precision/dispersion. Although the ordering might be meaningfully tied to a third dimension of interest, such as the values of a measured covariate, the effectiveness of such displays is limited by the fact that the positioning of a plotted entity (a bar and whisker or an interval) is not uniquely determined by the corresponding point estimate and precision. In other words, two identical point estimate and precision pairs will result in two equally shaped entities being plotted at different, arbitrary locations in the display.

Other types of representations have been developed with specific applications in mind, although they could conceivably be employed in different, related settings as well. Two examples developed to summarize graphically Tukey’s method of multiple comparisons are the comparison circles of Sall (1992) and the mean-mean scatterplot of Hsu and Peruggia (1994). In the standard application to multiple comparisons, the point estimate is the estimated pairwise contrast corresponding to the difference between two of a set of $k > 2$ treatments and the precision estimate is the half-width of the corresponding,
multiplicity-adjusted, symmetric confidence interval. In both of these graphical representations there is a precise relation between the locations and dimensions of the entities being plotted and the values of the corresponding point and precision estimates, with the relation exploited by the mean-mean scatterplot leading to a more direct display of the estimated contrasts and confidence intervals. Both representations, however, become cluttered and break down visually when the number of entities to be plotted is larger than about ten.

The limitations described in the previous paragraphs affect most representations. Either a direct relation between the plotted entities and the point and precision estimates is lost, or some type of relation is maintained at the expense of a loss of visual effectiveness when many entities are plotted. From a practical perspective, these limitations are caused by the size of the graphical entities that are being plotted (bars, circles, intervals, etc.).

In Section 2 we use the *midpoint-radius* graphical representation of interval arithmetic introduced in Kulpa (2003) to develop a display, the *location-spread* display, that overcomes these limitations. Because the entities being plotted are points, this display allows for the effective representation of a large number of pairs of point and precision estimates or derived interval estimates. Because there is a precise relation between the plotted points and the corresponding pairs of point and precision estimates, essential visual connections can be readily made and overall interpretative summaries can be elaborated. In particular, for the case of an interval estimate, there is a natural way to relate the plotted point to a visualization of the corresponding interval and to determine if such an interval contains a given reference value (such as zero). Color can be used effectively to convey additional covariate information. In this section we illustrate the use of the display in the multiple comparisons context by considering an application to a well known data set on fuel efficiency. This gives us the opportunity to examine the pros and cons of the display and compare them to those of the mean-mean scatterplot in a situation where a limited number of location-spread pairs are being considered.

In Sections 3 and 4 we present two applications illustrating the great potential benefits of the
location-spread representation when very many location-spread pairs are being considered. The first application deals with a set of examination scores from pupils attending 38 inner London schools that is analyzed in Spiegelhalter, Thomas, Best, and Gilks (1996). The hierarchical Bayes model fit in Spiegelhalter et al. (1996) contains 38 school specific random effects that are utilized to derive the distributions of the school rankings. Here, we focus on the distributions of very many contrasts based on the school random effects and show how to employ the location-spread representation to construct an effective visual display. The second application is to the analysis of microarray data. In this example, the location-spread representation is used to display confidence intervals for the difference in gene expressions of 19641 genes under normal and infected conditions, normalized with respect to the average differential expression of five housekeeping genes. In Section 5 we summarize the benefits of the location-spread representation, discuss its limitations, and outline some additional applications and possible improvements. An elementary R code implementation of a function to draw a location-spread plot is presented in Appendix A.

2 The Location-Spread Display

The location-spread display is based on the midpoint-radius representation introduced in Kulpa (2003) to describe interval arithmetic. A summary review is contained in Hayes (2003). The midpoint-radius representation hinges on a mapping of the set of compact intervals into the Euclidean upper half-plane \( \mathbb{R} \times \mathbb{R}^+ \) defined as follows.

