PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 4361–4383
doi:10.1098/rsta.2009.0120

Statistical inference for exploratory data analysis and model diagnostics

By Andreas Buja[1], Dianne Cook[2,*], Heike Hofmann[2], Michael Lawrence[3], Eun-Kyung Lee[4], Deborah F. Swayne[5] and Hadley Wickham[6]

[1] *Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2] *Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA*
[3] *Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, Seattle, WA 98109, USA*
[4] *Department of Statistics, EWHA Womans University, 11-1 Daehyun-Dong, Seodaemun-Gu, Seoul 120-750, Korea*
[5] *Statistics Research Department, AT&T Laboratories, 180 Park Avenue, Florham Park, NJ 07932-1049, USA*
[6] *Department of Statistics, Rice University, Houston, TX 77005-1827, USA*

We propose to furnish visual statistical methods with an inferential framework and protocol, modelled on confirmatory statistical testing. In this framework, plots take on the role of test statistics, and human cognition the role of statistical tests. Statistical significance of 'discoveries' is measured by having the human viewer compare the plot of the real dataset with collections of plots of simulated datasets. A simple but rigorous protocol that provides inferential validity is modelled after the 'lineup' popular from criminal legal procedures. Another protocol modelled after the 'Rorschach' inkblot test, well known from (pop-)psychology, will help analysts acclimatize to random variability before being exposed to the plot of the real data. The proposed protocols will be useful for exploratory data analysis, with reference datasets simulated by using a null assumption that structure is absent. The framework is also useful for model diagnostics in which case reference datasets are simulated from the model in question. This latter point follows up on previous proposals. Adopting the protocols will mean an adjustment in working procedures for data analysts, adding more rigour, and teachers might find that incorporating these protocols into the curriculum improves their students' statistical thinking.

Keywords: permutation tests; rotation tests; statistical graphics; visual data mining; simulation; cognitive perception

*Author for correspondence (dicook@iastate.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsta.2009.0120 or via http://rsta.royalsocietypublishing.org.

One contribution of 11 to a Theme Issue 'Statistical challenges of high-dimensional data'.

1. Introduction

Exploratory data analysis (EDA) and model diagnostics (MD) are two data analytic activities that rely primarily on visual displays and only secondarily on numeric summaries. EDA, as championed by Tukey (1965), is the free-wheeling search for structure that allows data to inform and even to surprise us. MD, which we understand here in a narrow sense, is the open-ended search for structure not captured by the fitted model (setting aside the diagnostics issues of identifiability and influence). Roughly speaking, we may associate EDA with what we do to raw data *before* we fit a complex model and MD with what we do to transformed data *after* we fit a model. (Initial data analysis (IDA) as described in Chatfield (1995), where the assumptions required by the model fitting are checked visually, is considered a part of, or synonymous with, EDA.) We are interested here in both EDA and MD, insofar as they draw heavily on graphical displays.

EDA, more so than MD, has sometimes received an ambivalent response. When seen positively, it is cast as an exciting part of statistics that has to do with 'discovery' and 'detective work'; when seen negatively, EDA is cast as the part of statistics that results in unsecured findings at best, and in the over- or misinterpretation of data at worst. Either way, EDA seems to be lacking something: discoveries need to be confirmed and over-interpretations of data need to be prevented. Universal adoption of EDA in statistical analyses may have suffered as a consequence. Strictly speaking, graphical approaches to MD deserve a similarly ambivalent response. While professional statisticians may resolve their ambivalence by resorting to formal tests against specific model violations, they still experience the full perplexity that graphical displays can cause when teaching, for example, residual plots to student novices. Students' countless questions combined with their tendencies to over-interpret plots impress on us the fact that reading plots requires calibration. But calibrating inferential machinery for plots is lacking and this fact casts an air of subjectivity on their use.

The mirror image of EDA's and MD's inferential failings is confirmatory statistics' potential failure to find the obvious. When subordinating common sense to rigid testing protocols for the sake of valid inference, confirmatory data analysis risks using tests and confidence intervals in assumed models that should never have been fitted, when EDA before, or MD after, fitting could have revealed that the approach to the data is flawed and the structure of the data required altogether different methods. The danger of blind confirmatory statistics is therefore 'missed discovery'. This term refers to a type of failure that should not be confused with either 'false discovery' or 'false non-discovery', terms now often used as synonyms for 'Type I error' and 'Type II error'. These confirmatory notions refer to trade-offs in deciding between *pre-specified* null hypotheses and alternative hypotheses. By contrast, 'missed discovery' refers to a state of blindness in which the data analyst is not even aware that alternative structure in the data is waiting to be discovered, either in addition or in contradiction to present 'findings'. Statistics therefore needs EDA and MD because only they can force unexpected discoveries on data analysts.

It would be an oversimplification, though, if statistics were seen exclusively in terms of a dichotomy between the exploratory and the confirmatory. Some parts of statistics form a mix. For example, most methods for non-parametric modelling and model selection are algorithmic forms of data exploration, but some

are given asymptotic guarantees of finding the 'truth' under certain conditions, or they are endowed with confidence bands that have asymptotically correct coverage. Coming from the opposite end, confirmatory statistics has become available to ever larger parts of statistics due to inferential methods that account for multiplicity, i.e. for simultaneous inference for large or even infinite numbers of parameters. Multiplicity problems will stalk any attempt to wrap confirmatory statistics around EDA and MD, including our attempt to come to grips with the inferential problems posed by visual discovery.

The tools of confirmatory statistics have so far been applied only to features in data that have been captured algorithmically and quantitatively and our goal is therefore to extend confirmatory statistics to features in data that have been discovered visually, such as the surprise discovery of structure in a scatterplot (EDA), or the unanticipated discovery of model defects in residual plots (MD). Making this goal attainable requires a re-orientation of existing concepts while staying close to their original intent and purpose. It consists of identifying the analogues, or adapted meanings, of the concepts of (i) test statistics, (ii) tests, (iii) null distribution, and (iv) significance levels and p-values. Beyond a one-to-one mapping between the traditional and the proposed frameworks, inference for visual discovery will also require considerations of multiplicity due to the open-ended nature of potential discoveries.

To inject valid confirmatory inference into visual discovery, the practice needs to be supplemented with the simple device of duplicating each step on simulated datasets. In EDA, we draw datasets from simple generic null hypotheses; in MD, we draw them from the model under consideration. To establish full confirmatory validity, there is a need to follow rigorous protocols, reminiscent of those practised in clinical trials. This additional effort may not be too intrusive in the light of the inferential knowledge acquired, the sharpened intuitions and the greater clarity achieved.

Inference for visual discovery has a pre-history dating back half a century. A precursor much ahead of its time, both for EDA and MD, is Scott *et al.* (1954). Using astronomical observations, they attempted to evaluate newly proposed spatial models for galaxy distributions (Neyman *et al.* 1953), by posing the following question: 'If one actually distributed the cluster centres in space and then placed the galaxies in the clusters exactly as prescribed by the model, would the resulting picture on the photographic plate look anything like that on an actual plate…?' In a Herculean effort, they proceeded to generate a synthetic $6^\circ \times 6^\circ$ 'plate' by choosing reasonable parameter values for the model, sampling from it, adjusting for 'limiting magnitude' and 'random 'errors' of counting', and comparing the resulting 'plate' of about 2700 fictitious galaxies with a processed version of the actual plate, whose foreground objects had been eliminated. This was done at a time when sampling from a model involved working with published tables of random numbers, and plotting meant drawing by hand—the effort spent on a single instance of visual evidence is stunning! The hard work was of course done by 'computers', consisting of an office with three female assistants whose work was acknowledged as requiring 'a tremendous amount of care and attention'. (The plots, real and synthetic, are reproduced in Brillinger's 2005 Neyman lecture (Brillinger 2008), albeit with undue attributions to Neyman. Scott *et al.* (1954) acknowledge Neyman only 'for his continued interest and for friendly discussions'.) Much later, when computer-generated

