Modeling Customer Survey Data

Linda A. Clark, Bell Labs, Murray Hill, NJ
William S. Cleveland, Bell Labs, Murray Hill, NJ
Lorraine Denby, Bell Labs, Murray Hill, NJ
Chuanhai Liu, Bell Labs, Murray Hill, NJ

ABSTRACT In customer value analysis (CVA), a company conducts sample surveys of its customers and of its competitors' customers to determine the relative performance of the company on many attributes ranging from product quality and technology to pricing and sales support. The data discussed in this paper are from a quarterly survey run at Lucent Technologies.

We have built a Bayesian model for the data that is partly hierarchical and has a time series component. By "model" we mean the full specification of information that allows the computation of posterior distributions of the data — sharp specifications such as independent errors with normal distributions and diffuse specifications such as probability distributions on parameters arising from sharp specifications. The model includes the following: (1) survey respondent effects are modeled by random location and scale effects, a t-distribution for the location and a Weibull distribution for the scale; (2) company effects for each attribute through time are modeled by integrated sum-difference processes; (3) error effects are modeled by a normal distribution whose variance depends on the attribute; in the model, the errors are multiplied by the respondent scale effects.

The model is the first full description of CVA data; it provides both a characterization of the performance of the specific companies in the survey as well as a mechanism for studying some of the basic notions of CVA theory.

Building the model and using it to form conclusions about CVA, stimulated work on statistical theory, models, and methods: (1) a *Bayesian theory of data exploration* that provides an overall guide for methods used to explore data for the purpose of making decisions about model specifications; (2) an approach to *modeling random location and scale effects* in the presence of explanatory variables; (3) a reformulation of integrated moving-average processes into *integrated sum-difference models*, which enhances interpretation, model building, and computation of posterior distributions; (4) *post-posterior modeling* to combine certain specific exogenous information — information from sources outside of the data — with the information in a posterior distribution that does not incorporate the exogenous information; (5) *trellis display*, a framework for the display of multivariable data.

1 Introduction

1.1 CVA

In customer value analysis (CVA), a company conducts sample surveys of its customers and of its competitors' customers to determine the relative performance of the company on many *attributes* ranging from product quality and technology to pricing and sales support (Gale, 1994; Nauman and Kordupleski, 1995). The goal is to use CVA to look across the business to prioritize where resources would best be committed to improve performance. The basic premise of CVA is that the market share of a company is directly linked to the company's performance on the polled attributes.

The concept of CVA goes back to the 1970s and is often attributed to Sidney Schoeffler at General Electric. But the practice of CVA by major corporations is relatively new and is still in the stage of gaining acceptance. It represents a broadening of the quality movement which for the last few decades has focused largely on process improvement to meet engineering specifications. CVA adds information about customers' opinions so that in addition to quality as conformance to specifications, there is customer-perceived quality. And CVA introduces the practice of measuring competitors' customers.

1.2 Data

The data discussed in this paper are from a quarterly survey of supplier companies in the information technology industry. Customers are polled and rate companies on attributes on a scale of 1 to 10. The data of the paper cover 9 attributes and there are 3505 pollings. If there were no missing data we would have $3505 \times 9 = 31545$ ratings. But because of missing data there are 25309 ratings. In all, 19 suppliers are rated over a period of 9 quarters. Most respondents are polled just once but a small fraction are polled more than once.

These data are a subset of a much larger data set; 7 of the 9 attributes studied here have sub-attributes, questions that clarify issues of the main attribute. In the interest of space and readability we study only the 9 main attributes here, rather than the full set of 45 attributes. Also, not all of the companies in the survey are included here; for example, Lucent Technologies is not included. Nor are all of the currently available quarters included.

The information in our survey is highly sensitive; individual companies compete, but they also form many partnerships — for example, partnerships to develop and market certain product lines, and industry-wide partnerships to establish standards. For this reason we must regard the company names as confidential. We will rename the companies with three letter words such as "Cat" and "Jet".

1.3 The Two Goals of the Data Analysis

Ambitious surveys of customers such as ours are instituted to characterize the relative performance of companies on the polled attributes; this provides a basis for actions to improve performance from a customer perspective. Thus one goal of our data analysis is to provide a rigorous mechanism for going from the survey data to characterizations of performance. The mechanism involves the combination of information from the data with the basic tenets of customer-value-analysis theory. The tenets amount to premises about the perceptions of customers on a variety of attributes and how these perceptions are structured. Gale (1994) and Nauman and Kordupleski (1995) present the current status of CVA theory that is adopted by most practitioners. A second goal of our analysis is to study CVA theory both to test old tenets and to develop new ones. Thus the outcome of our analysis is (1) characterizations of relative performance and (2) statements about CVA theory. While most of this paper is devoted to model development, estimation, and computation, we briefly illustrate the two-part outcome of our work at the end of the paper.

2 Overviews

We cannot in this one paper describe in full detail our analysis of the survey data. In this section we will summarize (1) the contents of papers on theory, models, and methods that were stimulated by work on the CVA data; (2) the contents of this paper; (3) aspects of the data analysis not conveyed here.

By the word “model” we mean the full specification of information that allows the computation of posterior distributions of the data — sharp specifications such as independent errors with normal distributions and diffuse specifications such as probability distributions on parameters arising from sharp specifications.

By the phrase “exogenous information” we mean information about the subject under study from sources outside of the data. Some might use the term “prior information” but we do not simply to avoid the unfortunate connotation that prior information is formally specified prior to the exploration and study of the data.

2.1 Theory, Models, and Methods

Philosophies of Model Building

We did not simply write down a model for the customer data and then carry out diagnostic checks based on a complete model to see if it appeared to be reasonable. The sources of variability in our data are too varied and complex to allow such an approach to succeed. Instead, we carried out a process of building up a model by a detailed exploration of the structure of the data, starting with a minimum of specifications, and ending with a complete model. At each stage of our model building there were specifications in place that guided the exploration, and at the end there was exploration based on a tentative complete model.

Since data visualization methods are the single most powerful approach to studying the structure of data, visualization played a fundamental role in the model building.

The origins of this philosophy of searching the data for structure as a part of model building go back to the emerging ideas on diagnostic checking of models (e.g., Anscombe and Tukey, 1961; Box and Hunter, 1965; Daniel and Wood, 1971).

But recent evolutions of this philosophy have had significant changes (Cleveland, 1993). In recent decades computing technology has changed so profoundly that it calls for revised fundamentals for approaching data. There has been an enormous increase in the power of visualization tools for exploration. This has led to a significant increase in the potential reliance we can place on visualization as a basis of modeling decisions. With this increase has been a corresponding decrease in the amount of reliance that we need to put on using more formal tests of sharp hypotheses. Rather it means that the myriad decisions that must be made in building a model for data are often made more judiciously and expeditiously via data exploration and less formal methods. There has been an enormous increase in the speed of numerical computations and in the ease and flexibility of specifying and fitting models. Consequently, data exploration can be more than simply a look at the raw data. It can involve exploration based on many tentative structures for the data, some partial and some complete.

A Theory of Data Exploration for Model Building

One result that we hope will excite some is that it is possible to see model building via data exploration quite explicitly as part of the overall Bayesian paradigm.

Cleveland and Liu (1998a) have developed a theory of data exploration that invokes Bayesian principles for its rationale. The theory begins with a number of premises about data analysis in practice, both what is actually done and what is desirable. Many of these premises have been expressed in a number of quite disparate works, both Bayesian and frequentist (Box, 1980; Draper, 1970; Dempster, 1970; Draper, Hodges, Mallows, and Pregibon, 1993; Edwards, Lindman, and Savage, 1963; Gelman, Meng, and Stern, 1996; Good, 1957; Hill, 1986; Hill, 1990; Kass and Raftery, 1995; Rubin, 1984; Savage, 1961). From these premises comes a key concept of the theory: exploration methods allow us to reason about the likelihood of the data given specifications. We combine our judgment of likelihood based on the data exploration with our judgment of likelihood based on exogenous information to form a new likelihood that is the basis for decisions about specifications. The combination is analogous to the use of Bayes theorem when we have exogenous information and likelihood described mathematically. But, for data exploration, in place of the combining by mathematical computation, we use a process of direct, intuitive reasoning. While we lose the precision of mathematically described inductive inferences, we gain an enormous breadth of coverage of competing specifications that is simply not feasible to capture mathematically.

Casting data exploration as a vehicle for judging the likelihood of the data given specifications has important implications for how we carry out exploration. Here are four of many: (1) The theory encourages accompanying exploratory methods with assessments of the variability of the displayed functions of the data, either by simulating the displays with data generated from credible specifications, or by simple mathematical probabilistic calculations based on these specifications. This may seem like a break with data exploration as it is practiced by many, because we invoke notions of probability, but we believe it is only a break with how data exploration is often portrayed. (2) The theory encourages the use of visualization methods to explore the data because this often provides powerful assessments of likelihood. Visualization methods can reveal much information in the data and thus allow us to judge the likelihood of the data given a wide variety of specifications seen to be credible based on exogenous information. (3) The theory discourages the use of test statistics that reduce the vast information in the data to a small amount of information because this provides exceedingly limited assessments of likelihood. For example, the theory explains why a normal probability plot is a far more powerful tool for deciding on a sharp specification of normality than a chi-squared test (even a chi-squared test calibrated by a posterior predictive distribution). (4) The theory encourages the fitting of many alternative models, some to the full data, and others to functions of the data, as a vehicle for assessing likelihood.

Building Models with Random Locations and Scales

Many sets of data consist of cases, each of which has an associated set of measurements of a response variable. The response depends on explanatory variables, and the goal is to describe the dependence, but the cases have an effect as well, which complicates the study of dependence.

For our survey data, the response is the ratings, the cases are the respondents, and the explanatory variables are attribute, company, and time.

It is common to model a case effect by a random location effect. But for many case effects, scales vary as well. For example, it is widely acknowledged that for rater data, respondents have varying locations and scales. But it is uncommon to model such scale effects by random effects, in part because model building, model fitting, and diagnostic checking are complex. Often, the issue is ignored, or effects are estimated as fixed which uses up too many degrees of freedom unless each rater does ratings of many cases (Longford, 1995).

Cleveland, Denby, and Liu (1998) have developed an approach to modeling random location and scale effects in the presence of explanatory variables. An overall framework is posited that can be verified by diagnostic checking. Within this framework, they have developed procedures for exploring the data to get insight into the form of the location, scale, and error distributions as well as the dependence of the response on the explanatory variables. The problem is that the functions of the data used in the exploration convolve the population location and scale distributions with error distributions and with sampling distributions. So their methods, which have an empirical Bayes flavor, amount to deconvolution procedures.

Integrated Sum-Difference Models

We employ an integrated moving-average time series model (Box and Jenkins, 1970) to describe changes through time in company-attribute effects. To do this we developed a class of models — integrated sum-difference models, or ISD(d,q) — which are IMA(d,q) models but are structured differently; the new structure enhances interpretation, model building, and computation of posterior distributions (Cleveland and Liu, 1998b). The new structure models a differenced series by a sum of orthogonal series, each the output of applying, to white noise, filters that are made up of products of powers of first-difference filters and first-sum filters.

Post-Posterior Exploration, Analysis, and Modeling

The posterior distribution of a complicated model such as ours is a complex numerical object. It is in many ways like the data themselves. Just as with data, it can require extensive exploratory study, for example by visualization methods, to comprehend its structure. We can find ourselves, simply as an exploratory device, fitting simplified mathematical functions to the posterior information as a way of summarizing the information. Even more, it can be judicious to impose even further model specifications, using exogenous information not incorporated at the outset and combining the posterior and exogenous information directly rather than returning to the original model specification. Cleveland and Liu (1998a) discuss in detail these issues of *post-posterior exploration, analysis, and modeling.*

Trellis Display

Our model building, our model diagnostics, and our post-posterior exploration, analysis, and modeling depend heavily on trellis display, a framework for the visualization of multivariable data (Becker, Cleveland, and

Shyu, 1996; Becker and Cleveland, 1996). Its most prominent aspect is an overall visual design, reminiscent of a garden trelliswork, in which panels are laid out into rows, columns, and pages. On each panel of the trellis, a subset of the data is graphed by a display method such as a scatterplot, curve plot, boxplot, 3-D wireframe, normal quantile plot, or dot plot. Each panel shows the relationship of certain variables conditional on the values of other variables.

2.2 Contents of the Paper

Section 3 describes the data. Section 4 attacks the scale of measurement of the data. The respondent ratings are integers from 1 to 10. We had to make an important decision: build a discrete model whose rating scale allows only integer values from 1 to 10, or build a model whose ratings emanate from an underlying continuous scale that is rounded to discrete values by the measurement process. Section 4 describes the decision and the basis for it.