**Definition 2.1** Let \( \mathcal{I} \) denote the set of compact, non-degenerate, interval subsets of the real line \( \mathbb{R} \). For any interval \( I = [a, b] \in \mathcal{I} \), the midpoint-radius mapping is defined as

\[
g(I) = \left( \frac{a + b}{2}, \frac{b - a}{2} \right).
\]

Figure 1 illustrates the mapping and can be used as a guide to understand some of the subsequent developments.
There are several features of the midpoint-radius mapping that make it potentially useful in various applications. The following definitions will help us to state succinctly some of these features. Given two real numbers \( k \) and \( s \), with \( s \neq 0 \), let \( \ell(k; s) \) denote the ray of slope \( s \) emanating from the point \((k, 0)\) and extending into \( \mathbb{R} \times \mathbb{R}^+ \). For \( s = \infty \), let \( \ell(k; \infty) \) denote the vertical ray emanating from \((k, 0)\) and extending into \( \mathbb{R} \times \mathbb{R}^+ \).

For the purpose of building and interpreting our location-spread displays, we have identified the following three features as being most relevant.

**Feature 1.**

The function \( g \) maps any interval into a point with abscissa equal to the midpoint of the interval and ordinate equal to the half-width of the interval.

**Feature 2.**

For a given interval \( I = [a, b] \in \mathcal{I} \), \( g(I) \) coincides with the point of intersection between the two rays \( \ell(a; 1) \) and \( \ell(b; -1) \).

**Feature 3.**

Let \( [a, b] \in \mathcal{I} \) and \( [c, d] \in \mathcal{I} \) denote the pre-images through \( g \) of any two points \( P_1 \) and \( P_2 \) belonging to \( \mathbb{R} \times \mathbb{R}^+ \) and not vertically aligned, and let \( k \in \mathbb{R} \) and \( s \in \mathbb{R} \setminus \{0\} \) be given. Then, the condition

\[
\frac{b - a}{a + b - k} = \frac{d - c}{c + d - k} = s
\]

is satisfied if and only if \( P_1 \) and \( P_2 \) lie on the ray \( \ell(k/2; s) \).

Feature 1 is simply a restatement of the definition of the midpoint-radius mapping and its usefulness is apparent. Because points can be drawn as small graphical entities (as compared to the graphical entities needed to draw intervals), a display that makes use of the midpoint-radius mapping can represent a large number of intervals and still remain intelligible.

The usefulness of Features 2 and 3 (which are immediate to verify) is perhaps less obvious but just as relevant. As a consequence of Feature 2, by tracing the lines with slopes 1 and \(-1\) through a given point in \( \mathbb{R} \times \mathbb{R}^+ \) and identifying their intersections with the horizontal axis it is possible to determine
the endpoints of the interval constituting the pre-image under \( g \) of the given point. Closely related is a straightforward means of determining if a given reference value \( r \) is contained in the interval constituting the pre-image of a given point in \( \mathbb{R} \times \mathbb{R}^+ \). In fact, all points included in the closure of the quadrant of vertex \((r, 0)\) and delimited by the two diagonal rays \( \ell(0; -1) \) and \( \ell(0; 1) \) have interval pre-images that contain \( r \). All remaining point in \( \mathbb{R} \times \mathbb{R}^+ \) have interval pre-images that do not contain \( r \). Often, a reference value of interest is \( r = 0 \). Then, to determine if the interval pre-image of a given point in \( \mathbb{R} \times \mathbb{R}^+ \) contains or does not contain \( r = 0 \), it is enough to ascertain the location of the point relative to the quadrant defined by the main diagonal rays emanating from \((0, 0)\). For example, in Figure 1, the point \(((a + b)/2, (b - a)/2)\) lies inside the quadrant and its interval pre-image \([a, b]\) contains \( r = 0 \) while the point \(((c + d)/2, (d - c)/2)\) lies outside the quadrant and its interval pre-image \([c, d]\) does not contain \( r = 0 \).