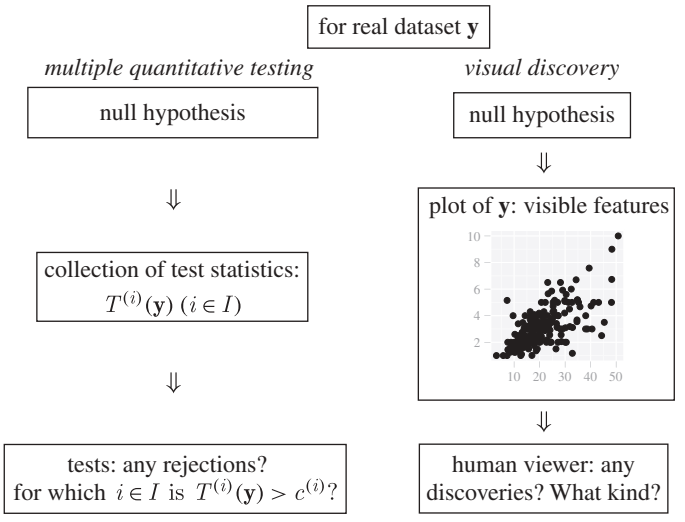

Figure 1. Depiction of the parallelism between multiple quantitative testing and visual discovery. Potential features of data that can be visible in the plot are thought of as a stylized collection of test statistics. The actually observed features in a plot (the 'discoveries') correspond to the test statistics that result in rejection.

plotting and sampling started to make headway into data analytic practice, there are more examples, and also several voices urging analysts to gauge their sense for randomness by looking at plots of random data. Daniel (1976) has 40 pages of null plots in his book on statistics applied to industrial experiments. Diaconis (1983) describes 'magical thinking' as the natural human tendency to over-interpret random visual stimuli. Davison & Hinkley (1997) in their book on bootstrap methods (§4.2.4) recommend overlaying lines corresponding to quantiles from random samples, of the same size as the data, to normal probability plots. The practice was implemented in an early visualization system, Dataviewer, and described in Buja *et al.* (1988; in the 'Informal Statistical Inference' section), some of which was revisited in the XGobi system (Swayne *et al.* 1992; Buja *et al.* 1996, 1999). Recently, it is Bayesians who have systematized the idea of MD as visualizing datasets simulated from statistical models (Gelman 2004 and references therein). In spatial statistics, Brillinger (2008) can be cited for talks that keep alive the idea of 'synthetic plots', plots of data simulated from a complex model. In a related tenor are the ideas of Davies (2008), which are discussed further in §5. Also relevant to this paper is the discussion, comparison or even recommendation of good practice of visual methods for data analysis, for which there exists a rich literature (e.g. Tufte 1983; Cleveland 1993; Buja *et al.* 1996; Card *et al.* 1999; Wilkinson 1999; Chen *et al.* 2007).

In §2, we first outline the parallelism between established tests of quantitative features and proposed inference for qualitative discovery. In §3, we briefly mention approaches to reference distributions from which 'null datasets' can be sampled (with further details deferred to the electronic supplementary material). We then discuss protocols that specify how simulated null datasets are to be used to attain inferential validity (§§4 and 5). In the remaining sections, we illustrate our preferred protocol with several practical examples.

2. Discoveries as rejections of null hypotheses

Here, we outline a parallelism between quantitative testing and visual discovery. The steps are depicted in figure 1. The initial step is to take seriously the colloquial identification of the term 'discovery' with 'rejection of a null hypothesis'. This is entirely natural for MD, where the model constitutes the null hypothesis, and any model defect found with diagnostics can be interpreted as rejection of the model in the sense of statistical testing. It requires some elaboration for EDA, because data analysts may not usually think of their discoveries as rejections of null hypotheses. The 'discovery' of skewness, for example, in univariate distributions can be taken as a rejection of symmetry or normality, and the discovery of an association can be taken as the rejection of independence. Such null hypotheses help sharpen the understanding of a 'discovery', because they provide a canvas of unstructured data situations upon which to judge the discovered structure.

In EDA, the same null hypothesis can be rejected for many reasons, i.e. in favour of many possible alternatives. For example, the null assumption of independence between two variables can be rejected by the discovery of linear, nonlinear but smooth, discontinuous or clustered association. Similarly, in MD the fitted model can be rejected for many reasons; in standard linear models, such reasons may be nonlinearities, skew residual distributions, heterogeneous error variances or lurking variables. In the parallelism, this can be interpreted to be that many 'discoveries' can be contrasted against a background provided by the same null hypothesis.

One arrives quickly at a critical divergence between quantitative testing and visual discovery: quantitative testing requires the *explicit prior specification* of the intended 'discoveries'; by contrast, the *range of visual discoveries* in EDA and MD is *not pre-specified explicitly*. This difference is critical because the absence of prior specification is commonly interpreted as invalidating any inferences as post hoc fallacies. This interpretation is correct if what is criticized is the naive tailoring of a quantitative statistical test to a previously made qualitative discovery on the same data, as when the discovery of two clusters is 'confirmed' with a post hoc two-sample test. We address this by stylizing the set of discoverable features in a plot as a *collection of test statistics*, call them $T^{(i)}(\mathbf{y})$ $(i \in I)$, where $\mathbf{y}$ is the dataset and I is as yet a nebulous set of all possible features. Each test statistic $T^{(i)}(\mathbf{y})$ measures the degree of presence of a feature in the data to which the human viewer of the plot may respond. This collection of discoverable features, and thus, test statistics, is (i) potentially very large and (ii) not pre-specified. Of these two issues, the second is the more disconcerting because it appears to be fatal for statistical inference.

A way to address pre-specification, for a given type of plot, would be to form a list as comprehensive as possible of discoverable features and to formalize them in terms of test statistics. Such an approach has indeed been attempted for scatterplots by Tukey & Tukey (1985) who coined the term 'scagnostics'; more recently Wilkinson *et al.* (2005) revived the idea with a list of features that includes 'outlying', 'skewed', 'clumpy', 'sparse', 'striated', 'convex', 'skinny', 'stringy' and 'monotonic'. Although these were not framed as formal test statistics, they are defined quantitatively and could be used as such under any null distribution that does not have these features. Yet, any

such list of features cannot substitute for the wealth and surprises latent in real plots. Thus, while cumulative attempts at pre-specifying discoverable features are worthwhile endeavours, they will never be complete. Finally, because few data analysts will be willing to forego plots in favour of scagnostics (which in fairness was not the intention of either group of authors), the problem of lack of pre-specification of discoverable features in plots remains as important and open as ever.

Our attempt at cutting the Gordian Knot of prior specification is by proposing that there is no need for pre-specification of discoverable features. This can be seen by taking a closer look at what happens when data analysts hit on discoveries based on plots: they not only register the occurrence of discoveries, but also describe their nature, e.g. the nature of the observed association in a scatterplot of two variables. Thus data analysts reveal what features they respond to and hence, in stylized language, which of the test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ resulted in rejection. In summary, among the tests that we assume to correspond to the possible discoveries but which we are unable to completely pre-specify, those that result in discovery/rejection will be known.

The next question we need to address concerns the *calibration of the discovery process* or, in terms of testing theory, the control of Type I error. In quantitative multiple testing, one has two extreme options: for marginal or one-at-a-time Type I error control, choose the thresholds $c^{(i)}$ such that $P(T^{(i)}(\mathbf{y}) > c^{(i)} \mid \mathbf{y} \sim H_0) \leq \alpha$ for all $i \in I$; for family-wise or simultaneous Type I error control, raise the thresholds so that $P(\text{there exists } i \in I : T^{(i)}(\mathbf{y}) > c^{(i)} \mid \mathbf{y} \sim H_0) \leq \alpha$. False discovery rate control is an intermediate option. Pursuing the parallelism between quantitative testing and visual discovery further, we ask whether the practice of EDA and MD has an equivalent of Type I error control. Do data analysts calibrate their declarations of discovery? Do they gauge their discoveries to guarantee a low rate of spurious detection? They usually declare discoveries by relying on past experience and trusting their judgement. In clear-cut cases of strong structure, dispensing with explicit calibration is not a problem, but in borderline cases there is a need to calibrate visual detection without resorting to the pseudo-calibration of post hoc quantitative tests tailored to the discovery.

We argue in favour of a protocol that attacks the problem at the level of plots as well as data analysts' reactions to plots. We propose to consider data analysts as black boxes whose inputs are plots of data and whose outputs are declarations of discoveries and the specifics thereof. To calibrate the discovery process, simultaneously for all discoverable features $T^{(i)}$ $(i \in I)$, the process is applied to 'null datasets' drawn from the null hypothesis, in addition to the real dataset. In this manner, we learn the performance of the discovery process when there is nothing to discover, which is the analogue of a null distribution. We also escape the post hoc fallacy because we avoid the retroactive calibration of just the feature $T^{(i_o)}$ that the data analyst considers as discovered. In essence, we calibrate the family-wise Type I error rate for the whole family of discoverable features $T^{(i)}$ $(i \in I)$, even though we may be unable to completely enumerate this family. If data analysts find structure of any kind in the 'null plots', they will tell, and we can (i) tally the occurrences of spurious discoveries/rejections, and more specifically we can (ii) learn the most frequent types of features $T^{(i)}$ that get spuriously discovered.