Section 5 briefly describes the model building process. The result of the process is a Bayesian model: sharp specifications and prior distributions on parameters arising from the sharp specifications. The model is partly hierarchical and has a time series component. The model specifications include the following: (1) Survey respondent effects are modeled by random location and scale effects, a t-distribution for the location and a Weibull distribution for the scale. (2) Company effects for each attribute through time are modeled by integrated sum-difference processes. (3) Error effects are modeled by a normal distribution whose variance depends on the attribute; in the model, the errors are multiplied by the respondent scale effects. Section 6 describes the full model.

Earlier we stated two goals in analyzing the data: (1) characterizing the relative performances of the companies on the attributes, and (2) studying the basic tenets of CVA. In Section 7 we briefly describe how information in the posterior distribution is used to achieve these two goals.

2.3 What is Not in the Paper

Our actual model building was a highly iterative process in which we accumulated specifications for a complete model, fitted the complete model, carried out a variety of diagnostic checks (including posterior predictive checks), found inadequacies, altered the model, carried out more diagnostics, and so forth. The actual process was far too complicated to remember let alone recount. But by the end of our process we had discovered a logical path of accumulating assumptions that serves as one form of validation of our model; as we move along the path, we are able to use previous assumptions together with exploration of the data and our exogenous knowledge to form additional assumptions. The logical path, a series of steps, does begin with a specification of overall structure, but this was checked by a series of diagnostics. Section 5 describes the logical path rather than the actual model building process.

We carried out posterior predictive checking (Rubin, 1984; Gelman, Meng, and Stern, 1996). Because of the extensive model building process we did not need to rely fully on such checking to justify our model, but we did generate data from the posterior predictive distribution and plot it in a variety of ways. This is not discussed in the paper.

Computation of the posterior was a challenging matter because of the complexity of the model. We report briefly on computational matters in Section 6 but far more care than described there needed to be given. For example, convergence was a matter that needed much attention but is not discussed here.

3 The Data Studied Here

Our survey is administered on a quarterly basis. For the data studied here, 19 supplier companies are rated over a period of 9 quarters. Altogether there are 3505 rating sessions, or pollings; in each polling, a person rates one company on 9 attributes on a scale of 1 to 10 where 1 = poor and 10 = excellent.

A small number of people participated more than once. There are 3385 people. 3271 people were polled once, 106 were polled twice, and 6 were polled three times, which results in

$$3505 = 3271 + 2 \times 106 + 3 \times 6$$

pollings.

The 9 polled attributes cover quality, technology, price, and the interaction between the customer and the company. The attributes and their abbreviated names are the following:

prod-qual product quality

over-qual overall quality of company processes

delivery delivery of the product

cost cost of the product to the customer and the process of working with the customer to establish the cost

features technological excellence of product

pre-sup providing product information before delivery to support customer processes

response responsiveness of the company to customer needs

value value of the product relative to the cost of the product

service service provided by the sales team

The survey is designed with a skip pattern that results in no respondent rating all nine attributes; only purchasing agents rate `delivery` and only product and process designers rate `pre-sup`. The remaining 7 attributes are intended to be rated by all respondents, but some respondents choose not to rate all attributes about which they are asked. Table 1.1 shows counts of the number of attributes rated in the 3505 pollings:

Number of Attributes Rated	1	2	3	4	5	6	7	8
Number of Pollings	5	9	26	64	150	350	1106	1795

TABLE 1.1. Counts of Pollings According to Number of Attributes Rated

The number of respondents selected for polling for each company in the survey is determined by a number of factors but a chief one is the market share. Companies with a larger share have a larger number of customers.

Table 1.2 shows the number of ratings for all combinations of companies and attributes. The rows are ordered so that the totals for the companies increase from top to bottom. The columns are ordered so that the totals for the attributes increase from left to right.

4 Modeling: The Measurement Scale

The survey response scale is an integer scale from 1 to 10. We must make a critical decision here, at the onset, that will set the strategy of the ensuing model development. Should we build an ordinal-scale model that allows only integer values from 1 to 10, or should we suppose respondents have an underlying continuous scale that is rounded to an integer by the measurement process?

With a continuous underlying scale, we could divide the measurement scale into disjoint rounding intervals and estimate the interval endpoints (Albert and Chib (1996), Bradlow and Zaslavsky (1997), Johnson (1997)). Or, we could ignore the rounding with a supposition that a continuous scale is a good approximation to the measurement scale. And, in the continuous case, we would search for a transformation of the data to make the model simpler.

For a continuous model it is likely that the analysis would be simpler and results would be easier to comprehend. The rating scale would then, implicitly, be given a stronger metric interpretation. Model development, model diagnostics, and model interpretation would focus on a comparison of the relative values of ratings on a continuous scale, rather than focusing

	deli	pre-	prod	serv	cost	resp	valu	feat	over
Nut	15	15	25	29	28	32	33	33	32
Gas	17	25	26	34	40	43	42	44	46
Ham	13	35	38	43	42	47	46	50	50
Pub	32	32	56	59	61	67	68	70	71
Key	35	41	49	70	72	76	77	77	76
Cab	25	74	65	85	88	104	107	113	111
Ear	28	74	75	92	94	102	106	106	107
Bee	38	76	87	114	119	121	128	131	132
Mug	58	71	106	116	120	132	136	138	137
Jet	26	128	102	138	147	153	158	162	162
Rug	54	100	124	148	163	169	174	167	174
Log	63	97	121	156	165	167	170	172	171
Toy	47	110	132	145	152	170	177	178	181
Oak	84	110	169	174	197	205	218	219	216
Ace	77	174	186	254	240	270	267	273	275
Inn	83	187	180	255	260	283	287	287	285
Ski	112	189	250	282	300	321	334	328	335
Duo	121	292	309	408	389	420	420	431	430
Fan	132	303	320	400	403	431	455	465	463

TABLE 1.2. Counts of Pollings According to Number of Questions Answered

on relative distributions of probabilities of ratings on an ordinal scale, which would likely lead to more easily fathomed results. And computations for the continuous scale would likely be simpler, or at the very least, involve better understood methods.

Figures 1 to 3 are a trellis display of the ratings. Each panel has a histogram of all ratings of one attribute for one supplier company; the company and attribute labels are in the strip labels at the tops of the panels. Each block of panels has the 19 company histograms for one attribute; there are three blocks on each page.

The histograms show substantial regularity in the ratings. The mode is typically at 7 or 8. Because the scale ranges from 1 to 10, more of the scale is available below the mode than above. This results in skewness to the left. But there does not appear to be an undue build-up at the high end of the scale, in particular, at 10.

The regularity of the ratings and the lack of build-up at either end of the scale are a positive signal that a continuous model is at least worth attempting. The regularity and lack of build-up also suggest that we can take the simple route of supposing that each response k arises from rounding a continuous value in an interval extending from $k - 0.5$ to $k + .5$. The rounding process has a variance of $1/12$, exceedingly small compared with the variability that we observe in the histograms, so we will attempt a first model that treats the ratings as if they were continuous variables from 1 to