As a consequence of Feature 3, compact interval subsets of the real line whose half-widths are proportional to the distance of their centers from a fixed reference point \( k/2 \in \mathbb{R} \) are mapped into points that line up along a ray and vice versa. The constant of proportionality coincides with the slope of the ray and the intersection of the ray with the horizontal axis is given by the point \((k/2, 0)\).

We can now give a formal constructive definition of the location-spread display.

**Definition 2.2** Given a set \( S = \{(l_i, v_i)\}_{i=1}^N \) of location estimates \( l_i \in \mathbb{R} \) and associated measures of variability \( v_i \in \mathbb{R}^+ \), the location-spread display is constructed as the scatterplot of the points in \( S \), with \( l_i \) plotted along the horizontal axis and \( v_i \) plotted along the vertical axis.

When the location and variability estimates under consideration are the centers and the half-widths of a set of symmetric confidence intervals, the location-spread display is nothing but the graph of the images of the confidence intervals under the midpoint-radius mapping.

Figure 2, generated using the sample R code presented in Appendix A, exhibits the location-spread display of simultaneous confidence intervals for the Fuel data contained in the Splus data frame `fuel.frame`. These are data on 117 makes of cars published in the April 1990 issue of Consumer Re-
ports. The factor Type classifies makes of cars into six general categories: Small, Sporty, Compact, Medium, Large, and Van. For labeling purposes, in Figures 2 and 4, these six categories are denoted by S, X, C, M, L, and V, respectively. To analyze how the response variable Fuel (which represents the gallons of fuel consumed by each make of car to travel 100 miles) is affected by the factor Type, we fit a one-way anova model and computed Tukey-Kramer 95% simultaneous confidence intervals for all pairwise comparisons. Because the factor Type has six levels, this yields $(6 \times 5)/2 = 15$ intervals. The points in Figure 2 are all plotted to the right of the vertical line through the origin because we chose to consider the 15 differences in which the smaller estimated effect is subtracted off from the larger, so that all intervals are centered at a positive value.

We used color and task-specific graphics traits to emphasize the elements outlined in Feature 2 above. In particular, the background color of the quadrant delimited by the two main diagonals through the origin is a pale yellow (this is equivalent to the use of shading in Figure 4 of Hayes 2003). The four points falling inside the quadrant represent non-significant differences in mean fuel consumption according to car type (the corresponding intervals contain zero). All remaining points lie outside the quadrant and represent significant differences. Comparisons with reference values other than zero is facilitated by the drawing of equally spaced rays of slope 1 to the right of the origin and of slope $-1$ to the left of the origin. Each ray is color-coded according to a heat color map, going from pale yellow to dark red, based on how far from the origin the ray’s intersection with the horizontal axis is. Following the rays in Figure 2 we can see, for example, that only one interval lies entirely to the right of the reference value $r = 1$.

In situations like this, where the number of location-spread pairs to be drawn is small, there may be specialized displays that provide equally or even more effective representations. For the purpose of comparison we consider the mean-mean scatterplot representation of Hsu and Peruggia (1994). The mean-mean scatterplot of the 15 simultaneous confidence intervals corresponding to the 15 estimated positive differences for the Fuel data is presented in Figure 3. In the Xlispstat interactive implementation
developed by Hsu and Peruggia, the intervals are represented as segments whose opposite tips are color-coded so as to match up with different levels of a factor. The convention used is that a given segment corresponds to the confidence interval for the mean response difference between the level coded by the color of the right tip and the level coded by the color of the left tip. For example, the top segment in Figure 3 represents the interval for the mean difference between car types \textit{Van} (coded in red) and \textit{Large} (coded in orange). Within the mean-mean plane (identified by the two blue diagonal axes with the light gray vertical bisecting line) each segment is centered at the point of intersection of two color-coded lines corresponding to the estimated effects for the two factor levels involved in the comparison. Each segment is drawn perpendicularly to the gray vertical line and the scaling is chosen so that a segment will intersect the vertical gray line if and only if the corresponding confidence interval contains zero. The details of the implementation and additional interactive features are described in Hsu and Peruggia (1994).