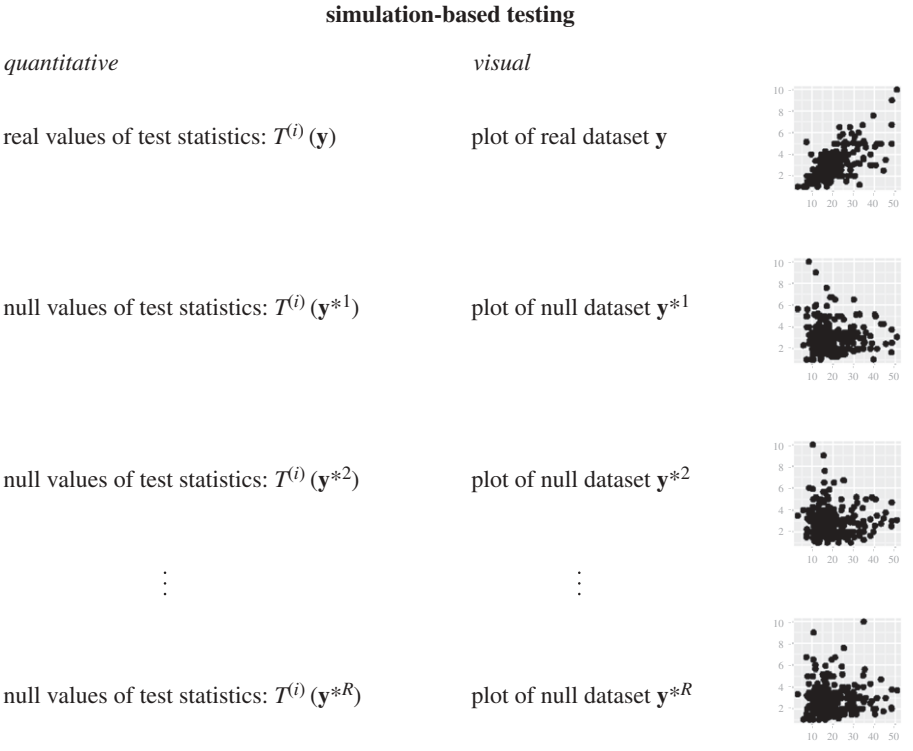

Figure 2. Depiction of simulation-based testing.

3. Reference distributions, null datasets and null plots

In visual inference, the analogue of a collection of test statistics is a plot of the data. Accordingly, we introduce the concept of a 'null distribution of plots' as the analogue of the null distribution of test statistics. This refers conceptually to the infinite collection of plots of 'null datasets' sampled from H_0. In practice, we sample a finite number (R, say) of null datasets $\mathbf{y}^{*1}, \mathbf{y}^{*2}, \ldots, \mathbf{y}^{*R}$ and generate a gallery of R 'null plots' (see figure 2 for a schematic depiction).

The question that arises next is what 'sampling from H_0' means because the null hypothesis H_0 rarely consists of a single distribution. Instead, H_0 is usually 'composite', i.e. a collection of distributions indexed by the so-called 'nuisance parameters'. Fortunately, the problem of reducing composite null hypotheses to single 'reference distributions' has found several solutions in statistical theory, and three principles that can be applied are (i) conditioning, (ii) plug-in, and (iii) posterior inference. Expressed as sampling schemes, they are:

(i) conditional sampling given a statistic that is minimal sufficient under H_0,
(ii) parametric bootstrap sampling whereby nuisance parameters are estimated under H_0, and
(iii) Bayesian posterior predictive sampling.

Of these approaches, the first is the least general but when it applies it yields an exact theory. It does apply to the examples used in this paper: null hypotheses of independence in EDA, and of normal linear models in MD. The resulting reference distributions are, respectively:

EDA: permutation distributions, whereby the observed values are subjected to random permutations within variables or blocks of variables; and
MD: 'residual rotation distributions', whereby random vectors are sampled in residual space with length to match the observed residual vector.

Of the two, the former are well known from the theory of permutation tests, but the latter are lesser known and were apparently explicitly introduced only recently in a theory of 'rotation tests' by Langsrud (2005). When H_0 consists of a more complex model where reduction with a minimal sufficient statistic is unavailable, parametric bootstrap sampling or posterior predictive sampling will generally be available. More details on these topics can be found in the electronic supplementary material. We next discuss two protocols for the inferential use of null plots based on null datasets drawn from reference distributions according to any of the above principles.

4. Protocol 1: 'the Rorschach'

We call the first protocol 'the Rorschach', after the test that has subjects interpret inkblots, because the purpose is to measure a data analyst's tendency to over-interpret plots in which there is no or only spurious structure. The measure is the family-wise Type I error rate of discovery, and the method is to expose the 'discovery black box', meaning the data analyst, to a number of null plots and tabulate the proportion of discoveries which are by construction spurious. It yields results that are specific to the particular data analyst and context of data analysis. Different data analysts would be expected to have different rates of discovery, even in the same data analysis situation. The protocol will bring a level of objectivity to the subjective and cultural factors that influence individual performance.

The Rorschach lends itself to cognitive experimentation. While reminiscent of the controversial Rorschach inkblot test, the goal would not be to probe individual analysts' subconscious, but to learn about factors that affect their tendency to see structure when in fact there is none. This protocol estimates the effective family-wise Type I error rate but does not control it at a desired level.

Producing a rigorous protocol requires a division of labour between a protocol administrator and the data analyst, whereby the administrator (i) generates the null plots to which the data analyst is exposed and (ii) decides what contextual prior information the data analyst is permitted to have. In particular, the data analyst should be left in uncertainty as to whether or not the plot of the real data will appear among the null plots; otherwise, knowing that all plots are null plots, the data analyst's mind would be biased and prone to complacency. Neither the administrator nor the data analyst should have seen the plot of the real data so as not to bias the process by leaking information that can only be gleaned from the data. To ensure protective ignorance of all parties, the administrator might programme the series of null plots in such a way that the plot of the real data is inserted with known probability in a random location. In this manner, the

administrator would not know whether or not the data analyst encountered the real data, while the data analyst would be kept alert because of the possibility of encountering the real data. With careful handling, the data analyst can in principle self-administer the protocol and resort to a separation of roles with an externally recruited data analyst only in the case of inadvertent exposure to the plot of the real data.

While the details of the protocol may seem excessive at first, it should be kept in mind that the rigour of today's clinical trials may seem excessive to the untrained mind as well, and yet in clinical trials this rigour is accepted and heavily guarded. Data analysts in rigorous clinical trials may actually be best equipped to work with the proposed protocol because they already adhere to strict protocols in other contexts. Teaching environments may also be entirely natural for the proposed protocol. Teachers of statistics can put themselves in the role of the administrator, while the students act as data analysts. Such teaching practice of the protocol would be likely to efficiently develop the students' understanding of the nature of structure and of randomness.

In the practice of data analysis, a toned-down version of the protocol may be used as a self-teaching tool to help data analysts gain a sense for spurious structure in datasets of a given size in a given context. The goal of the training is to allow data analysts to informally improve their family-wise error rate and develop an awareness of the features they are most likely to spuriously detect. The training is of course biased by the analysts' knowledge that they are looking exclusively at null plots. In practice, however, the need for looking at some null plots is often felt only after having looked at the plot of the real data and having found merely weak structure. Even in this event, the practice of looking at null plots is useful for gauging one's senses, though not valid in an inferential sense. Implementing this protocol would effectively mean inserting an initial layer into a data analysis—before the plot of the real data is revealed a series of null plots is shown.

5. Protocol 2: 'the lineup'

We call the second protocol 'the lineup', after the 'police lineup' of criminal investigations ('identity parade' in British English), because it asks the witness to identify the plot of the real data from among a set of decoys, the null plots, under the veil of ignorance. The result is an inferentially valid p-value. The protocol consists of generating, say, 19 null plots, inserting the plot of the real data in a random location among the null plots and asking the human viewer to single out one of the 20 plots as most different from the others. If the viewer chooses the plot of the real data, then the discovery can be assigned a p-value of 0.05 (=1/20)—under the assumption that the real data also form a draw from the null hypothesis there is a one in 20 chance that the plot of the real data will be singled out. Obviously, a larger number of null plots could yield a smaller p-value, but there are limits to how many plots a human viewer is willing and able to sift through. This protocol has some interesting characteristics.