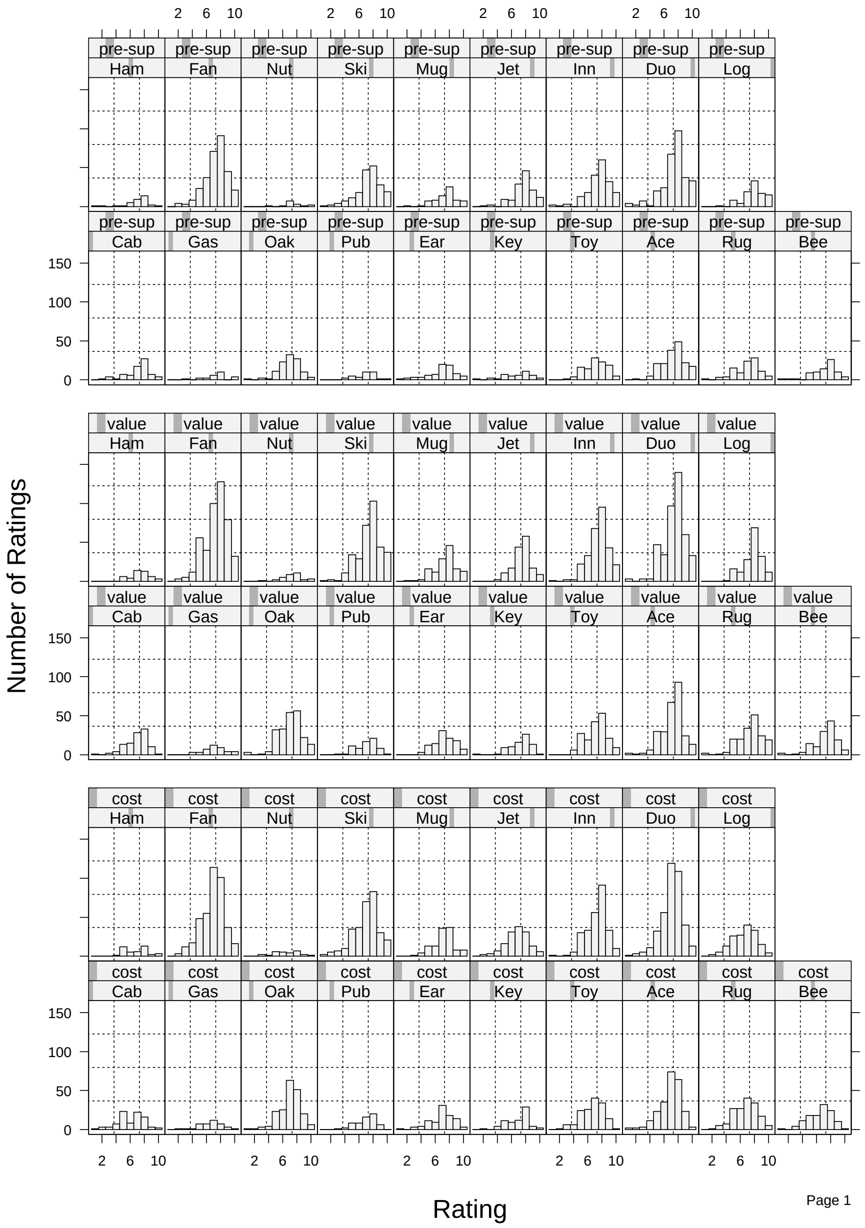

FIGURE 1. Histograms of Rating Given Company and Attribute

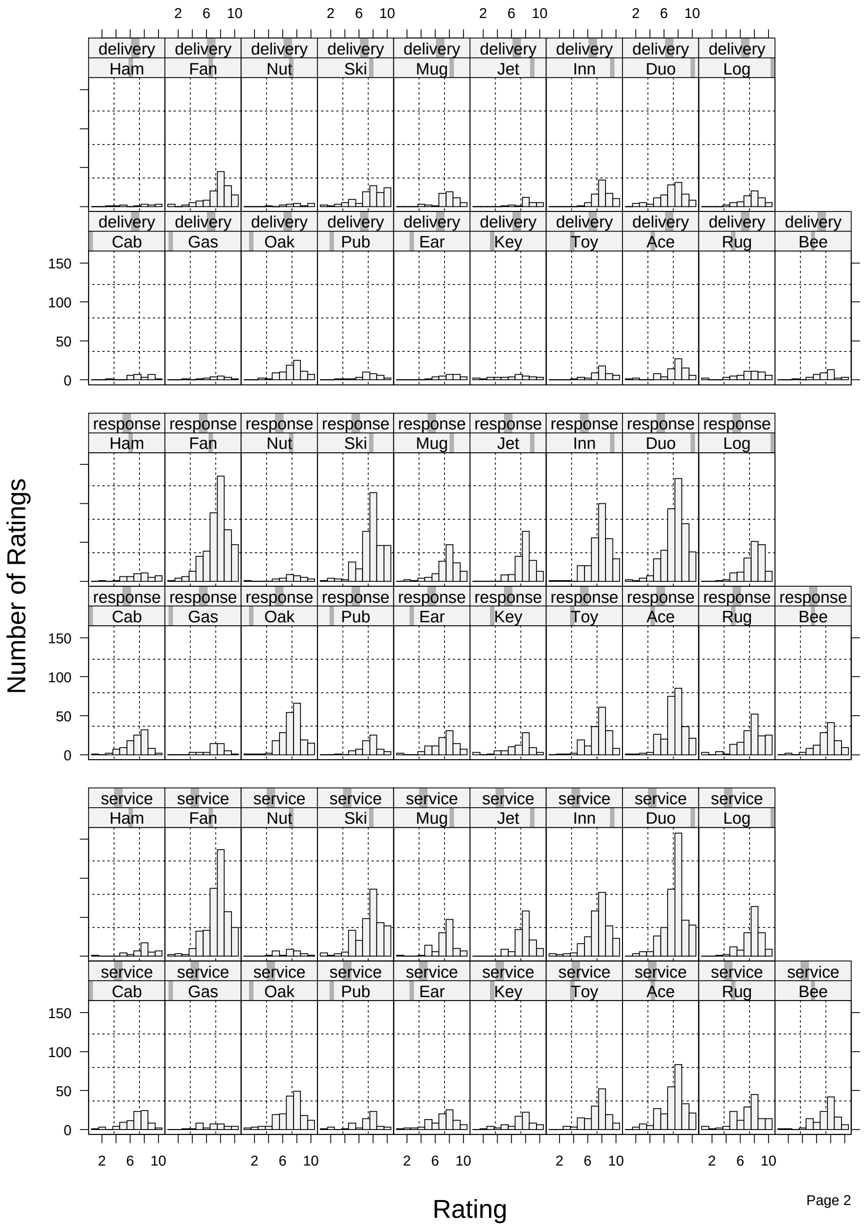

FIGURE 2. Histograms of Rating Given Company and Attribute

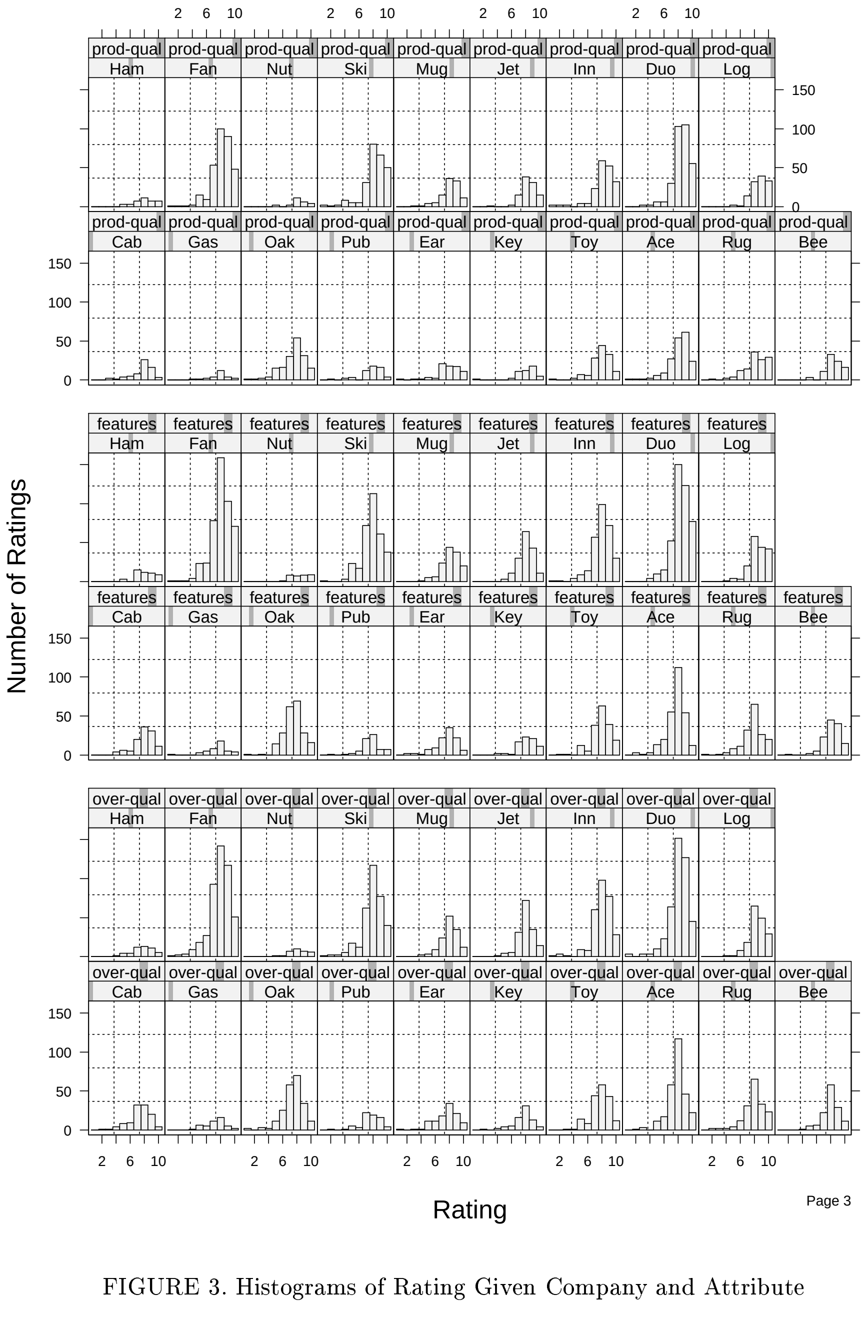

FIGURE 3. Histograms of Rating Given Company and Attribute

10 and ignore the rounding. But this approach would rest on a more solid foundation if the scale of measurement resulted in symmetric distributions rather than skewed.

Next, we attack the skewness problem. The histograms of Figures 1 to 3 provide less incisive assessments of distributional shapes and we move to normal quantile plots in Figure 4; in the interest of space we show the plots only for the `prod-qual` attribute, but the shapes for other attributes are quite similar. As with all of the quantile plots to come, we use the following procedure. Suppose the data for a particular quantile plot are x_i for $i = 1$ to n. Suppose F is a theoretical distribution. To check whether the empirical distribution of the data is well approximated by F we plot $x_{(i)}$, the i-th order statistic, against $F^{-1}((i - 0.5)/n)$.

The skewness of the data is apparent in Figure 4 but the discontinuity of the patterns arising from the discreteness of the data does interfere with our ability to visually judge the adequacy of an underlying continuous distribution. To assist our visual processes we will smooth the distribution function of the data by the following procedure and then display the smoothed data quantiles. Consider the ratings of one attribute for one company. Suppose, for $k = 1$ to 10, that there are $n(k)$ values equal to k, then we replace the $n(k)$ values of k by $k - 0.5 + (i - 0.5)/n(k)$ for $i = 1$ to $n(k)$. The result is shown in Figure 5. In addition, each panel has lines on it. The oblique line is drawn through two points: the lower quartile point and the upper quartile point. The horizontal lines are drawn at the 0.01, 0.1, 0.9, and 0.99 quantiles of the data; their purpose is to show where the majority of the data lie.

Figure 5 suggests that a transformation might well symmetrize the data. Since we must push in the lower tail and pull in the upper we will try the transformation $a \log(11 - \text{rating}) + b$, where a and b are chosen so that a rating of 1 becomes 1 on the transformed scale and a rating of 10 becomes 10. Figure 6 displays probability plots of the transformed, smoothed ratings. Unfortunately, the log transformation is too drastic; the data are now skewed to the right. We will use a square root transformation, $a\sqrt{(11 - \text{rating})} + b$. Figure 7 shows the result. The data are now symmetric but there is leptokurtosis: longer tails than a normal.

We have gone from discrete data, bounded by a rating scale of 1 to 10, to symmetric transformed data that suggest an underlying continuous scale for which the rating distributions stretch out in the tails even more than a normal. This makes it far more likely that a continuous model will prove a judicious choice. We will use the transformed scale exclusively in the remainder of the paper, referring to the transformed ratings as the "ratings".

The discussion in this section has attempted a certain level of verisimilitude, approximating our thinking as it was at the beginning of our analysis. In real life, at this stage, we had only a good hunch that a continuous approximation, an ignoring of the rounding, and a square root transformation