Because the displayed intervals correspond to the 15 positive estimated differences, all segments are centered to the right of the gray vertical line. The four intervals for the four non-significant differences are the only ones intersecting the gray vertical line. Following the color coding, it is immediate to verify that they correspond (from top to bottom) to the following four comparisons: \textit{Van} – \textit{Large} (red – orange), \textit{Large} – \textit{Medium} (orange – yellow), \textit{Medium} – \textit{Compact} (yellow – green), and \textit{Compact} – \textit{Sporty} (green – cyan). These four intervals correspond to the four points plotted inside the pale yellow quadrant in the location-spread representation of Figure 2. Unfortunately, in the location-spread representation, there is no simple way of determining what difference a point corresponds to, other than direct labeling. This can be easily accomplished using the function \texttt{identify()} interactively in R or Splus. For example, in Figure 2, we used direct labeling to identify the four non-significant comparisons. However, direct labeling of all plotted points would render the graph too cluttered and hard to decipher. This drawback of the location-spread representation is a price that we willingly pay in order to gain the ability to represent very large numbers of location-spread pairs in situations such as those illustrated in
the next two sections.

Color (and/or different plotting symbols) can be used effectively to obviate in part to this shortcoming. For example, if one wishes to emphasize one or more subsets of comparisons, a subset-specific color can be used to plot the points corresponding to the comparisons in each subset (with all remaining points plotted in black). This is done in Figure 2, where all comparisons involving car type Small are plotted in magenta. These are the same comparisons that are highlighted (using a thicker drawing line) in the mean-mean scatterplot of Figure 3.

3 An Application to School Comparison

In this section we illustrate the use of the location-spread display to represent a large number of comparisons concerning inner London schools. The comparisons are based on a Bayesian hierarchical model developed in Spiegelhalter et al. (1996) using a subset of the data originally analyzed in Goldstein, Rasbash, Yang, Woodhouse, Pan, Nuttall, and Thomas (1993). The response variable is an examination achievement score, averaged over study subjects, collected on 1978 pupils attending 38 different schools. There are both pupil level and school level covariates available. The pupil level covariates are gender and two measures of intake achievement given by the score on a common reading test taken at age 11 and by a categorical classification into one of three groups based on a verbal reasoning test also taken at age 11. The school level covariates are two categorical classification based on gender intake (3 levels: girls, boys, mixed) and denomination (4 levels: Church of England, Roman Catholic, State school, other).

Spiegelhalter et al. (1996) use the available covariates to specify a normal linear regression model for the examination achievement scores. The model includes fixed effect terms and random effect terms modeling variation at the school level. The model for the variation at the pupil level assumes a linear regression structure for the logarithm of the precision based on each pupil’s score on the reading test. The prior distributions are specified in such a way as to reflect vague prior knowledge. For each school
the regression model contains a random intercept term, \( \alpha_{1,j}, j = 1, \ldots, 38 \), that Spiegelhalter et al. (1996) regard as the residual school effect after adjusting for the covariates at both levels.

Based on a set of \( M \) MCMC draws from the posterior distributions of these school specific intercepts, Spiegelhalter et al. (1996) construct estimates of the posterior marginal distribution of each school’s ranking. Specifically, given the 38 school specific intercepts \( \alpha^{(m)}_{1,j} \) produced at iteration \( m \) of the MCMC algorithm, \( m = 1, \ldots, M \), they obtain a vector \( r^{(m)} = (r^{(m)}_1, \ldots, r^{(m)}_{38}) \) of school rankings by sorting the \( \alpha^{(m)}_{1,j} \) values and setting \( r^{(m)}_j = \text{rank} \left( \alpha^{(m)}_{1,j} \right), j = 1, \ldots, 38 \). Finally, the \( M \) vectors of school rankings constructed at the various iterations are combined to obtain an empirical estimate of the posterior distribution of the rankings. Marginally, the posterior distribution of the rank of each school is summarized by a point estimate and a 95% credible interval.