(i) It can be carried out without having the viewer identify a distinguishing feature. The viewer may simply be asked to find 'the most special picture' among the 20, and may respond by selecting one plot and saying 'this

one feels different but I cannot put my finger on why this is so'. This is a possibility in principle, but usually the viewer will be eager to justify his or her selection by identifying a feature with regard to which the selected plot stands out compared to the rest.

(ii) This protocol can be self-administered by the data analyst once, if he or she writes code that inserts the plot of the real data among the 19 null plots randomly in such a way that its location is not known to the data analyst. A second round of self-administration of the protocol by the data analyst will not be inferentially valid because the analyst will not only have seen the plot of the real data but in all likelihood have (inadvertently) memorized some of its idiosyncrasies, which will make it stand out to the analyst even if the real data form a sample from the null hypothesis.

(iii) Some variations of the protocol are possible whereby investigators are asked to select not one but two or more 'most special' plots or rank them completely or partially, with p-values obtained from methods appropriate for ranked and partially ranked data.

(iv) This protocol can be repeated with multiple independently recruited viewers who have not seen the plot of the real data previously, and the p-value can thereby be sharpened by tabulating how many independent investigators picked the plot of the real data from among 19 null plots. If K investigators are employed and k ($k \leq K$) selected the plot of the real data, the combined p-value is obtained as the tail probability $P(X \leq k)$ of a binomial distribution $B(K, p = 1/20)$. It can hence be as small as 0.05^K if all investigators picked the plot of the real data ($k = K$).

The idea of the lineup protocol is alluded to by §7 of Davies (2008) to illustrate his idea of models as approximations. He proposes the following principle: 'P approximates x_n if data generated under P look like x_n'. Davies illustrates with a univariate example where a boxplot of the real data is indistinguishable from 19 boxplots of $\Gamma(16, 1.2)$ data but stands out when mingled with boxplots of $\Gamma(16, 1.4)$ data. The ingredient that is missing in Davies (2008) is the general recommendation that nuisance parameters of the model be dealt with in one of several possible ways (see electronic supplementary material) and that a protocol be applied to grant inferential validity.

6. Examples

This section is structured so that readers can test lineup witness skills using the examples. Following all of the lineups, readers will find solutions and explanations. We recommend reading through this section linearly. Several of the datasets used in the examples may be familiar, and if so we suggest that the familiarity is a point of interest because readers who know the data may prove to themselves the disruptive effect of familiarity in light of the protocol.

The first two examples are of plots designed for EDA: scatterplots and histograms. For a scatterplot, the most common null hypothesis is that the two variables are independent, and thus null datasets can be produced by permuting the values of one variable against the other. Histograms are more difficult. The simplest null hypothesis is that the data are a sample from a normal

distribution, and null datasets can be simulated. The viewer's explanation of the structure will be essential, though, because often the plot of the real data will be so obviously different from those of the simulated sets. The next two examples involve time series of stock returns, to examine whether temporal dependence exists at all; here again permutations can be used to produce reference sets. MD is examined in the fourth example, where rotation distributions are used to provide reference sets for residuals from a model fit. The fifth example examines class structure in a large p, small n problem—the question being just how much separation between clusters is due to sparsity. This might be a good candidate for the Rorschach protocol, to help researchers adjust their expectations in this area. The last example studies MD in longitudinal data, where a non-parametric model is fitted to multiple classes. Permutation of the class variable is used to provide reference sets.

(i) *Places Rated data*: This example comes from Boyer & Savageau (1984) where cities across the USA were rated in 1984 according to many features. The data are also of interest because of two peculiarities: they consist of aggregate numbers for US metropolitan areas, and they form a census, not a random sample. In spite of these non-statistical features, it is legitimate to ask whether the variables in this dataset are associated, and it is intuitive to use random pairing of the variable values as a yardstick for the absence of association. The variables we consider are 'Climate-Terrain' and 'Housing'. Low values on Climate-Terrain imply uncomfortable temperatures, either hot or cold, and high values mean more moderate temperatures. High values of Housing indicate a higher cost of owning a single family residence. The obvious expectation is that more comfortable climates call for higher average housing costs. The null hypothesis for this example is

H_0: Housing is *independent* of Climate-Terrain.

The decoy plots are generated by permuting the values of the variable Housing, thus breaking any dependence between the two variables while retaining the marginal distributions of each. Figure 3 shows the lineup. The reader's task is to pick out the plot of the real data.

(a) Is any plot different from the others?
(b) Readers should explicitly note why they picked a specific plot.

(ii) *Tips Data*: This dataset was originally analysed in Bryant & Smith (1995). Tips were collected for 244 dining parties. Figure 4 shows a histogram of the tips using a very small bin size corresponding to 10 cent widths. The null plots were generated by simulating samples from a normal distribution having the same range as the real data. Which histogram is different? Explain in detail how it differs from the others.

(iii) *HSBC* ('*The Hongkong and Shanghai Banking Corporation*') *daily stock returns*: two panels, the first showing the 2005 data only, the second the more extensive 1998–2005 data (figure 5). In each panel, select which plot is the most different and explain why.

(iv) *Boston Housing Data*: This dataset contains measurements on housing prices for the Boston area in the 1970s. It was discussed in Harrison & Rubinfeld (1978), used in Belsley *et al.* (1980) and is available at Vlachos (2005).

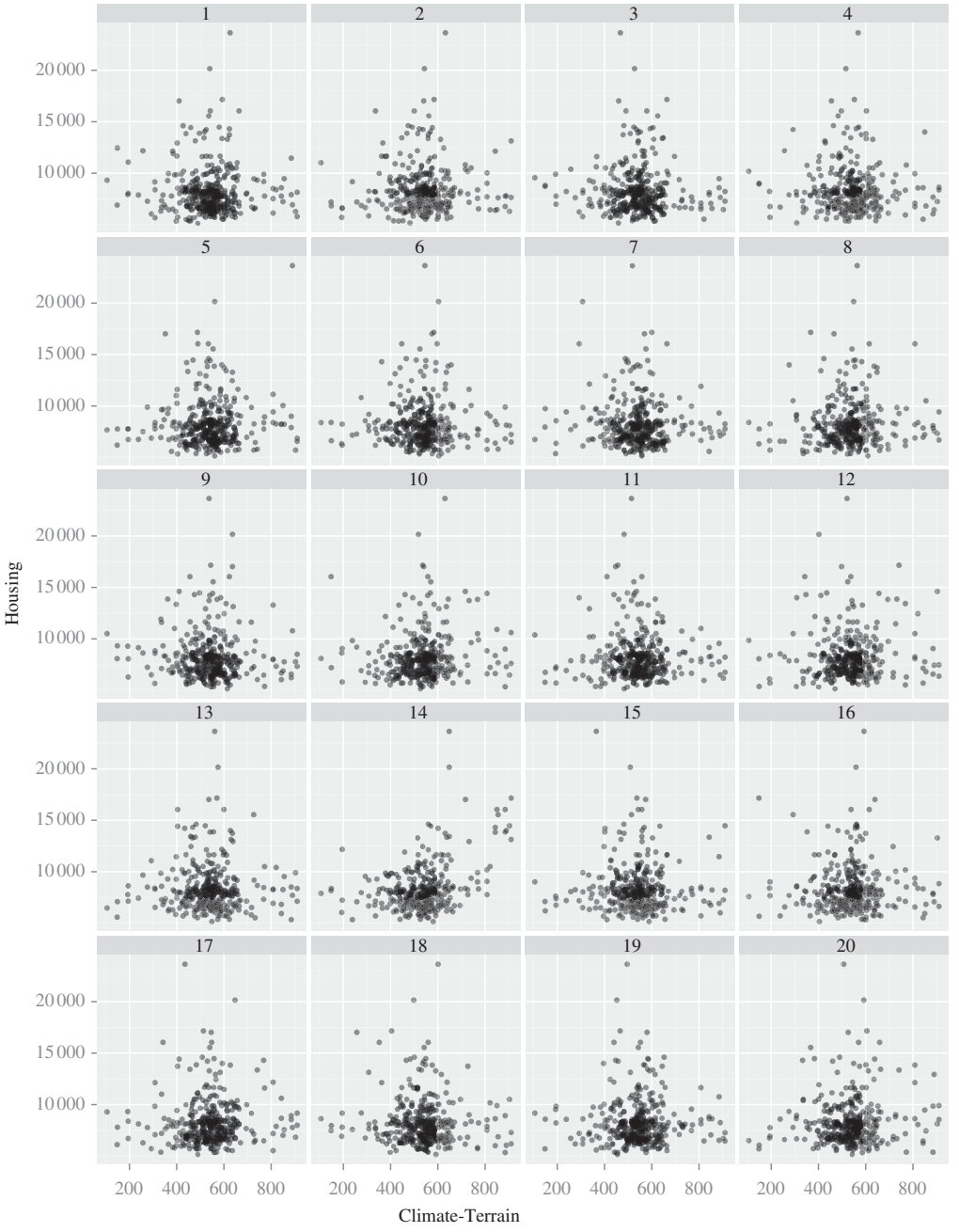

Figure 3. Permutation lineup: Places Rated data, Housing versus Climate-Terrain. The plot of the real data is embedded among permutation null plots. Which plot shows the real data? What features make it distinctive? Does knowledge of the meaning of the variables influence the choice?