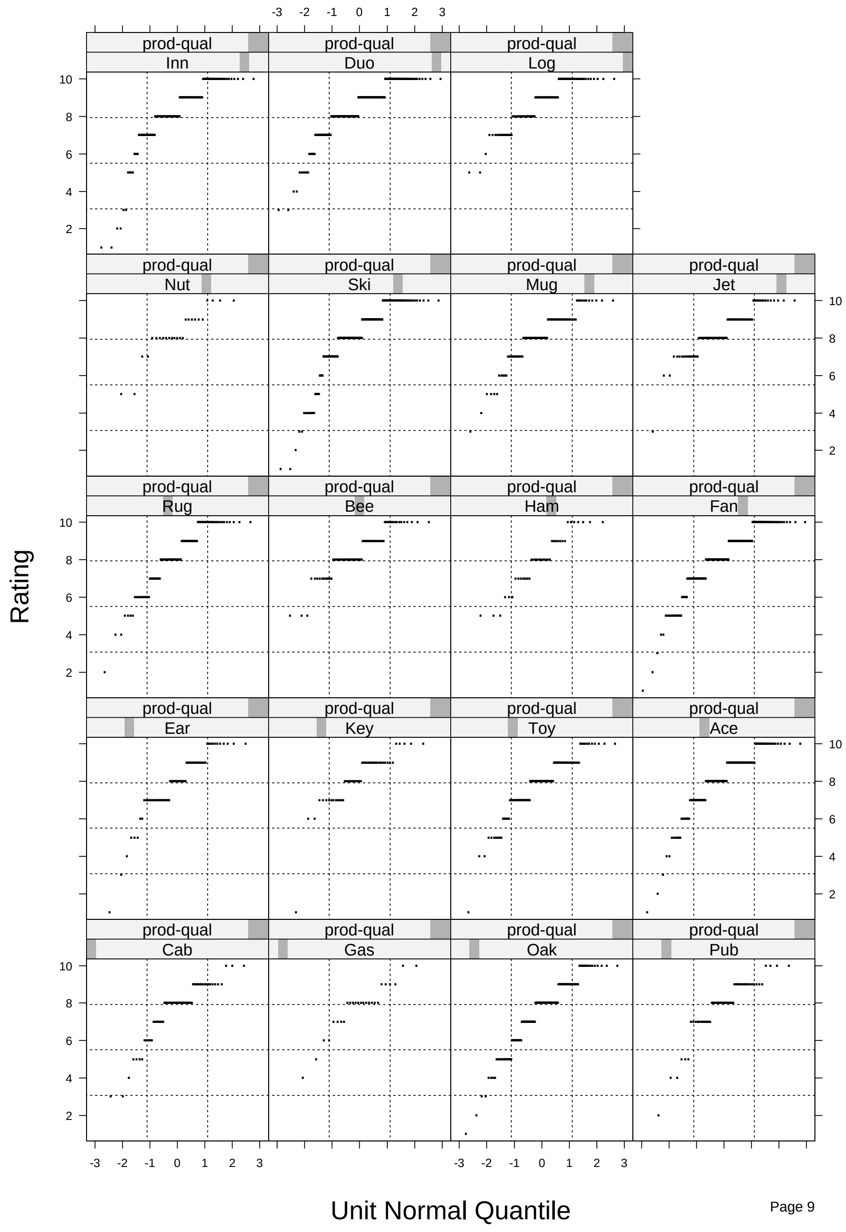

FIGURE 4. Normal Quantile Plots of Rating Given Company and Attribute

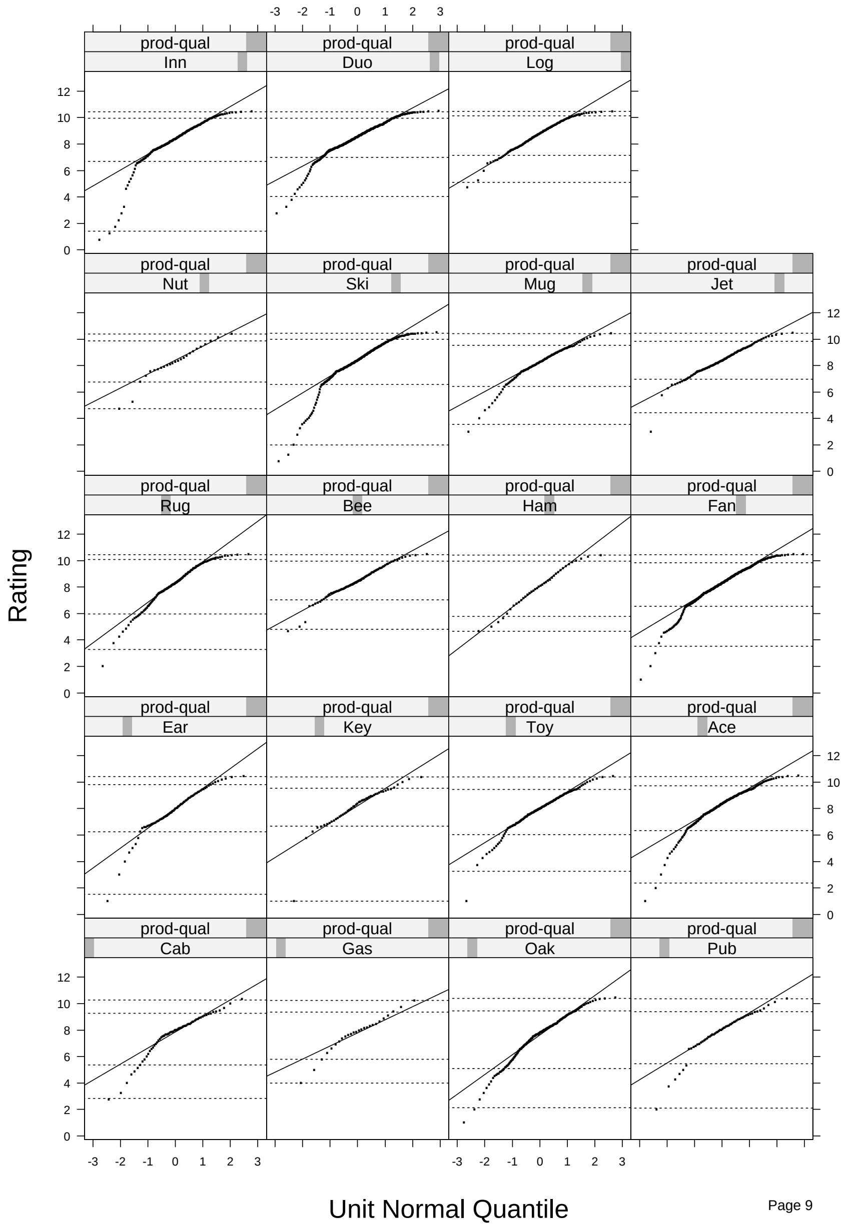

FIGURE 5. Normal Quantile Plots of Smoothed Rating Given Company and Attribute

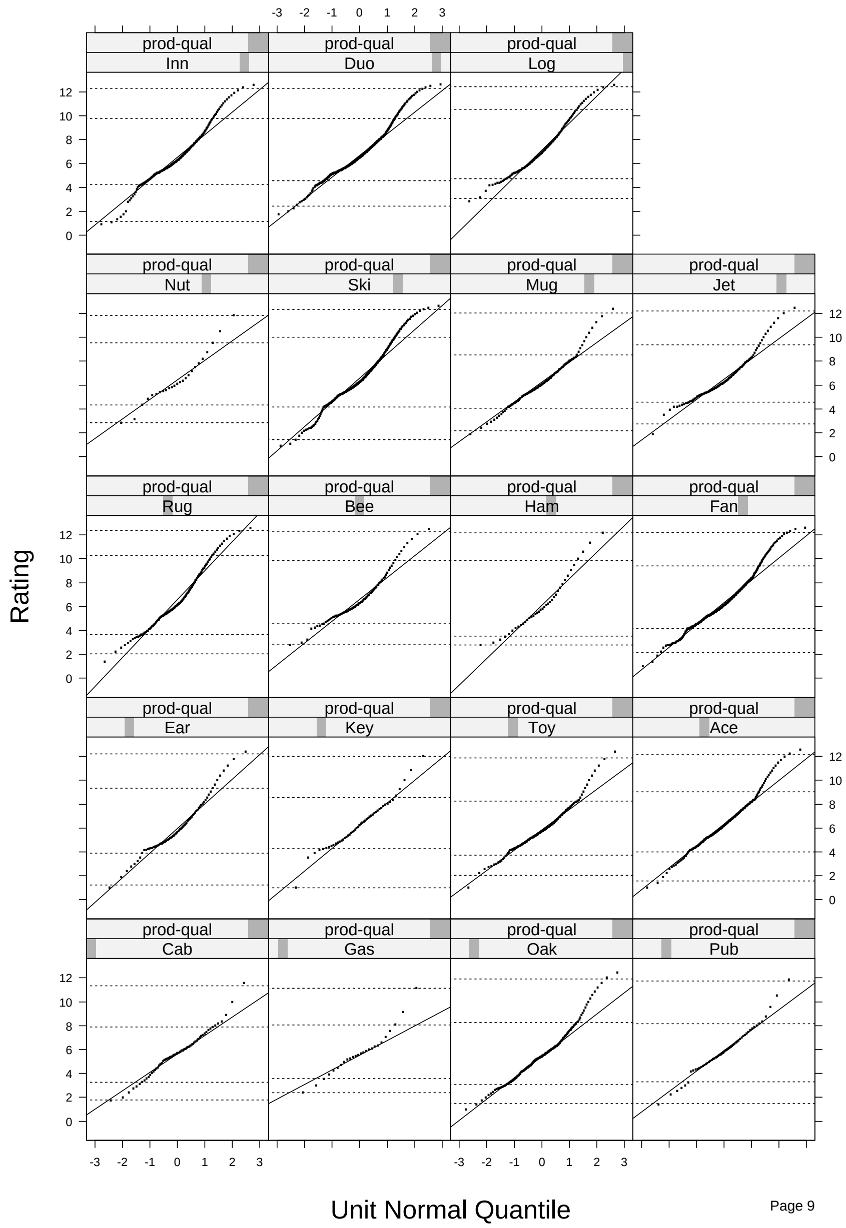

FIGURE 6. Normal Quantile Plots of Log Transformed Smoothed Rating Given Company and Attribute

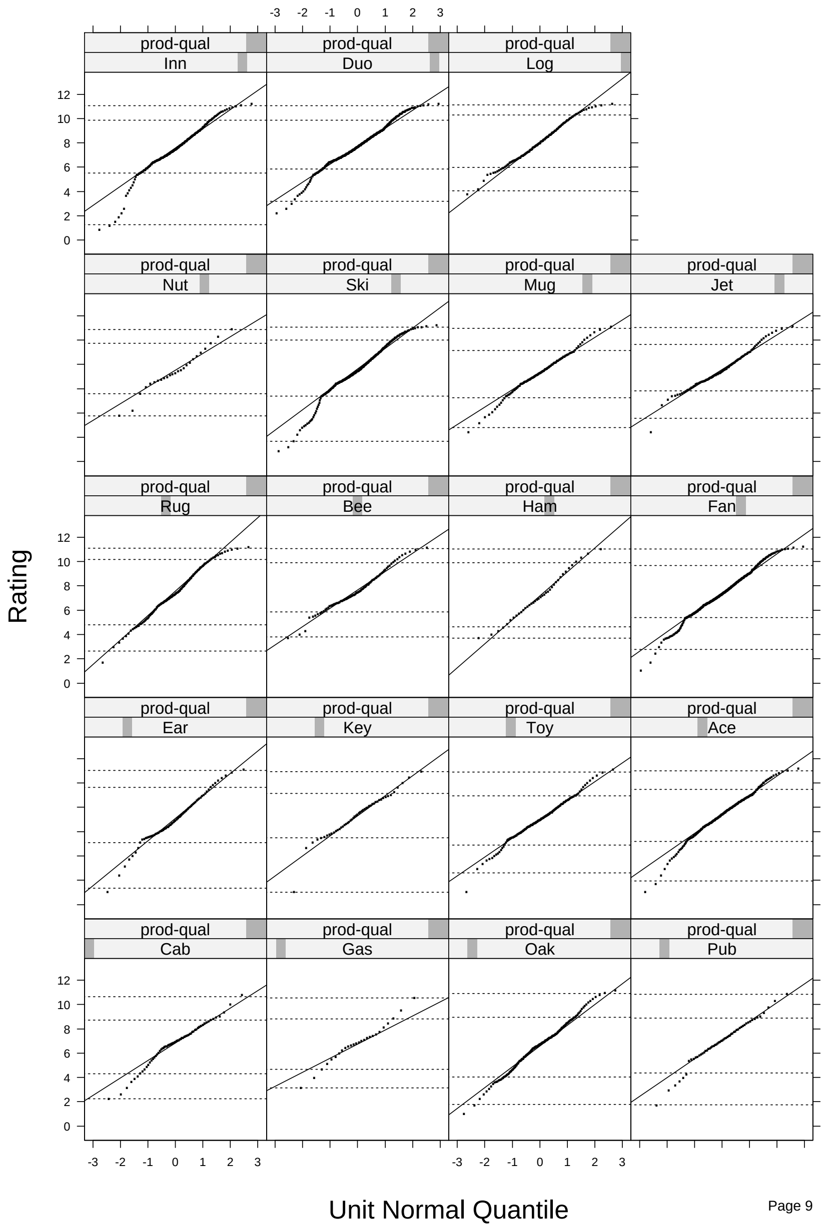

FIGURE 7. Normal Quantile Plots of Square-Root Transformed Smoothed Rating Given Company and Attribute

would lead to a model that was parsimonious, that accorded with our exogenous knowledge about CVA, and that appeared to fit the data. We can now assert its success but only because of the ensuing data analysis.

5 A Summary of the Model Building

The supplier company data of this paper are an example of case data, which are pervasive in the sciences, engineering, business, and elsewhere. Such data consist of a collection of cases and for each case there are measurements of one or more variables. For our survey, the customers are the cases. Cleveland, Denby, and Liu (1998) present a collection of basic models and model building tools for case data, and it is this methodology that we invoked in building an initial model for our survey data.

In the survey, a person in a single polling rates one company in one quarter on the 9 attributes. As we saw in Section 3, there were 3505 pollings. Most of the pollings involved people who participated just once. Only 3.3% of the pollings involved a person who had already participated. To simplify the model building process, we supposed that a repeat is a different person. (Later in the paper, when we describe the complete model, we will we will take account of repeats, specifying them in the model from exogenous information.)

Step 1: Overall Structure to Guide the Data Exploration

We entertained an initial structure on which to base our exploration of the data:

$$r_{ac} = \theta_{asq} + \mu_c + \alpha_c \epsilon_{ac}$$

where r_{ac} is the rating (on the transformed scale) by person (case) c on attribute a. θ_{asq} is the rating performance for supplier company s on attribute a in quarter q. (In our notation, r_{ac} does not explicitly show a dependence on s and q; because we have assumed that each polling is a separate person, s and q are determined by c.) μ_c is a person location effect, and α_c is a person scale effect. ϵ_{ac} is an error term. We allow for a dependence of the error variance on attribute by supposing $E\epsilon_{ac}^2 = \sigma_a^2$.

The equation provides a framework for data exploration that entertains as few specifications as possible. It allows us to study the few specifications that are made and to develop new ones.

Step 2: Modeling Company, Time, and Attribute Effects

We visualized two functions of the data to study specifications for θ_{asq}. The first was the sample means across c for each combination of a, s, and q. The second was the sample means across c and q for each combination of a and s. This led to certain tentative decisions about modeling the θ_{asq}.

First, for fixed a and s, the θ_{asq} were taken to be an ISD(1,1) process as a function of the time parameter q. This accords with our exogenous information that customer opinions have a substantial low-frequency component, but also allows for a for high-frequency quarter-to-quarter variation.

Second, while we have substantial exogenous information and information from the data about the relationships of the θ_{asq} for fixed c and q — that is, relationships among the attributes — we chose not to impose strong specifications to reflect this information to its fullest. The reason for this is discussed in Section 7.

Step 3: Adjusting for Company, Attribute, and Time Effects

To get at person effects, we adjusted the data for the company, attribute, and time effects by fitting loess curves to the ratings as a function of q for each combination of a and c, and then subtracting the fitted curve. In other words, we adjusted to remove the effect of the θ_{asq}. Because a large number of observations are pooled to estimate the θ_{asq}, we assume the estimates are the true values and, with a slight abuse of notation, will take r_{ac} to denote the adjusted data:

$$r_{ac} = \mu_c + \alpha_c \epsilon_{ac}. \tag{5.1}$$

Step 4: Estimating Attribute Variances

We used an iterative method based on Equation 5.1 to estimate the σ_a^2. Because a large number of observations are pooled to estimate the σ_a^2, we assume the estimates are the true values.

Step 5: Least Squares Fitting

For each c, we fitted Equation 5.1 by least squares to get estimates $\hat{\mu}_c$ and $\hat{\alpha}_c$, and residuals, $\hat{\epsilon}_{ac}$.

Step 6: Modeling the Error Distribution

Beta-quantile plots of studentized residuals, $\hat{\epsilon}_{ac}/\hat{\alpha}_c$, led to a specification of a normal distribution for ϵ_{ac}.

Step 7: Modeling the α_c Distribution

The assumption of a normal distribution for the ϵ_{ac} implies that the $\hat{\alpha}_c^2$ have a distribution that is a multiplicative convolution of the α_c^2 distribution with a $\chi^2\{d(c)\}/d(c)$ distribution where $d(c)$ is the number of ratings by person c minus 1. We were able to deconvolve using quantile plots, which led to a specification of the distribution of α_c^2 as a unit exponential; thus the distribution of α_c is a Weibull.

Step 8: Modeling the μ_c Distribution

The assumption of a normal distribution for the ϵ_{ac} and the assumption of a Weibull distribution for $\hat{\alpha}_c$ implies that the $\hat{\mu}_c$ have a distribution that is an additive convolution of the distribution of μ_c with a product of a Weibull and a normal. We were able to deconvolve through quantile plots, which led to the specification of the distribution of μ_c as a $t\{\nu\}$-distribution with $\nu = 6$ degrees of freedom.

6 A Bayesian Model

6.1 Missing-Data Mechanism and Repeat Polling

In Section 3 we discussed two issues about respondents' participation that are critical for the modeling.

First, some respondents were polled more than once, sometimes rating the same company and sometimes rating a different one. The fraction of such people here is quite small. We could simply ignore this, as we did in the model building, and treat a repeat as a new person. However, the data studied here is a subset of our currently available data, and in our complete data set we have a much greater fraction of repeats. For this reason we will incorporate repeats in our model here, but we will rely on our exploration with the larger data set and ask the reader to take on faith that our modeled repeat mechanism is in accord with the data and exogenous information.