The school specific intercepts can also be used to estimate contrasts of interest involving specific schools. For the purpose of illustration we consider estimation of the posterior distributions of all pairwise contrasts \( \alpha_{1,j} - \alpha_{1,k}, j \neq k, j = 1, \ldots, 38, k = 1, \ldots, 38 \). The empirical distribution of each contrast can be simply constructed by combining the differences \( \alpha^{(m)}_{1,j} - \alpha^{(m)}_{1,k}, m = 1, \ldots, M \), based on the output at iteration \( m \) of the MCMC algorithm. This produces a total of \( 38 \times 37 = 1,406 \) non-degenerate empirical distributions that we would like to summarize graphically. There is, of course, redundancy in this set of distributions because \( \alpha^{(m)}_{1,j} - \alpha^{(m)}_{1,k} = - \left( \alpha^{(m)}_{1,j} - \alpha^{(m)}_{1,k} \right) \), but eliminating this redundancy would still leave us with 703 distributions, a very large number given that we wish to present a graphical summary of all of them at once.

We use the location-spread display to construct a global summary of interesting aspects of all distributions as follows. First, for each contrast \( \alpha_{1,j} - \alpha_{1,k} \), we determine the 0.025 quantile, \( q_{j,k}(0.025) \), the median, \( q_{j,k}(0.5) \), and the 0.975 quantile, \( q_{j,k}(0.975) \), of the corresponding empirical distribution. If the median is positive, we summarize the distribution by the location-variability pair given by \( (l_{j,k} = q_{j,k}(0.5), v_{j,k} = q_{j,k}(0.5) - q_{j,k}(0.025)) \). If the median is negative, we summarize the distribution by the location-variability pair given by \( (l_{j,k} = q_{j,k}(0.5), v_{j,k} = q_{j,k}(0.975) - q_{j,k}(0.5)) \). The
location-variability pairs thus determined are then plotted to construct the location-spread display of Definition 2.2. The resulting graph is shown in Figure 4.

There are a few features of Figure 4 worth mentioning. First, we note that the 95% credible intervals \([q_{j,k}(0.025), q_{j,k}(0.975)]\) need not be symmetric about \(q_{j,k}(0.5)\). Yet, by a straightforward generalization of Feature 1, our choice of summary location and variability measures guarantees that if a point in the location-spread display lies outside of the pale yellow quadrant, then the 95% credible interval does not contain zero, and vice versa. Thus, this choice of \((l_{j,k}, v_{j,k})\) pairs accomplishes the important goal of conveying direct visual information about whether or not a given contrast is significant.

Figure 4 represents summaries of all 1,406 non-degenerate contrasts (including the redundant ones). To reduce the blob-like effect of overplotting we used a very small plotting symbol. The foremost visual message conveyed by the figure is that the vast majority of points lie inside the pale yellow quadrant. This is an indication that most differences between schools are of little importance, confirming the statement in Goldstein et al. (1993) that “few schools can be separated reliably.”

Simple graphical devices allow us to use the location-spread display to concentrate easily on contrasts involving specific schools. In Figure 4 we employed a larger plotting character to emphasize all pairwise comparisons of the type \(\alpha_{1,17} - \alpha_{1,k}, k \neq 17, k = 1, \ldots, 38\), involving School 17. School 17 is the school that attains the second to last posterior median ranking using the approach of Spiegelhalter et al. (1996), with only School 5 attaining a worse posterior median ranking. This is reflected in the fact that all but one of the location estimates for the contrasts under consideration are negative and, consequently, all but one of the larger points in the display are plotted to the left of the vertical line through the origin. However, only a small number of these contrasts have a posterior distribution that is concentrated away from zero, as can be seen from the fact that most of the points fall inside the pale yellow quadrant.