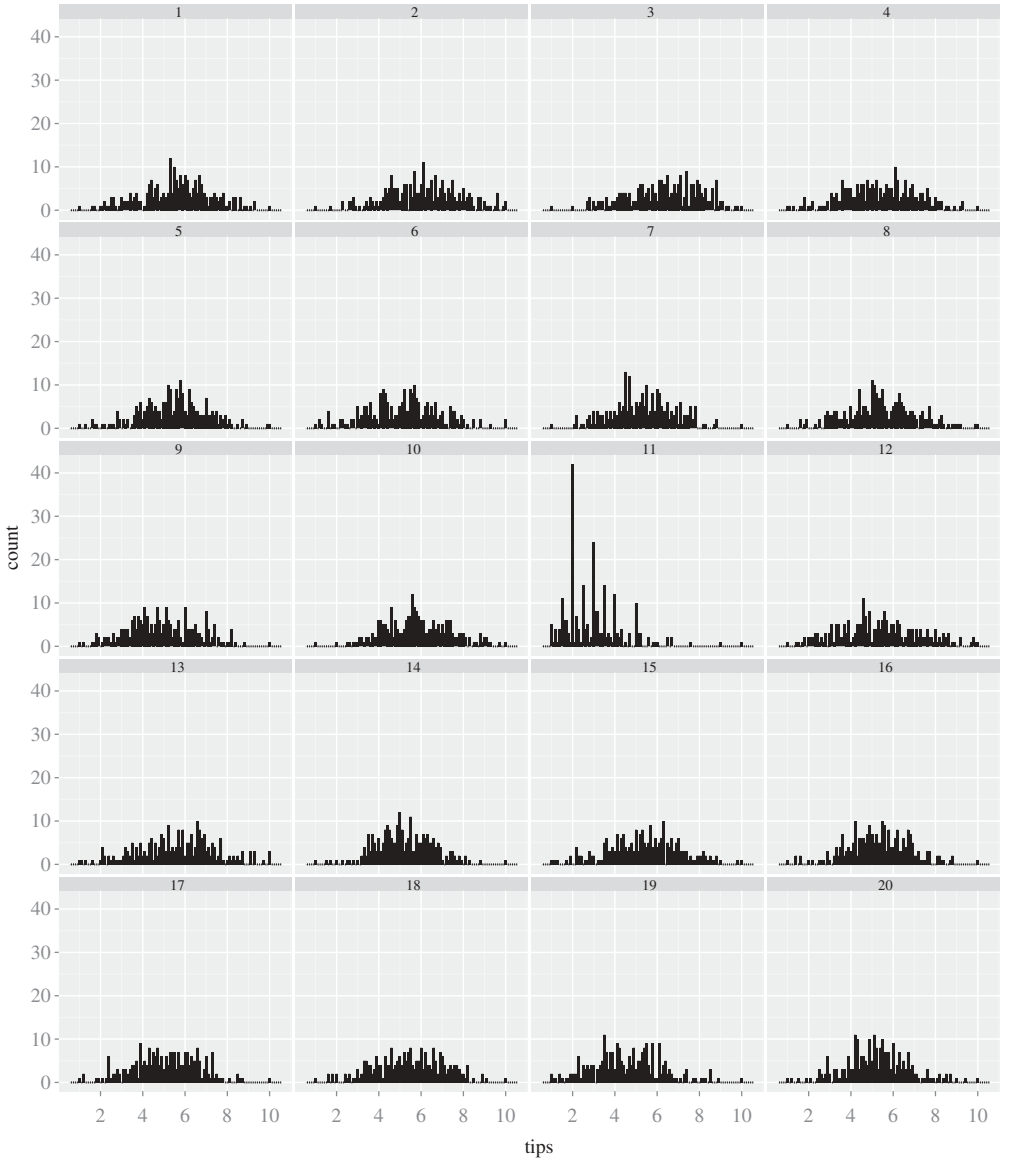

Figure 4. Simulation lineup: Tips Data, histograms of tip. Which histogram is most different? Explain what makes it different from the others.

Figure 6 shows residuals plotted against the order in the data. Structure in the real plot would indicate the presence of lurking variables. The plot of the real data is embedded among decoys, which were produced by simulating from the residual rotation distribution, as discussed in §3 and the electronic supplementary material. Which is the plot of the real data? Why? What may be 'lurking'?

(v) *Leukaemia data*: This dataset originated from a study of gene expression in two types of acute leukaemia (Golub *et al.* 1999), acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). The dataset consists of

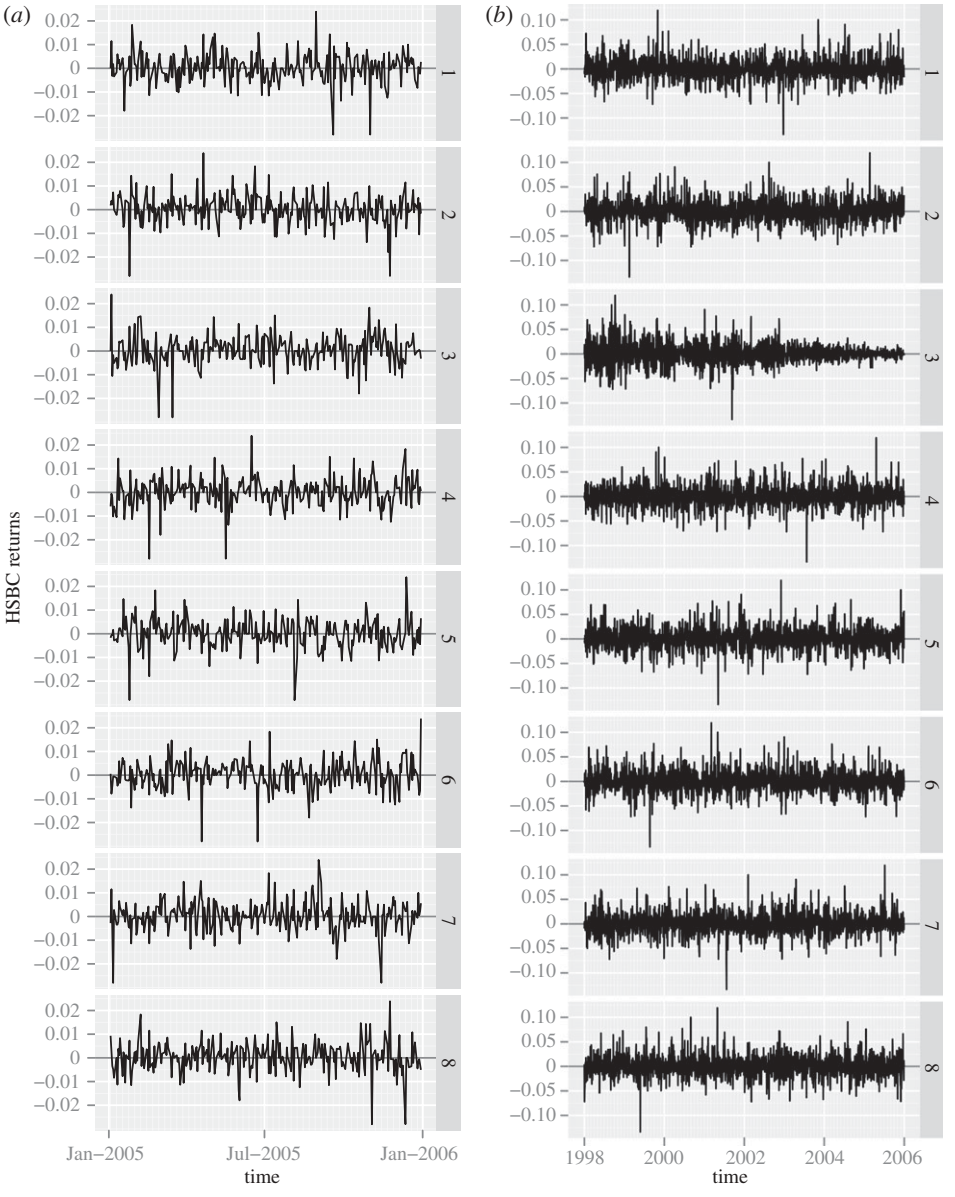

Figure 5. Permutation lineup: (*a*) HSBC daily returns 2005, time series of 259 trading days and (*b*) HSBC daily returns 1998–2005, time series of 2087 trading days. The plot of the real data is embedded among seven permutation null plots. Which plot is different from the others and why?