Second, there is missing data in the sense that respondents do not answer all questions, either as a result of the survey design or because a respondent decided not to rate an attribute. In our modeling we will suppose that the absence of a rating is ignorable (Gelman, King, and Liu, 1997; Rubin, 1976, 1977). In fact, in data exploration not discussed here, we have reason to believe that the mechanism is not ignorable, but the effect appears to be quite small and no threat to the validity of our analysis.

Except for these considerations above, we follow the sharp specifications of Section 5.

6.2 Notation

Notation is somewhat more cumbersome than in Section 5 because of the small fraction of people who repeat. As before, c denotes a person (case), a respondent in the survey. k denotes the repeat; $k = 1$ is the first polling of a person, $k = 2$ is the second polling for a person who repeats, and $k = 3$ is the third polling of a person who repeats twice. (No one participated more than three times.) s denotes the supplier company and q, the quarter. But company and quarter depend on person and repeat in the sense that if we know c and k, then we know s and q. Finally, as before, a denotes the attribute.

6.3 The Overall Form of the Model

We suppose

$$r_{ack} = \theta_{asq} + \mu_c + \rho_{ck} + \alpha_{ck}\epsilon_{ack} \tag{6.2}$$

where r_{ack} is the rating of attribute a by person c for the kth time. θ_{asq} is the effect of company s for attribute a in quarter q. μ_c, ρ_{ck}, and α_{ck} describe the person effects as described below, and ϵ_{ack} is the error term. To make the variables well-defined we suppose

$$E(\mu_c) = E(\epsilon_{ack}) = E(\rho_{ck}) = 0,$$

and

$$E(\alpha_{ck}^2) = 1. \tag{6.3}$$

6.4 Person (Case) Effects and Errors

Following, largely, Section 5, we will make the following specifications for the distributions of the person effects and errors:

$$\begin{aligned}
\mu_c &\sim \sigma(\mu)t\{6\} && (6.4)\\
\rho_{ck} &\sim \sigma(\rho)t\{6\} && (6.5)\\
\alpha_{ck}^2 &\sim exp\{1\} && (6.6)\\
\epsilon_{ack} &\sim \sigma_a(\epsilon)N\{0,1\}. && (6.7)
\end{aligned}$$

All of these random variables are independent of one another and of the θ_{asq}.

μ_c and ρ_{ck} are the person location effects. If $\sigma(\rho)^2 = 0$, then repeat locations for a person are the same. If $\sigma(\mu)^2 = 0$, then repeat locations for a person are distributed like those of different people. If both variances are positive then the location values for a person who repeats are not equal but tend to be nearer in value then those for different people. α_{ck} describe the person scale effects. Unlike for the location effects, repeat scales for a person are distributed like those of different people. ϵ_{ack} are the error variables.

6.5 Company, Attribute, Time Effects

For fixed a and c we suppose that for $q = 1$ to 9, the time series θ_{asq} is an ISD(1,1) process:

$$\begin{aligned}
\theta_{asq} &= \kappa_{asq} + \iota_{asq} && (6.8)\\
\kappa_{asq} - \kappa_{as(q-1)} &= \lambda_{asq} + \lambda_{as(q-1)}, && (6.9)
\end{aligned}$$

where the λ_{asq} and ι_{asq} are independent of one another, and each is a gaussian white noise series with variances $\sigma(\kappa)$ and $\sigma(\iota)$. This ISD(1,1) process is also an IMA(1,1). Formulating the structure as an ISD leads to a more fathomable parameterization that allows better incorporation of exogenous information and more efficient computational methods (Cleveland and Liu, 1998b).

In Section 7 we give reasons for not wanting to impose strong restrictions on the θ_{asq} across a for fixed s and q even though there is a body of exogenous information that we might invoke to do so. However, we will place some mild impositions on the values. The reason is this: some companies have a very small number of observations compared with others. Without at least mild imposition the result is a posterior distribution with mass for company-attribute effects for these small-sample companies that appears unrealistic. When a posterior distribution appears counter-intuitive it is typically an indication that we did not properly assess our exogenous information. Thus we are prepared to impose conditions that rein in the errant values, a hierarchical distribution on the 19×9 values of κ_{as1} for all combinations of a and s. We take them to be independent with

$$\kappa_{as1} \sim N\{\tau_a, 1\} \tag{6.10}$$

where the 9 τ_a to have improper uniform distributions.

6.6 Standard Deviations

The above model specifications have given rise to several variances: $\sigma_a^2(\epsilon)$, $\sigma^2(\mu)$, $\sigma^2(\rho)$, $\sigma^2(\kappa)$, and, $\sigma^2(\iota)$. We wish to place little restriction on the these parameters based on our exogenous information; we specify distributions for them by taking their inverses to be unit exponentials.

6.7 Posterior Computation

Advances in computational methods have made it possible to simulate posterior distributions of complex models such as ours using Markov chain Monte Carlo (MCMC). Among the large number of MCMC-related papers are Metropolis and Ulam (1949), Metropolis *et al.* (1953), Hastings (1970), Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990), which are key references that stimulated much work in this research area. The Gibbs sampler and the Metropolis-Hastings algorithm are used to simulate the posterior distribution of the variables in our model.

To draw the person location parameters, μ_c and ρ_{ck}, we use the conventional hierarchical representation of the t-distribution using the *normal/independent* distribution with gamma distributed *weights*. Detailed implementation is straightforward and relevant references appear in various places (see Liu (1995, 1996) and Pinheiro, Liu, and Wu (1997) for recent examples).

To draw person scale parameters, α_{ck}, we use the Metropolis-Hastings algorithm. The conditional distribution of ratings given the θ_{asq}, μ_c, and ρ_{ck} is a mixture of normals with mean zero where the inverse variance is mixed by an exponential, that is, a gamma. Because it is computationally easy to simulate these variables when the gamma is replaced with an inverse gamma, we make use of this fact and then employ the Metropolis-Hastings algorithm to ensure correct MCMC. The jumping rate of the Metropolis-Hastings algorithm is greater than 80% when the degrees of freedom for the inverse gamma approximating $\chi^2_{\nu_\tau}/\nu_\tau$ is $\nu_\tau = 6$.

To draw the components in the time series structure, we use the methods proposed by Cleveland and Liu (1998b) for $\{\kappa_{asq}\}$. Taking draws of $\{\theta_{asq}\}$ requires taking draws from the normal distributions that are determined by the shrinkage between the (conditional) normal distributions for the polling ratings and the (conditional) normal distribution of $\{\kappa_{asq}\}$.

7 The Posterior Distribution

In Section 1 we set out two goals in the analysis of the survey data — (1) characterize the *relative performance* of the companies on the attributes; (2) study *CVA theory*, both to test old tenets and to develop new ones. In this section we briefly illustrate how we use information in the posterior distribution together with exogenous information not incorporated into the posterior to move toward these goals. First, we will discuss CVA theory, then, relative performance, and finally, CVA theory again.

7.1 CVA Theory: Components of Variation

The model equation that relates a rating to supplier company, attribute, quarter, person, and error effects, breaks the ratings into four additive components of variation:

$$r_{ack} = \theta_{asq} + \mu_c + \rho_{ck} + \alpha_{ck}\epsilon_{ack}.$$

The θ_{asq} are broken further into two additive components of variation:

$$\theta_{asq} = \kappa_{asq} + \iota_{asq}.$$

For fixed values of θ_{asq}, ratings vary because of person and error effects: (1) the person location effects, μ_c, with variance $\sigma^2(\mu)$; (2) the person location repeat effects, ρ_{ck}, with variance $\sigma^2(\rho)$; (3) the person scale effects, α_{ck}, with variance 1, times the error effects, ϵ_{ack}, with variance $\sigma^2_a(\epsilon)$. For each combination of a and s, ratings vary because of company-attribute time effects: (1) a smooth, or low-frequency, component κ_{acq} where

$$\kappa_{asq} - \kappa_{as(q-1)} = \lambda_{asq} + \lambda_{as(q-1)},$$

and λ_{asq} is a white noise series with variance $\sigma^2(\lambda)$; (2) a white noise component, ι_{asq}, with variance $\sigma^2(\iota)$.

Table 1.3 shows the posterior means of the variances of the components of variation. To our knowledge, this delineation and quantification of variation is the first in the area of CVA customer opinion polling, and it has important implications for CVA theory and the design of CVA surveys.

$\sigma^2(\lambda)$	0.012
$\sigma^2(\iota)$	0.024
$\sigma^2(\mu)$	0.71
$\sigma^2(\rho)$	0.71
$\sigma^2_{over-qual}(\epsilon)$	0.85
$\sigma^2_{response}(\epsilon)$	1.27
$\sigma^2_{prod-qual}(\epsilon)$	1.28
$\sigma^2_{value}(\epsilon)$	1.45
$\sigma^2_{pre-sup}(\epsilon)$	1.66
$\sigma^2_{features}(\epsilon)$	1.70
$\sigma^2_{service}(\epsilon)$	1.86
$\sigma^2_{delivery}(\epsilon)$	1.96
$\sigma^2_{cost}(\epsilon)$	2.07

TABLE 1.3. Posterior Means of Variances of Components of Variation

For fixed θ_{asq}, the posterior mean of the variance of a rating depends on the attribute and ranges from

$$E\{\sigma^2(\mu) + \sigma^2(\rho) + \sigma^2_{over-qual}(\epsilon)\} = 0.71 + 0.71 + 0.85 = 2.27$$

to

$$E\{\sigma^2(\mu) + \sigma^2(\rho) + \sigma^2_{cost}(\epsilon)\} = 0.71 + 0.71 + 2.07 = 3.49$$

The variation about θ_{asq} is substantial and has sobering implications about overall survey design. To illustrate, consider the `over-qual` attribute. In the information industry, scores appear to range from about 6.5 to 7.5, so worst to best has a range of 1. (We will see this shortly in our displays of company performance.) What sample size does it take to put the posterior probability of a company's performance in an interval of ± 0.3 to provide insight about which companies are at the top and which at the bottom? Instead of a full posterior predictive simulation we will be satisfied here with a back-of-the-envelope calculation to get a rough picture. Suppose the survey only asks about `over-qual` for one company and that the true variance is 2.27, whose square root is 1.51 Then with a diffuse exogenous distribution for the company performance, the posterior mass would be contained, roughly, within

$$\pm 1.96 \times 1.51/\sqrt{n} = 2.95/\sqrt{n}.$$

Thus we need a sample size of about 100 to achieve a ± 0.3 interval.

7.2 Relative Performance

To study relative performance we will focus on the κ_{acq}, the smooth component of κ_{acq}. The posterior distribution of the κ_{acq} conveys information about the relative performance of the companies on the attributes. Extensive study of this posterior has yielded important insight into performance. The following briefly illustrates some of the study.

The nine-quarter average performances are

$$\kappa_{as} = \frac{\sum_q \kappa_{asq}}{9}.$$

We use the normal distribution to approximate their posterior distribution. The means and 95% intervals of the approximating distribution are plotted in the trellis displays of Figures 8 and 9. Each plotted dot is the posterior mean for one attribute and one company. The line segment about the dot is the 95% posterior interval. The values on the two displays are the same but are organized differently. Figure 8 graphs the posterior values against attribute given company; each panel shows the 9 attribute means for one company. Figure 9 graphs the posterior values against company given attribute; each panel shows the 19 company means for one attribute. To enhance our perception of patterns in the data, both the attributes and the companies are ordered. The 19 companies are ordered by the 19 arithmetic averages of the posterior means across attributes. The 9 attributes are ordered by the 9 arithmetic averages of the posterior means across companies. In Figure 8, the attributes are ordered, going from bottom to top in the panels, so that the attribute averages increase. And, as we move from left to right and then bottom to top through the panels, the company averages increase. In Figure 9, the companies are ordered, going from bottom to top in the panels, so that the company averages increase. And as we move from left to right and then bottom to top through the panels, the attribute averages increase.