The other graphical device employed in Figure 4 to emphasize specific comparisons with School 17 is color, which is used to code the school denomination of the other school entering the comparison (blue for Church of England, purple for Roman Catholic, olive green for State school, and tan for
other). From the color coding some inferences can be readily made at a visual level. For example, it is immediate to notice that School 17 (a Church of England school) is not significantly different than any of the schools in the other category. Also, interestingly, the most prominent difference is the one with another Church of England school, corresponding to the blue point toward the left hand side of the display.

4 A Microarray Application

Living organisms need proteins to function. Genes are translated into proteins, after first being transcribed into messenger RNA (mRNA). Microarrays allow the expression of thousands or tens of thousands of genes to be screened simultaneously at the transcription stage, based on the abundance of mRNA in samples.

Possible uses of gene expression analysis from microarray data include the following.

• **Designer medicine.** One potential use is to tailor medicine or treatment to individuals, taking gene expressions as explanatory or predictor variables. Microarrays may be used to profile individual patients genetically, to classify patients into more refined disease categories, or to predict an individual’s predisposition to a certain disease.

• **Patient targeting.** The populations from which the samples are drawn might also be subpopulations of patients who have or have not had adverse reactions to a particular molecular entity. Some molecular entities that can potentially benefit many patients fail to get approval to be marketed as drugs because a small but significant percent of the patients experience adverse reactions. If the subpopulation prone to adverse reactions can be identified in terms of differential gene expressions, then, subject to not giving the drug to this subpopulation, a molecular entity can perhaps be approved in order to benefit the vast majority of the patients.

• **Drug discovery.** The populations from which the samples are drawn may be healthy or diseased
tissues. A third possible use of the comparison of gene expressions is to find protein targets that might intervene in the disease process.

Currently, the most popular form of statistical analysis of microarray data is to infer the presence of differential expressions. This form of inference is often presented in terms of p-values for testing the null hypotheses of equality of gene expressions. However, p-values from tests of equality are non-informative regarding the magnitude of the differences. Hsu, Chang, and Wang (2004b) proposed that simultaneous confidence intervals for differential expressions be given instead. A graphical representation of the confidence intervals would be useful.

Measurements of gene expressions may not have equal variances (Churchill 2001), so that a large estimated differential expressions may or may not be statistically significant. Jin, Riley, Wolfinger, White, Passador-Gurgel, and Gibson (2001) proposed the “volcano plot” to represent both the estimated magnitude of the differential expressions and their statistical significance. The horizontal axis of a volcano plot is log₂ of the estimated ratio of the expressions under the two conditions for a gene, while the vertical axis is log₁₀ of the p-value for testing the null hypothesis of no differential expression for that gene.

However, because p-values are non-informative regarding the magnitude of the differences, it is impossible to discern from a volcano plot which genes are significantly down-regulated by at least two folds, and which genes are up-regulated by more than three folds, for example. Such information can be represented by our location-spread plot, as illustrated below.

Blader, Manger, and Boothroyd (2001) employed 2-channel cDNA microarrays to investigate changes in gene expression in human foreskin fibroblasts (HFFs) in response to toxoplasma gondii infection. Each of the microarray in their experiment was spotted with approximately 22,000 known genes and uncharacterized expressed sequence tags (ESTs). Samples of mRNA were taken at 1 hour, 2 hours, 4 hours, 6 hours, 14 hours, and 24 hours after infection and reverse transcribed to their complementary DNA (cDNA). The cDNAs from infected cells were then stained with green dye (Cy3), while the cDNAs
from uninfected cells were stained with red dye (Cy5). The cDNA samples were then allowed to bind (hybridize) to the cDNA of the known genes and the ESTs spotted on the microarrays. A scanner subsequently read separately the red and green intensities of each spot on the microarray, providing for the gene spotted there a measure of its expression in each of the two samples. We analyzed gene expressions from 14 hours after infection, because that is the only time point at which the sample gene expressions were measured with duplicate microarrays. Gene expressions, represented as we said by red and green light intensities, were measured on a scale from 0 to $2^{16} - 1$. To analyze the gene expression data, we first took $\log_2$ of the measured intensities, as is customary.