25 cases of AML and 47 cases of ALL (38 cases of B-cell ALL and 9 cases of T-cell ALL), giving 72 cases. After pre-processing, there are 3571 human gene expression variables.

To explore class structure, one may seek a few interesting low-dimensional projections that reveal class separations using a projection pursuit method, such as LDA (linear discriminant analysis) and PDA (penalized discriminant analysis) indices (Lee 2003; Lee *et al.* 2005).

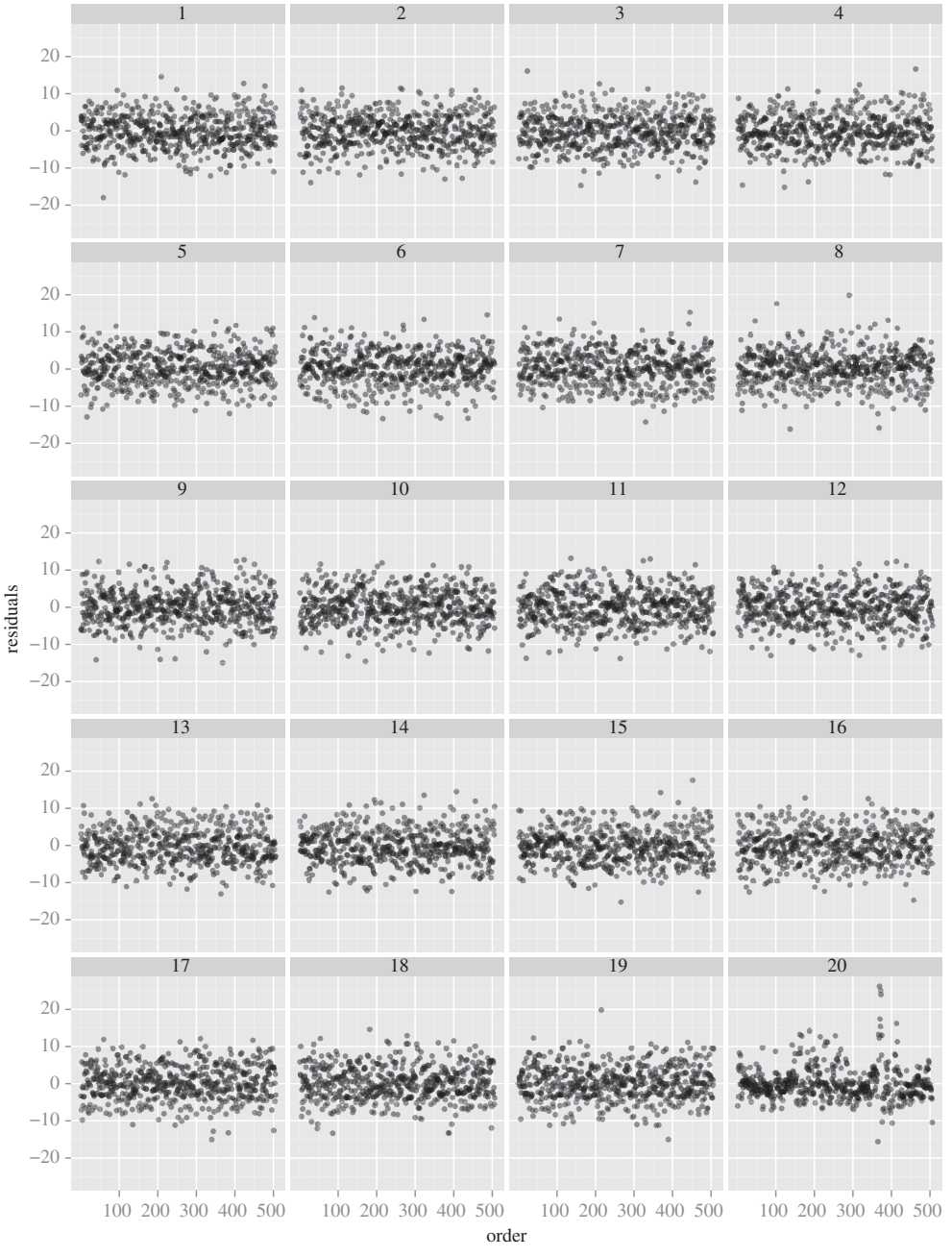

Figure 6. Residuals of the Boston Housing Data plotted against order in the data. What does the structure in the real plot indicate?

In supervised classification problems, especially when the size of the sample is small in relation to the number of variables, the PDA index is suitable. This index is designed to overcome the variance estimation problems that arise in the LDA index. When the sample size is small and the number of variables is large, the LDA index produces unreliable results, which express themselves as

Figure 7. These 20 plots show two-dimensional projection pursuit projections of the best separation of three classes in gene expression data, having 72 samples and 3571 variables. Which plot is most different from the others? Why? Group: pink, AML; green, B-ALL; blue, T-ALL.

data piling. To avoid this problem, the PDA index uses a penalization term regulated by a parameter λ. In this example we examine the two-dimensional optimal projection using the PDA index with $\lambda = 0.8$. Figure 7 shows the plot of the real data mingling among nineteen plots of data using permuted classes. Which plot shows the real data? Why?

(vi) *Wages data*: In this example, we study the relationship between wages and experience in the workforce by race for a cohort of male high-school dropouts. The data are taken from Singer & Willett (2003) and contain longitudinal measurements of wages (adjusted to inflation), years of experience in the workforce and several covariates, including the subject's race. A non-parametric approach for exploring the effect of race is to fit a smoother separately to each racial subgroup in the data. If there appears to be a difference between the curves, how can we assess the magnitude and significance of the difference? The null scenario is that there is no difference between the races. To generate null sets, the race label for each subject is permuted. The number of longitudinal measurements for each subject varies from one to 13. Each subject has an id and a race label. These labels are re-assigned randomly. There will be the same number of subjects in each racial group, but the number of individual measurements will differ. Nineteen alternative datasets are created. For each dataset, a loess smooth (Cleveland *et al.* 1992) is calculated on each subgroup, and these curves are plotted using different line types on the same graph, producing 20 plots, including the original (figure 8). The plots also have the full dataset shown as points underlying the curves, with the reasoning being that it is helpful to digest the difference between curves on the canvas of the variability in the data. Here is the question for this example.

> These 20 plots show smoothed fits of log(wages) to years of experience in the workforce for three demographic subgroups. One uses real data, and the other 19 are produced from null datasets, generated under an assumption that there was no difference between the subgroups. Which plot is the most different from the others, paying particular attention to differences in areas where there are more data?

This next part discusses the lineups, revealing the location of the real data, and explaining what we would expect the viewers to see.

(i) *Places Rated data* (*figure 3*): There is a slight positive association, but it is not strong. Also, there are two clusters on the right, coastal California and the Pacific Northwest. The so-called Climate-Terrain index is really just a measurement of how extreme versus how moderate the temperatures are, and there is nothing in the index that measures differences in cloud cover and precipitation.

Solution: The real data are shown in plot 14.

(ii) *Tips Data* (*figure 4*): Three features are present in the real data: skewness, multi-modality (with peaks at dollar and half-dollar amounts) and three outliers. Of these the first two are most obviously different from the null sets, and we would expect these to be reported by the viewer.

Solution: The real data are shown in plot 11.