Figure 8 shows that a strong attribute main effect produces similar orderings of the posterior means for the companies. The variation in the values for each company is typically 2 or somewhat less. `features` and `prod-qual` are typically first and second in rank order. `over-qual` is typically third or close to third. The next four are `delivery`, `response`, `service`, and `pre-sup`. `cost` is typically last. `value` is typically next to last or close to it. This result accords with our exogenous information. For example, the order reflects a natural grouping. `features` and `prod-qual` are attributes of the physical product. `delivery`, `response`, `service`, and `pre-sup` involve the interaction of the customer with the company. `over-qual` is a mixture of all attributes in these two groups and so ranks toward the middle although it tends to be closer in value to the product attributes because

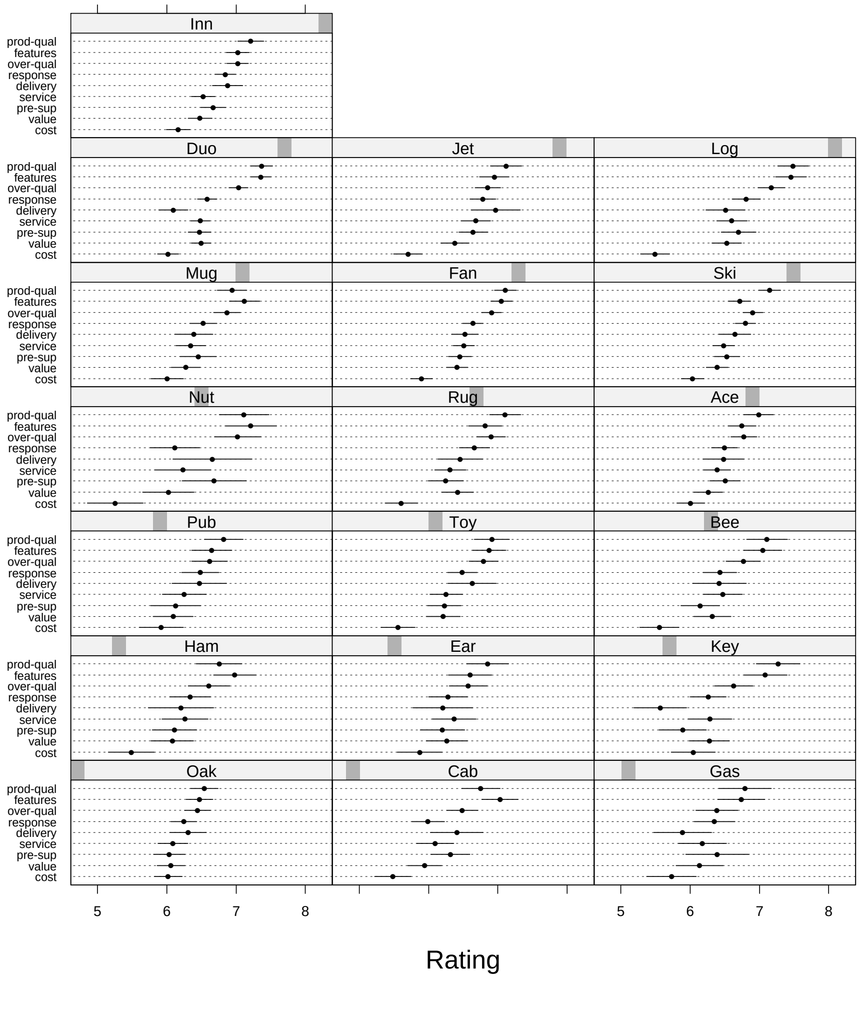

FIGURE 8. Posterior Distribution: Nine-Quarter Average Rating Given Company

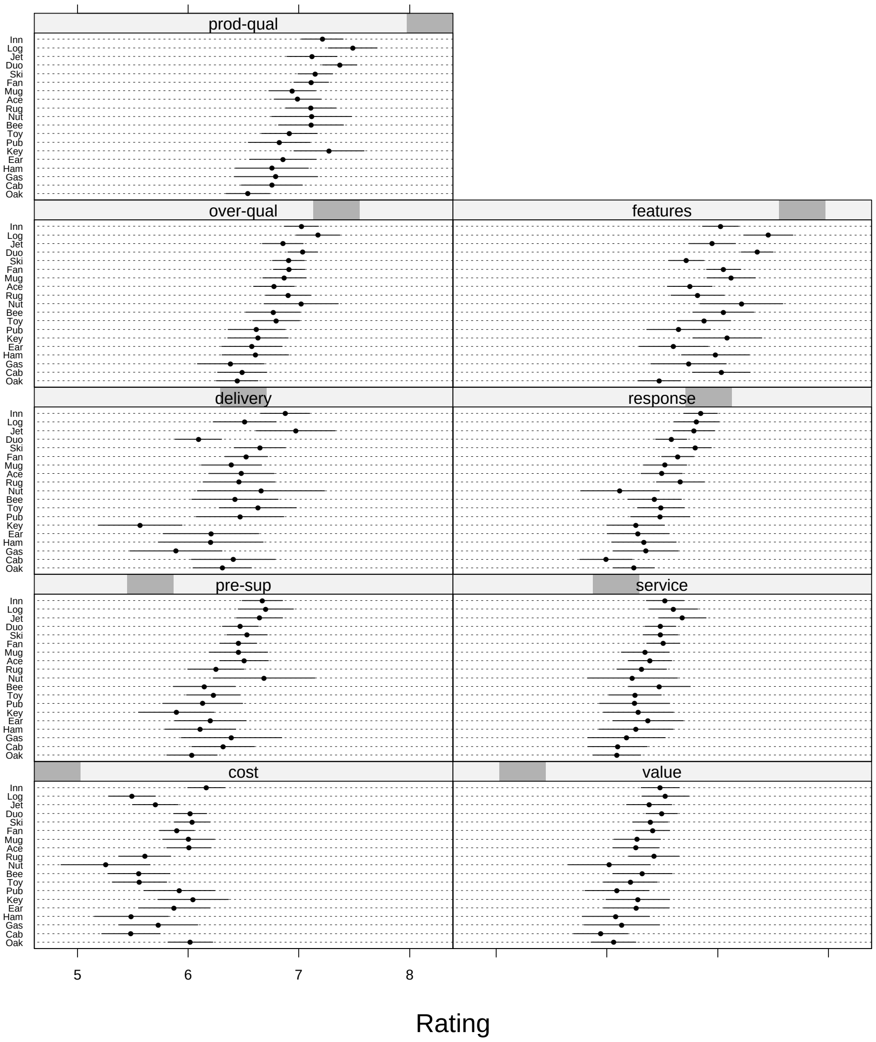

FIGURE 9. Posterior Distribution: Nine-Quarter Average Rating Given Attribute

it is most strongly linked in the minds of customers with `prod-qual`. `cost` and `value` are associated with the pricing of the product and services and its resulting value to the customer. The attribute that breaks with the ordering most frequently is `delivery`, but its precision is the lowest among the attributes because it has the smallest sample sizes, so greater variation is expected.

Figure 9 shows there is a noticeable company main effect as well, but not as strong as the attribute main effect; the variation in the values for each attribute is typically 1 or less . For each attribute, there is a similar ordering of companies for the different attributes with the strong exception of `cost` and the mild exception of `delivery`. To see the company effects and the interactions more clearly, we will adjust for attribute. For each attribute we define a weighted average

$$\kappa_a = \frac{\sum_s w_{as}\kappa_{as}}{\sum_s w_{as}}$$

where w_{as} is the inverse of the posterior variance of κ_{as}. The adjusted posterior means are $\kappa_{as} - \kappa_a$.

Figures 10 and 11 are trellis displays of the adjusted posterior means with 95% posterior intervals. Figure 10 shows quite clearly the company main effects; for example, overall, `Log`, `Duo`, and `Inn` are at the top of the industry, and `Cab`, `Gas`, and `Oak` are at the bottom. The figure also shows interactions in the form of isolated values that stand out from the others on a panel. These are cases where a company has a market position on an attribute that deviates by a significant amount from its overall market position.

Most of these isolated values occur for `delivery` and `cost`. A few isolated values involve other attributes. For example, `Ski` has a reduced market position on `features` while `Ham` has an increased one. It is displays such as these that are carefully studied to determine actions to improve performance. The tendency is to take action on an attribute for which a company is doing most poorly relative to competitors.

We turn now to the quarterly values κ_{asq} We will also adjust them by the weighted attribute averages, studying

$$\kappa_{asq} - \kappa_a.$$

Figures 12 to 16 are a trellis display of the posterior means of the adjusted quarterly values and 95% posterior intervals. The companies are ordered by the number of observations for each; as we go from bottom to top and through the pages the number of observations decreases. The effect of the ordering is a general increase in the sizes of the 95% posterior intervals we go from bottom to top and through the pages. Only a few trends change by amounts that are not small compared with the precisions. For example, `Log`, at the top in `prod-qual` overall, has had a decrease that threatens its position.

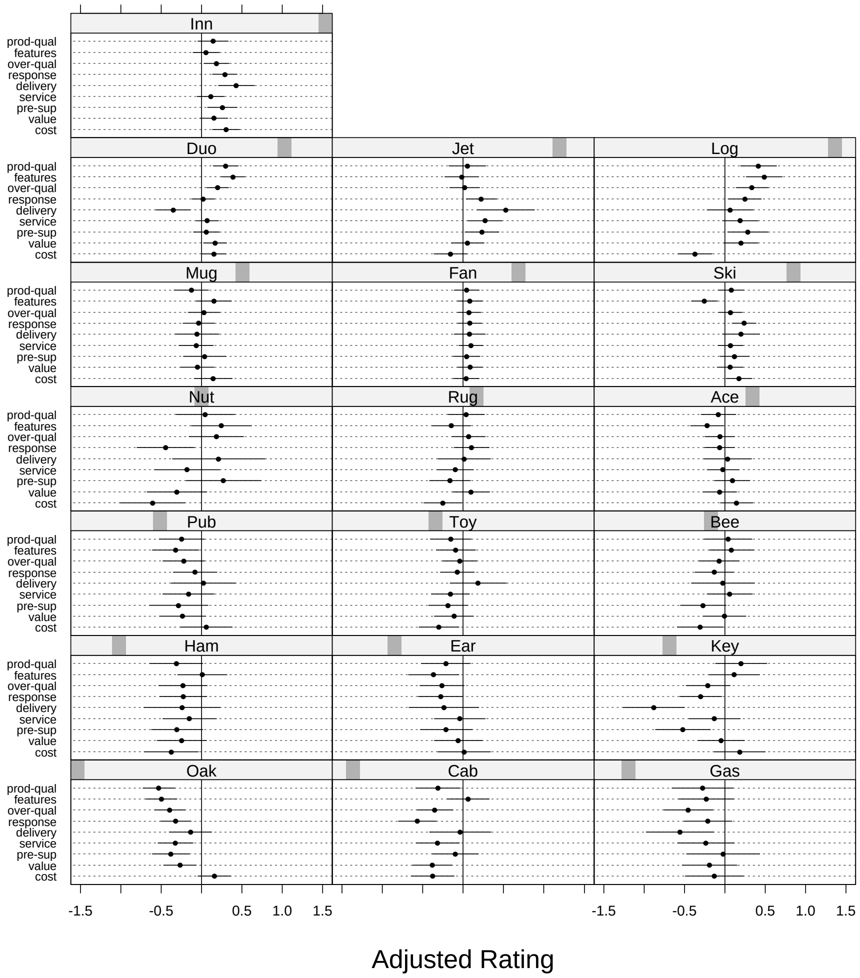

FIGURE 10. Posterior Distribution: Adjusted Nine-Quarter Average Rating Given Company

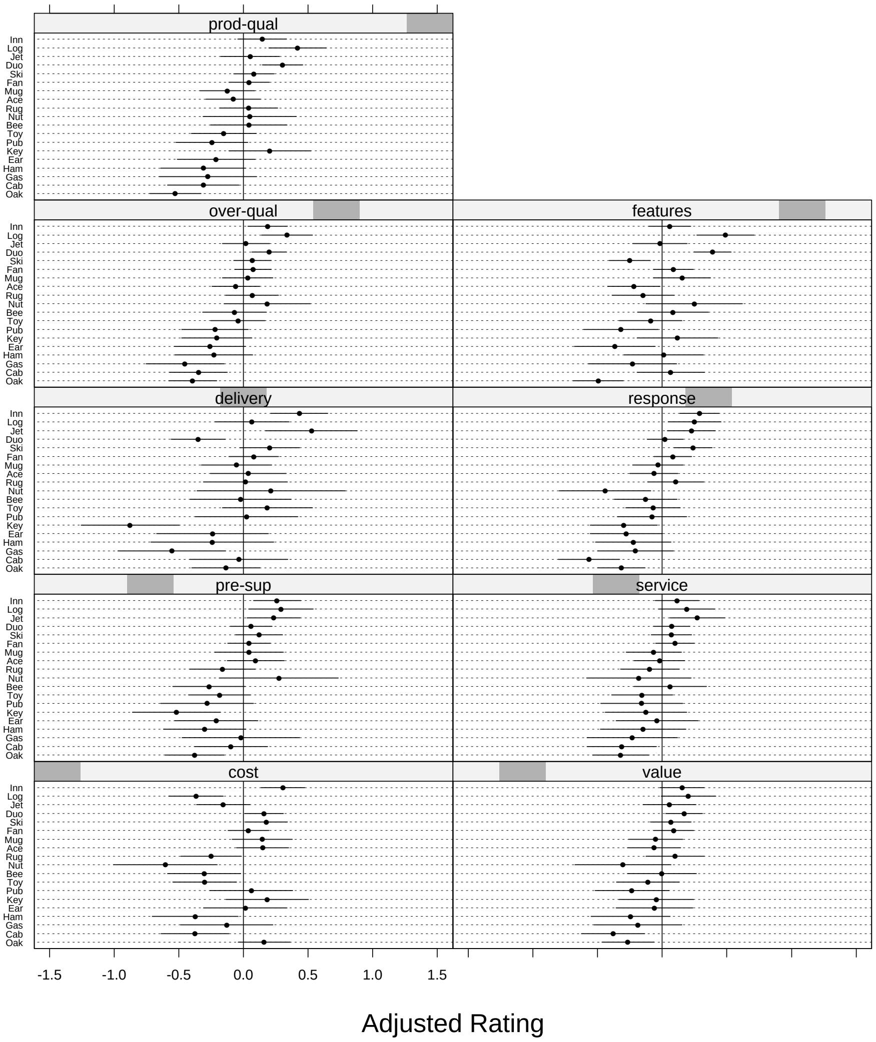

FIGURE 11. Posterior Distribution: Adjusted Nine-Quarter Average Rating Given Attribute