Measured gene expressions are not only potentially affected by infection, but also by unequal amounts of cDNA spotted on different arrays. Thus, each log intensity was further normalized by subtracting from it the average log intensities measured on that array under the same condition. For each gene, the mean difference in log intensity between normal and infected conditions was then estimated using the linear model in Wolfinger, Gibson, Wolfinger, Bennett, Hamadeh, Bushel, Afshari, and Paules (2001), separately for each gene (allowing for different variances).

Housekeeping genes are genes serving maintenance functions and, therefore, should be active (expressed) under all conditions. Each may have different expression levels under different conditions. However, if the differences in expression levels among a group of housekeeping genes remain relatively constant under a variety of conditions, then their average expression level may serve well as a control expression level. (Hsu, Chang, and Wang 2004a suggest that this is similar to using a placebo control in clinical trials.)

Using Real-Time Polymerase Chain Reaction (RT-PCR), which is more accurate than microarray gene expression analysis, Vandesompele, De Preter, Pattyn, Poppe, Van Roy, De Paepe, and Speleman (2002) identified nine such housekeeping genes, five of them happen to have been included in the Blader et al. (2001) experiment. Thus, we refined the estimated differential expressions by subtracting the average of the estimated housekeeping gene differentials from each of the estimated differential
expressions.

Due to the small number of replications for most genes in the Blader et al. (2001) data, it is unrealistic to compute accurate simultaneous confidence intervals. Instead, Figure 5 shows the location-spread display constructed by plotting each estimated differential gene expression versus three times its estimated standard error. The five housekeeping genes are highlighted. This illustrates how one can easily use the location-spread display to judge which genes have estimated differential expressions larger than three times their standard errors. More importantly, one can see the potential of the location-spread display to represent simultaneous confidence intervals, when microarray experiments are designed with enough replications for such confidence intervals to be reliably calculated.

5 Discussion

In this article we introduced a graphical tool, the location-spread display, that can be used to represent simultaneously a very large number of location estimates and associated measures of accuracy and we illustrated its effectiveness in three substantive applications. There are, of course, many other situations in which the display might prove useful. Besides gene expression analysis, other applications in bioinformatics may also involve a large number of inferences that one wishes to represent graphically. In drug discovery, for example, thousands or tens of thousands of molecular compounds may be screened for activities against cancer cell lines. We have also found the display very useful for illustrating to medical researchers the impact of the common practice of dichotomizing a continuous response and fitting a logistic regression model for the derived 0 – 1 variable. Such a display can be constructed by dichotomizing the response at several different levels and plotting selected estimated coefficients against the half-widths of their corresponding confidence intervals. By doing so, it often becomes apparent that the coefficient for a given term might be significant at one cutoff level and non-significant at a different level. The display can also illustrate clearly that, in many cases, accuracy decreases as changes in the cutpoint render the grouping of the observations more unbalanced. In particular, on account of
Feature 3, if the rate of increase of the standard errors is roughly linear in the size of the estimates, the points in the display will tend to fall along straight lines.

Quite naturally, the method has some limitations. As previously mentioned, direct labeling of the points is required for proper identification and this can often be accomplished effectively through the use of different colors and plotting characters. If many of the point estimates share a common estimated accuracy (which often happens for balanced designs), then the display will present unappealing horizontal streaks. Clearly, overplotting may become an issue when thousands of points are plotted. Reducing the size of the plotting character and increasing the size of the display is often helpful. However, more specialized plotting techniques, such as binning (Carr 1991), will be required to deal with very large sets of location-spread pairs.