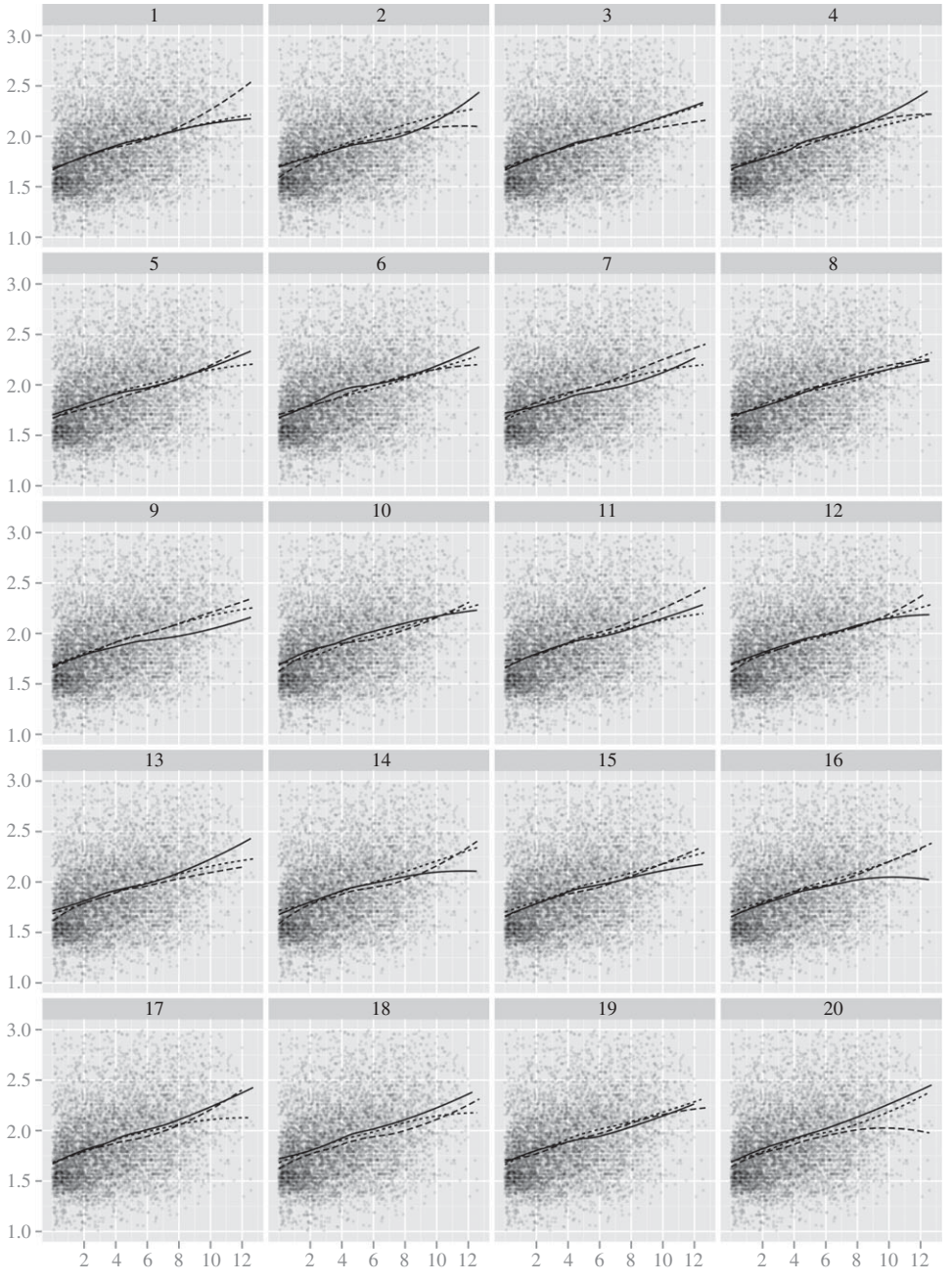

Figure 8. These 20 plots show loess smooths on measurements of log(wages) and workforce experience (years) for three subgroups in a sample of high-school dropouts. The soft grey points are the actual data used for the smoothing, before dividing into subgroups. One of the plots shows the actual data, and the remainder have had the group labels permuted before the smoothing. Which plot is the most different from the others, with particular attention paid to more differences between curves where there are more data? What is the feature in the plot that sets it apart from the others? Race: solid line, Black; dotted line, White; dashed line, Hispanic.

(iii) *HSBC* (‘*The Hongkong and Shanghai Banking Corporation*’) *daily stock returns* (*figure 5*): For the short 2005 series, the reader should have had difficulty discerning the real data. This is a year of low and stable volatility. Because volatility changes are the major time dependencies in this type of data, the real and permuted stock returns are quite indistinguishable. The long series for the years 1998–2005, however, features quite drastic changes in volatility, such as two volatility bursts, one in 1998 due to the Russian bond default and the LTCM collapse, the other in 2001 due to the 9/11 event. Thereafter, volatility peters out and stabilizes at a low level.

Solution: In both lineups the real data are shown in plot 3.

(iv) *Boston Housing Data* (*figure 6*): The real residuals show pronounced structure compared to the null residuals. For most parts, the real residuals move more tightly than the simulated ones, except for two major excursions in the high 300s. We may surmise that the order of the census tracts in the data is spatially coherent in that tracts nearby in the data order tend to be nearby geographically. If correct, the deficiency established in this type of plot points to spatially coherent lurking variables that are not in the data.

Solution: The real data are shown in plot 20.

(v) *Leukaemia data* (*figure 7*): Most plots with permuted data show separable class structure. However, the plot of the real data shows the most separable class structure, which suggests that there appears to be stronger distinction between the classes than would occur by chance with so many variables and so few cases.

Solution: Plot 1 shows real data.

(vi) *Wages data* (*figure 8*): In the plot of the real data, the three curves differ in two ways: (A) the solid line flattens out at about six years of experience and never increases to be equal to the other two lines and (B) the three curves are all different in the later years, around 10 years. Now which of these features is present in the null data plots? Although a mid-experience flattening is present in several other plots (e.g. 18, 7), feature A is more extreme in the plot of the real data than in any of the other plots. Feature B is present in almost every other plot and is more pronounced in a plot of the simulated data (e.g. 16, 1, 20).

Solution: Plot 9 shows real data.

In this example, the artefact in the null data is more pronounced than the feature of interest in the real data. In fact, this pattern, where the smoothed line dips around 8 years of experience, is so pronounced we suspect that many readers will have selected this as the most different plot. A seasoned data analyst might not be fooled, though, because it is well known that there can be edge effects with smoothers particularly when there are fewer points in the extremes. Without the careful wording of the question to point out that we are looking for differences, some readers may have been tempted to select plot 8, where the curves are almost all identical. The presence of multiple features can be both beneficial

or detrimental to the process of assessing significance and exploring data. If a pronounced feature is present in the null data, readers may be distracted from the feature of interest. However, in some cases, the readers may report a different feature in the real data than the analyst had noticed, thus leading to discovery of new information about the data.

Aside: An alternative approach to graphical inference, to assess the significance of the difference between smoothers, is to produce 'null bands' based on permutations of the race labels (Atkinson 1981; Bowman & Azzalini 1997; Buja & Rolke 2009). The race labels of the subjects are permuted many times and a smoother is fitted to each of the resulting samples. These different smooths are combined to produce an envelope for the smoother of the actual data, produced under the assumption that there is no difference between the races. If we did this for these data, we would find that for Whites and Hispanics the fit stays within the null bands, but for Blacks the actual fit dips low, out of the limits between experience values 6–10, suggesting that there is something different about the wage experience for this group. The benefit of the graphical inference approach over the null bands approach is that the entire curve from each permutation is examined, so that curve to curve variation can be seen. Davison & Hinkley (1997) and Pardoe (2001) have an intermediate solution between the null band and graphical inference, in which the curves are overlaid on the one plot.

7. Conclusions and caveats and future investigations

This paper proposes two protocols to formalize the process of visual discovery, the Rorschach and lineup. Each helps to quantify the strength of signal versus noise, similar to numerical hypothesis testing. The Rorschach protocol provides a prior-to-analysis visual calibration training for the data analyst, and the lineup provides an inferentially valid test of whether what we see in a plot is really there. Here are some of the issues that need consideration before more general implementation.

(i) The wording of the instructions to viewers can affect their plot choices. (This is a familiar issue in survey design (Dawes 2000).) Additional information might be included to help viewers make wise choices. For example, the introduction to the wages data example guided viewers towards the higher data density area in the middle of the x range. The purpose was to pre-empt the naive choice of selecting the plot exhibiting the strongest end-effects of the smoothed lines, which a trained statistician would probably avoid. Should this methodological effect factor into the family-wise Type I error rate? In the other direction, how much personal baggage is brought to the problem by informing the viewer about the nature of the demographic subsets being based on race? Would it be better to mask the race variable, and call them groups 1, 2, 3?

(ii) Ancillary information related to the viewer's response will be useful. For example, the time that an individual takes to arrive at a choice might be included in the analysis of the results. When the choice comes quickly, it

might suggest that the pattern is strong (tips example, figure 4), but when it takes longer it might suggest that the signal is weak (HSBC 2005 data, figure 5), or the viewer is not as confident in their choice. It may be useful to ask the viewer to rate their confidence in their answer. Limiting the time it takes to answer may produce less reliable results.