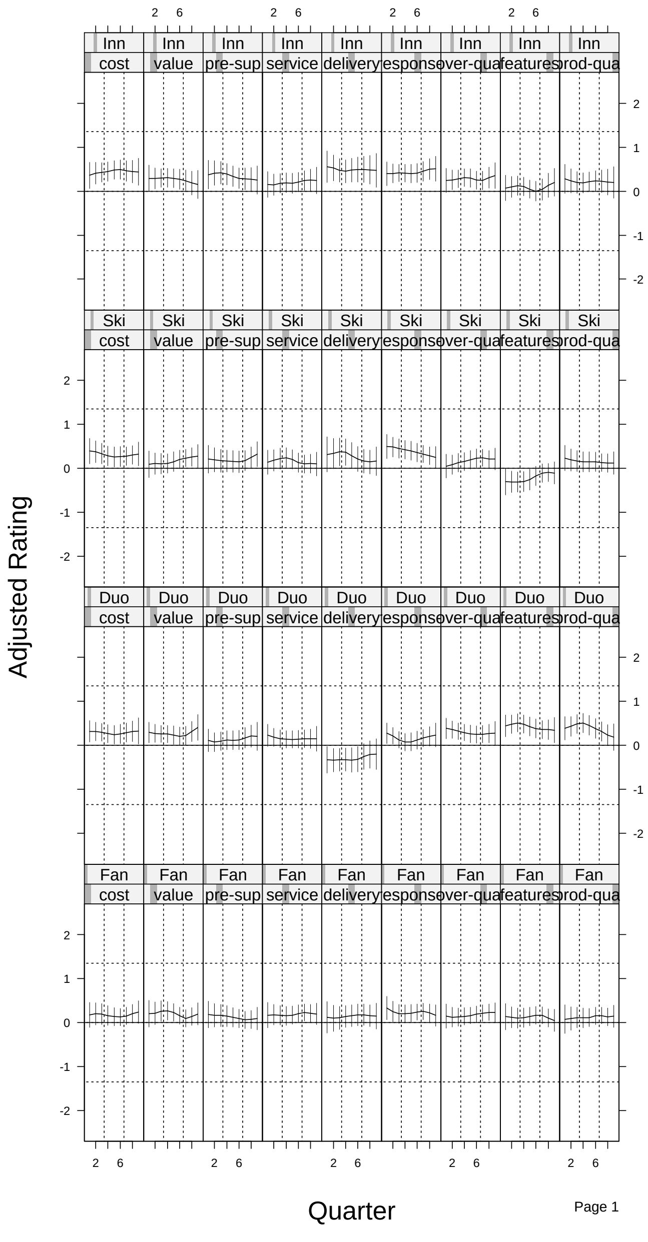

FIGURE 12. Posterior Distribution: Adjusted Quarterly Rating Given Attribute and Company

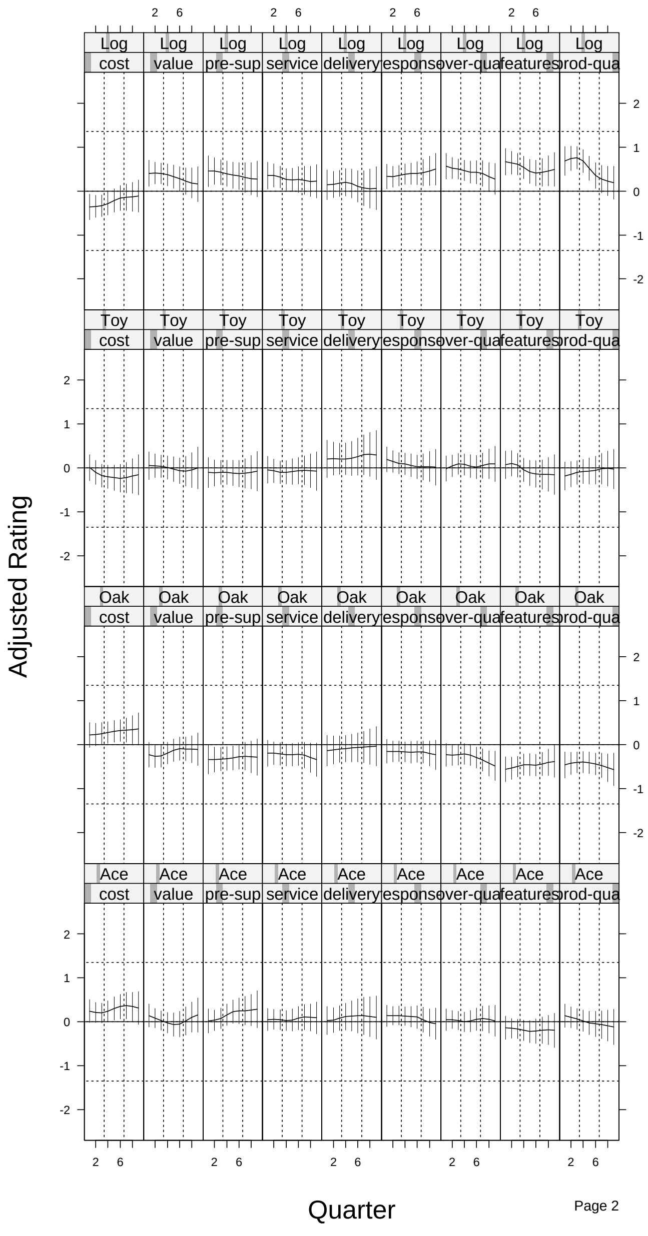

FIGURE 13. Posterior Distribution: Adjusted Quarterly Rating Given Attribute and Company

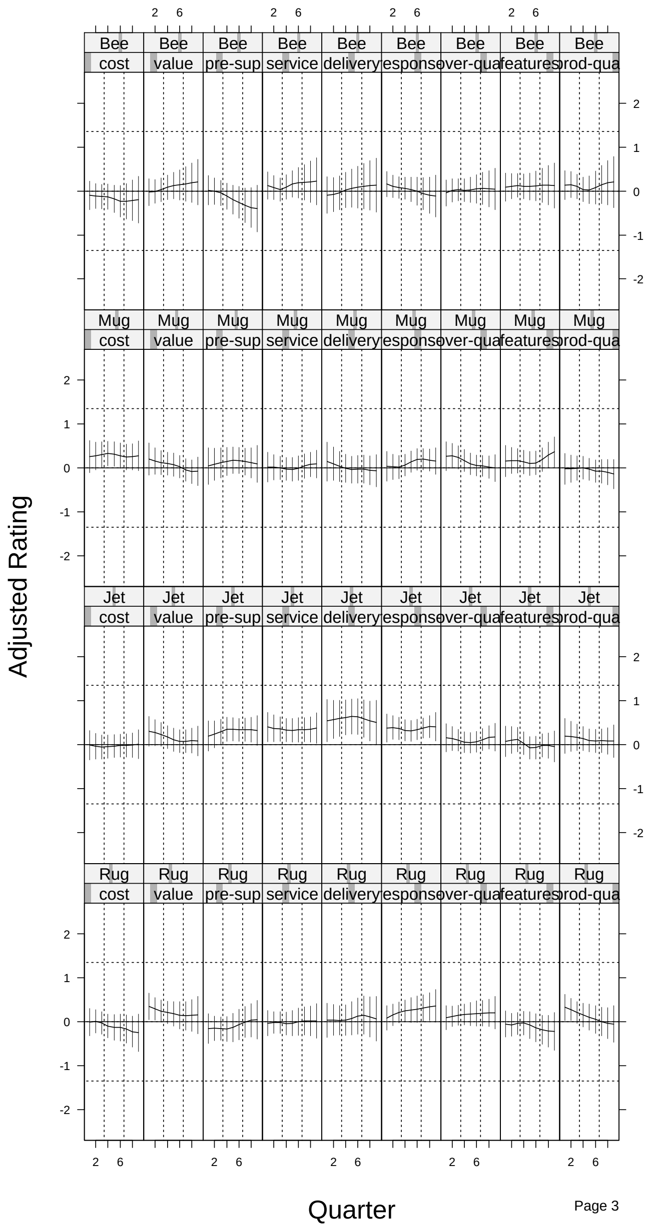

FIGURE 14. Posterior Distribution: Adjusted Quarterly Rating Given Attribute and Company

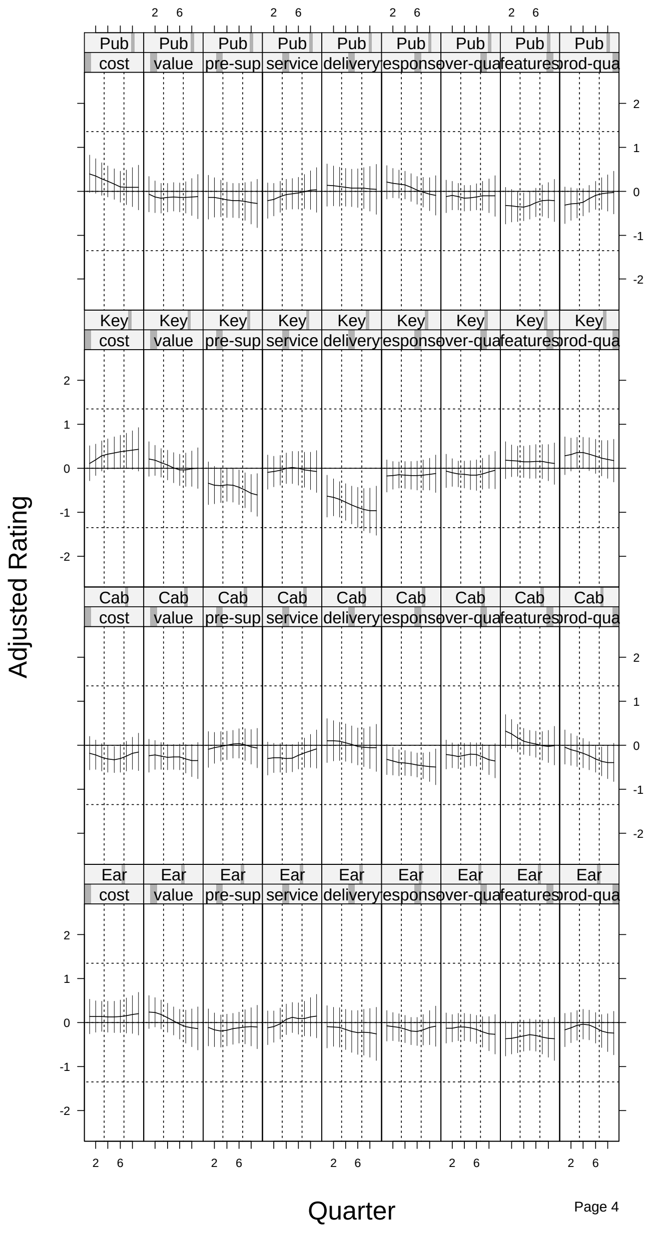

FIGURE 15. Posterior Distribution: Adjusted Quarterly Rating Given Attribute and Company

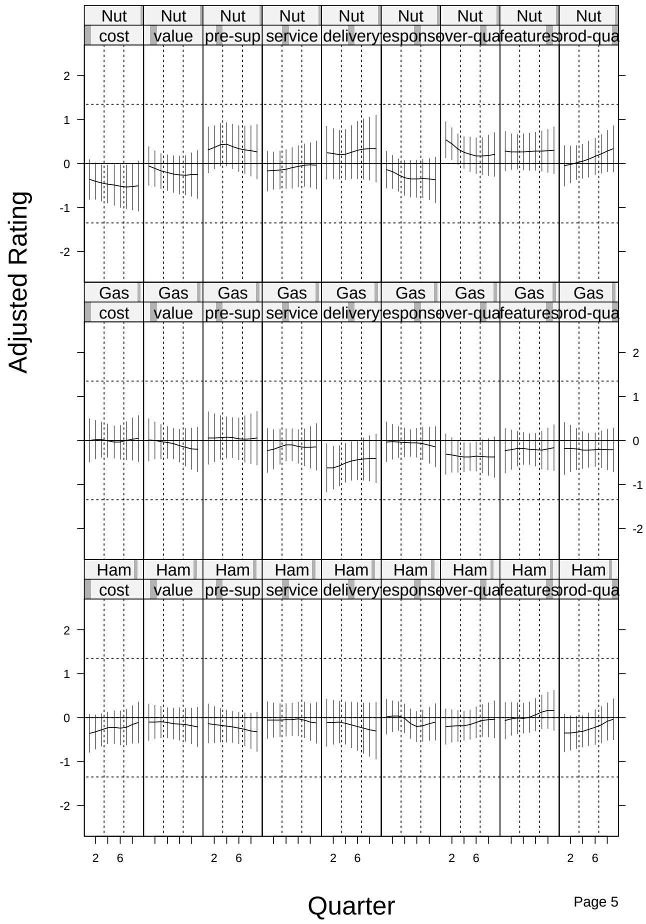

FIGURE 16. Posterior Distribution: Adjusted Quarterly Rating Given Attribute and Company

7.3 CVA Theory: Relationships of Attributes

There is a substantial amount of exogenous information about relationships among the attributes in CVA surveys. The information comes from a knowledge of the nature of the attributes that are rated. It comes from an experience base of interacting with customers and carrying out *discovery*, a process of asking customers what is important and how attributes relate to one another. It comes from surveys other than ours.