In conclusion, we observe that the location-spread display is a convenient graphical summary device, with several features that make it a powerful exploratory data analysis tool. In particular, as noted above, Feature 3 is useful to identify situations in which point estimates exhibit a constant coefficient of variation. This is accomplished by looking for straight lines in a plot of the point estimates against (some multiple of) their standard errors. One can conjecture, for example, that Figure 5 hints at the presence of several such straight lines. Determination of some corresponding subsets of genes might be a useful aspect to consider to further refine the model. Incidentally, in a possible interactive implementation of the display, identification of interesting subsets of location-accuracy points could be aided by the addition of a brushing tool.
location.spread.plot <-

function(loc.in,  # vector of location values
    sprd.in,  # vector of spread values
    # loc.in and sprd.in must be of equal length
    ww = pretty(c(loc.in,sprd.in))[length(pretty(c(loc.in,sprd.in)))],  
    # nice plotting range
    n.l = 10,  # number of equally spaced, 45 degree reference rays
    # to be drawn. Set n.l <= 20 for best results
    col.plot = rep(1,length(loc.in)),  # vector of plotting colors
    pch = 16,  # plotting character
    xlab = "Location",  # x-axis label
    ylab = "Spread",  # y-axis label
    main = "Location-Spread Plot",  # main title
    pin = 3  # height and half-width of the plot in inches
    ){  

# store the current values of the graphics parameters

    op <- par(pin=c(2*pin,pin))

# set up the axes and labels

    plot(loc.in,sprd.in,xlim=c(-ww,ww),ylim=c(0,ww),
        xlab=xlab,ylab=ylab,main=main,type="n")

# set up the colors for the 45 degree reference rays according to a heat
# color map

    c.offset <- floor(n.l/4)
    col.l <- heat.colors(n.l+2*c.offset)[(n.l+c.offset):(c.offset+1)]

# Make the background color of the quadrant delimited by the two main
# diagonals through the origin a pale yellow

    a <- 0
    b <- 0

d <- ww/(n.l-1)

x <- c(a,a+ww,a-ww,a)
y <- c(b,b+ww,b+ww,b)
polygon(x,y, xpd=FALSE, col = heat.colors(30)[29],
     lty=2, lwd=2, border = NA)

# Draw the 45 degree reference rays
for(j in seq(1,n.l)){
x <- c(a+(j-1)*d,ww)
y <- c(0,ww-(j-1)*d)
lines(x,y,lty=1,col=col.l[j])
x <- c(a-(j-1)*d,-ww)
y <- c(0,ww-(j-1)*d)
lines(x,y,lty=1,col=col.l[j])
}

# Draw a vertical reference ray through the origin
lines(c(0,0),c(0,ww),col=col.l[n.l])

# Plot the location-spread points
points(loc.in,sprd.in,pch=pch,col=col.plot)

# Restore the original values of the graphics parameters
par(op)
}
References


Newsletter 3(1), 27–32.


Figure 1: Illustration of the Midpoint-Radius Mapping. A segment is mapped into the vertex opposite the base of an isosceles triangle with 45 degree base angles. The base of the triangle is given by the segment itself.
Figure 2: The Location-Spread Display for the Fuel Consumption Data. For each point in the display the location dimension is the center of a Tukey-Kramer simultaneous confidence interval and the spread dimension is the half-width of the interval. The comparisons involving Small cars are highlighted.
Figure 3: The Mean-Mean Display for the Fuel Consumption Data. The display represents all Tukey-Kramer simultaneous confidence intervals and numerical summaries for the comparisons involving Small cars.
Figure 4: The Location-Spread Display for the School Comparison Data. For each point in the display the location dimension is the median of the posterior distribution of a pairwise contrast and the spread dimension is either the median minus the 0.025 quantile or the 0.975 quantile minus the median, depending on whether the point is to the right or to the left of the vertical ray through the origin. The contrasts involving School 17 are highlighted and color-coded according to the denomination of the other school entering the comparison.
Figure 5: The Location-Spread Display for the Gene Expression Data. For each point in the display the location dimension is the estimated differential gene expression and the spread dimension is three times the standard error. The five housekeeping genes are highlighted.