(iii) Readers will respond differently to the same plots, depending on training and even state of mind (Whitson & Galinsky 2008). There are, however, common traits that we should be aware of and expect to see from all viewers; for example, our visual perception responds strongly to gaps, colour inconsistencies and effects in the edges of plot regions, but may not pick up smaller deviations in large patterns. Subjectivity of results in visual test procedures is unavoidable. The Rorschach protocol may help to determine the baseline for each individual.

(iv) Use of these protocols might have positive effects on improving statistical graphics used in the community. Because analysts are forced to think about the null hypothesis associated with a plot, it may hone their abilities to choose appropriate graphics for their tasks. With additional work, the use of good principles in constructing plots might also be improved: pre-attentive plot elements for the data, attentive plot elements for grids and axes to allow look up only when needed.

(v) In Tukey's approach to EDA, analysis was sometimes done in an iterative manner: strong patterns are removed and the residuals are re-tested to reveal fine scale patterns. To use this approach, care might be needed to avoid bias of the secondary tests by exposure to the primary-stage plots.

(vi) One way to reach a substitute for a jury could be the use of a Web service such as Amazon's (2008) Mechanical Turk. Sets of plots based on the lineup protocol will be evaluated by the so-called human 'turkers', thus enabling us to gauge family-wise Type I error rates for each data situation. It also allows us to easily capture time until completion of the task, an explanation for the individual's pick in a lineup, together with a (subjective) confidence rating. While 'turkers' do not have the make-up of a representative population sample, we can collect some demographic information with the results and try to correct for that. The Mechanical Turk has also been used to collect data for the Fleshmap project by Viègas & Wattenberg (2008).

These points suggest directions for future research. We hope the paper provides something of a road-map to the incorporation of graphical discoveries as an integral part of statistical data analysis, consequently enhancing our ability, as statisticians, to handle increasingly difficult data problems. As an example, the plots in this paper were made using the R package, `ggplot2` (Wickham 2008), using new functions which semi-automate the lineup plot format.

This work has been partly supported by National Science Foundation grant DMS0706949.

References

Amazon. 2008 Mechanical Turk. See http://aws.amazon.com/mturk/.

Atkinson, A. 1981 Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13–20. (doi:10.1093/biomet/68.1.13)

Belsley, D. A., Kuh, E. & Welsch, R. E. 1980 *Regression diagnostic: identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bowman, A. W. & Azzalini, A. 1997 *Applied smoothing techniques for data analysis*. Oxford, UK: Oxford University Press.

Boyer, R. & Savageau, D. 1984 *Places rated almanac*. Chicago, IL: Rand McNally.

Brillinger, D. R. 2008 The 2005 Neyman lecture: dynamic indeterminism in science. *Stat. Sci.* **23**, 48–64. (doi:10.1214/07-STS246)

Bryant, P. G. & Smith, M. A. 1995 *Practical data analysis: case studies in business statistics*. Homewood, IL: Richard D. Irwin Publishing.

Buja, A. & Rolke, W. 2009 Calibration of simultaneity: (re-)sampling methods for simultaneous inference with applications to function estimation and functional data, pp. 277–308. See http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-sim.pdf.

Buja, A., Asimov, D., Hurley, C. & McDonald, J. A. 1988 Elements of a viewing pipeline for data analysis. In *Dynamic graphics for statistics* (eds W. S. Cleveland & M. E. McGill), pp. 277–308. Monterey, CA: Wadsworth.

Buja, A., Cook, D. & Swayne, D. 1996 Interactive high-dimensional data visualization. *J. Comput. Graph. Stat.* **5**, 78–99. (doi:10.2307/1390754)

Buja, A., Cook, D. & Swayne, D. 1999 Inference for data visualization. Talk given at Joint Statistical Meetings. See http://www-stat.wharton.upenn.edu/~buja/PAPERS/visual-inference.pdf.

Card, S. K., Mackinlay, J. D. & Schneiderman, B. 1999 *Readings in information visualization*. San Francisco, CA: Morgan Kaufmann Publishers.

Chatfield, C. 1995 *Problem solving: a statistician's guide*. London, UK: Chapman and Hall/CRC Press.

Chen, C.-H., Härdle, W. & Unwin, A. (eds) 2007 *Handbook of data visualization*. Berlin, Germany: Springer.

Cleveland, W. S. 1993 *Visualizing data*. Summit, NJ: Hobart Press.

Cleveland, W. S., Grosse, E. & Shyu, W. M. 1992 Local regression models. In *Statistical models in S* (eds J. M. Chambers & T. J. Hastie), pp. 309–376. New York, NY: Chapman and Hall.

Daniel, C. 1976 *Applications of statistics to industrial experimentation*. Hoboken, NJ: Wiley-Interscience.

Davies, P. L. 2008 Approximating data (with discussion). *J. Korean Stat. Soc.* **37**, 191–240. (doi:10.1016/j.jkss.2008.03.004)

Davison, A. C. & Hinkley, D. V. 1997 *Bootstrap methods and their applications*. Cambridge, UK: Cambridge University Press.

Dawes, J. 2000 The impact of question wording reversal on probabilistic estimates of defection/loyalty for a subscription product. *Market. Bull.* **11**, 1–7.

Diaconis, P. 1983 Theories of data analysis: from magical thinking through classical statistics. In *Exploring data tables, trends and shapes* (eds D. Hoaglin, F. Mosteller & J. Tukey), pp. 1–36. New York, NY: Wiley.

Gelman, A. 2004 Exploratory data analysis for complex models. *J. Comput. Graph. Stat.* **13**, 755–779. (doi:10.1198/106186004X11435)

Golub, T. R. *et al.* 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **268**, 531–537. (doi:10.1126/science.286.5439.531)

Harrison, D. & Rubinfeld, D. L. 1978 Hedonic prices and the demand for clean air. *J. Environ. Econ. Manag.* **5**, 81–102. (doi:10.1016/0095-0696(78)90006-2)

Langsrud, Ø. 2005 Rotation tests. *Stat. Comput.* **15**, 53–60. (doi:10.1007/s11222-005-4789-5)

Lee, E. 2003 Projection pursuit for exploratory supervised classification. PhD thesis, Iowa State University.

Lee, E., Cook, D., Klinke, S. & Lumley, T. 2005 Projection pursuit for exploratory supervised classification. *J. Comput. Graph. Stat.* **14**, 831–846. (doi:10.1198/106186005X77702)

Neyman, J., Scott, E. L. & Shane, C. D. 1953 On the spatial distribution of galaxies: a specific model. *J. Astrophys.* **117**, 92–133. (doi:10.1086/145671)

Pardoe, I. 2001 A Bayesian sampling approach to regression model checking. *J. Comput. Graph. Stat.* **10**, 617–627. (doi:10.1198/106186001317243359)

Scott, E. L., Shane, C. D. & Swanson, M. D. 1954 Comparison of the synthetic and actual distribution of galaxies on a photographic plate. *Astrophys. J.* **119**, 91–112. (doi:10.1086/145799)

Singer, J. D. & Willett, J. B. 2003 *Applied longitudinal data analysis*. Oxford, UK: Oxford University Press.

Swayne, D. F., Cook, D. & Buja, A. 1992 XGobi: interactive dynamic graphics in the X window system with a link to S. In *Proc. Section on Statistical Graphics at the Joint Statistical Meetings, Atlanta, GA, 18–22 August 1991*, pp. 1–8.

Tufte, E. 1983 *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tukey, J. W. 1965 The technical tools of statistics. *Am. Stat.* **19**, 23–28. (doi:10.2307/2682374)

Tukey, J. W. & Tukey, P. A. 1985 Computer graphics and exploratory data analysis: an introduction. In *Proc. 6th Annual Conf. and Exposition of the National Computer Graphics Association, Dallas, TX, 14–18 April 1985*, pp. 773–785.

Viègas, F. & Wattenberg, M. 2008 Fleshmap. See http://fernandaviegas.com/fleshmap.html.

Vlachos, P. 2005 Statlib: data, software and news from the statistics community. See http://lib.stat.cmu.edu/.

Whitson, J. A. & Galinsky, A. D. 2008 Lacking control increases illusory pattern perception. *Science* **322**, 115–117. (doi:10.1126/science.1159845)

Wickham, H. 2008 ggplot2: an implementation of the grammar of graphics in R. R package version 0.8.1. See http://had.co.nz/ggplot2/.

Wilkinson, L. 1999 *The grammar of graphics*. New York, NY: Springer.

Wilkinson, L., Anand, A. & Grossman, R. 2005 Graph-theoretic scagnostics. In *Proc. 2005 IEEE Symp. on Information Visualization* (*INFOVIS*), pp. 157–164.