The exogenous knowledge has evolved into a CVA theory. Here is one key piece of the theory. A customer's decision to purchase a product or service is assumed to depend strongly on the customer's perception of the attribute `value`, the worth of the product or service relative to its price. `value` is assumed to depend directly on `over-qual` and `cost`, and then `over-qual` is assumed to depend on the remaining attributes: `prod-qual`, `delivery`, `features`, `pre-sup`, `response`, and `service`.

As stated in earlier sections, we chose not to impose strong model specifications about the relationships of attributes. More specifically, while we made strong assumptions about the behavior of the θ_{asq} across q for fixed a and s, we made mild assumptions about the behavior across a and s for fixed q. For example, we did not build into our model any causal dependence of `value` on `over-qual` and `cost`. There were two reasons for this. First, we were reluctant to add to the complexity of an already complex model. Second, the exogenous information about the relationships are subject to substantial debate, and so, in our primary analysis, we wanted to withdraw from this debate by making no strong assumptions. Our reasons are nicely encapsulated by Gelman, Carlin, Stern, and Rubin (1995, pp. 56-57):

> In almost every real problem, the data analyst will have more information than can be conveniently included in the statistical model. This is an issue with the likelihood as well as the prior distribution. In practice, there is always compromise for a number of reasons: to describe the model more conveniently; because it may be difficult to express knowledge accurately in probabilistic form; to simplify computations; or perhaps to avoid using a possibly unreliable source of information.

Instead of building strong assumptions about attribute relationships into our model at the outset, we decided to employ *post-posterior* analysis (Cleveland and Liu, 1998a). To make our analysis fathomable — specifically, to bring in our exogenous knowledge about relationships in a reliable way — we focused just on the κ_{as}. We studied their posterior distribution. We combined the posterior information with reasonable specifications about relationships based on our exogenous information to form a post-posterior distribution of the κ_{ac}. This amounted to a model building exercise using the posterior information and exogenous information.

Of course, another approach would have been to return to the full, original model, and add the exogenous information. We judge this to be a task that is too complex given our current state of knowledge, and therefore unreliable. Our more conservative approach of narrowing our analysis of attribute relationships appears to us more prudent. By dealing just with the κ_{as} we are able to more reliably specify exogenous information and to more clearly separate the influence of the data (in the form of the posterior distribution) and the influence of the exogenous information in drawing inferences about the attribute relationships. But we do hope as our experience base increases that a comprehensive model can be developed.

Acknowledgments

We are grateful to Eric Bradlow for a number of discussions about Bayesian models and customer survey data, and for a number of very helpful comments on the paper; to Don Rubin for discussions that at the beginning of our project helped us to chart our course; to Nick Fisher for astute comments about what material to include in the paper; to Brad Carlin for overseeing our participation in the Workshop and providing important comments on the paper; and to the other organizers of the Workshop for carrying on this fine tradition of a conference on applications.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association*, **88**, 669–680.

Anscombe, F. J. and Tukey, J. W. (1961). The examination and analysis of residuals, *Technometrics* **5**, 141–160.

Becker, R. A. and Cleveland, W. S. (1996). *Trellis Graphics User's Manual*, MathSoft, Seattle. Internet: b&w or color postscript (224 pages) available from site cm.bell-labs.com/stat/project/trellis.

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996). The Design and Control of Trellis Display, *Journal of Computational and Statistical Graphics* **5**, 123–155. Internet: b&w or color postscript (36 pages) available from site cm.bell-labs.com/stat/project/trellis.

Box, G. E. P. and Hunter, W. G. (1965). The Experimental Study of Physical Mechanisms, *Technometrics* **7**, 23–42.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis Forecasting and Control*, 2nd ed., Holden-Day, Oakland, CA.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness, *Journal of the Royal Statistical Society A* **143**, 383–430.

Bradlow, E. T. and Zaslavsky (1997). A Hierarchical Latent Variable Model for Ordinal Data from a Customer Satisfaction Survey with "No Answer" Responses. Technical Report, Department of Marketing, Wharton School, University of Pennsylvania.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, books@hobart.com.

Cleveland, W. S., Denby, L., and Liu, C. (1998). Models and Model Building Tools for Case Data, Bell Labs Technical Report: cm.bell-labs.com/stat/doc/case.ps.

Cleveland, W. S. and and Liu, C. (1998a). An Approach to Model Building Based on a Bayesian Theory of Data Exploration, in preparation.

Cleveland, W. S. and and Liu, C. (1998b). Integated Sum-Difference Time Series Models, in preparation.

Daniel, C. and Wood, F. S. (1971). *Fitting Equations to Data*, Wiley, New York.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty, *Journal of the Royal Statistical Society, B* **57** 45–97.

Dempster, A. P. (1970). Foundation of Statistical Inference, *Proceedings of the Symposium of the Foundations of Statistical Inference, March 31 to April 9*, 56–81.

Draper, D., Hodges, J. S., Mallows, C. L., and Pregibon, D. (1993). Exchangeability and Data Analysis, *Journal of the Royal Statistical Society, A* **156**, Part 1, 9–37.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian Statistical Inference for Psychological Research, *Psychological Review* **70**, 193–242.

Gale, B. T. (1994). *Managing Customer Value*, MacMillan, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., King, G., and Liu, C. (1998). Multiple Imputation for Multiple Surveys (with discussion), *Journal of the American Statistical Association*, to appear.

Gelman, A., Meng, X-L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies, *Statistica Sinica* **6**, 733–807.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and The Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Good, I. J. (1957). Mathematical Tools, *Uncertainty and Business Decisions*, edited by Carter, C. F., Meredith, G. P., and Shackle, G. L. S., 20–36, Liverpool University Press.

Hill, B. M. (1986). Some Subjective Bayesian Considerations in The Selection of Models, *Econometric Reviews* **4**, 191–251.

Hill, B. M. (1990). A Theory of Bayesian Data Analysis, *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geiser, J. S. Hodges, S. J. Press and A. Zellner (Editors), 49–73, Elsevier Science Publishers B. V. (North-Holland).

Johnson, V. E. (1997). An Alternative to Traditional GPA for Evaluating Student Performance (with discussion), Statistical Science, **12**, 251–278.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors, *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. E., Tierney, L., and Kadane, J. B. (1989). Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis, *Biometrika* **76**, 663–674.

Liu, C. (1995). Missing Data Imputation Using the Multivariate t-distribution, *The Journal of Multivariate Analysis* **53**, 139–158.

Liu, C. (1996). Bayesian Robust Multivariate Linear Regression with Incomplete Data, *Journal of the American Statistical Association* **91**, 1219–1227.

Longford, N. T. (1995). *Models for Uncertainty in Educational Testing*, Springer, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics* **21**, 1087–1091.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo Methods, *Journal of the American Statistical Association* **44**, 335–341.

Naumann, E. and Kordupleski, R. (1995). *Customer Value Toolkit*, International Thomson Publishing, London.

Pinheiro, J., Liu, C., and Wu, Y. (1997). Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t-distribution, Technical Report, Bell Labs.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika* **63**, 581–592.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for The Applied Statistician, *The Annals of Statistics* **12** 1151–1172.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Savage, L. J. (1961). *The Subjective Basis of Statistical Practice*, unpublished book manuscript.

Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion), *Journal of the American Statistical Association* **82**, 528–550.

Young, F. W. (1981). Quantitative Analysis of Qualitative Data, *Psychometrika* **46**, 357–388.

Rejoinder

We are grateful to the discussants for their careful attention to the paper. They reviewed our work and made a number of excellent suggestions. To discuss their discussion we will invoke the email-style comment and response, with the discussant in italics.

Best

However, many of the steps taken during the exploratory stage would appeal equally to committed non-Bayesians. This is not so much a criticism of the method, but rather a comment that there is nothing inherently Bayesian about what was done. Yes, we fully agree that those who declare themselves non-Bayesians would be attracted to the methods. This does not make the methods non-Bayesian but rather says that non-Bayesians in word are typically Bayesian in deed when building models. A non-Bayesian who is a competent data analyst combines exogenous information and information from the data to build models.

I do not believe that such exploratory analysis can completely replace the need to examine posterior sensitivity of the model to the assumptions made. Nor do we. Our full analysis of the data included sensitivity analyses, model enlargements, and posterior predictive checking. Our paper did not report on the full scope of the analysis and iterative model building. If we had a record of the complete path it would fill a book. For example, we incorporated explanatory variables not discussed in the paper, but found them to have little influence. There were several complete Bayesian models that were tested and ultimately altered. The model building summarized in the paper and described in full by Cleveland, Denby, and Liu (1998) constitutes a partial demonstration of validity for part of the model.

... no mention is made of how convergence was assessed, nor how many samples were used to estimate the joint posterior distribution. Yes, assessing the convergence of the iterative simulation is very important in Bayesian computation using MCMC methods. Many convergence-diagnostic methods have been proposed in the literature (Cowles and Carlin 1996). Our numerical model building tools such as the many little case regressions (Cleveland, Denby, and Liu 1998) and intermediate computational results for the model such as maximum likelihood estimates gave us guides that could be used to assess convergence of posterior distributions. Of course, we also monitored the posterior means and variances of the parameters of interest by sequences of various lengths. The final results were calculated based on the last 10,000 draws of a single sequence of 12,000 iterations. Running multiple sequences (Gelman and Rubin, 1992) is currently under consideration. The major difficulties include creating over-dispersed start-

ing distributions; we are taking the approach of Liu and Rubin (1996, 1998).

CCDL continue their theme of sequential model building by carrying out a regression and factor analysis based on the posterior distributions estimated above....the data appear to be used twice — once as direct input into the posterior model, and secondly to estimate the posterior parameter distributions which are in turn used as 'data' in the a posterior model. It is best to think of what we did as combining information. Let us break the exogenous information into two sets: A and B. First, we combined A with the data to form the posterior information, more precisely, the combined information in A and the data. Then we combined the information in the posterior with B. In effect, we added more specifications but worked through the posterior to keep it simple. The results need to thought of as very tentative, awaiting a more thorough analysis after we gain more experience from other data.

A graphical representation of CCDL's customer survey model is shown. This is an appealing diagram for conveying the structure of the model.

... ignoring such measurement error in the final Bayesian model may result in over-precise estimates of other model quantities. We recant. Our descriptor of the data as a result of rounding seems less sensible now than simply viewing our model as a continuous approximation of discrete data. The recantation occurred when we realized that the concept of rounding implies that the continuous rating must have a smaller variance than the discrete rating. It would not hurt, of course, to run some simulations to check the effects on our quantities of interest, but various back-of-the-envelope calculations suggest the approximation is excellent for the estimation of company-attribute effects.

An alternative model for the observed customer survey data is to treat the response as an ordered categorical variable, where each category corresponds to an interval with unknown endpoints on some underlying continuous latent scale. This is a good point. In our paper, we transform and then model distributions. If one uses the categorical-interval approach, a transformation of the data is estimated. But we must make a distributional specification of some sort because any monotone transformation of the data obeys such a model as well. To further check our model, we could use our square root transformation, make the distributional assumptions from our model building process, fit the endpoints, and see if they are consistent with our model.

Bradlow and Kalyanam

The other natural consideration is whether one should be analyzing absolute value or relative value. The literature suggests that managerial interest is in the latter measure. It is correct that we need a performance measure that takes the companies' performance measures on each attribute and corrects for an industry average. We do this is Section 7. Figures 10 and 11 are trellis displays of posterior means minus an industry average for each attribute. We could have divided rather than subtracting but we question the meaning of a ratio for a subjective rating scale. The discussants add: *This latter measure can be obtained in the current data by simply dividing each respondent's ratings by the average of the respondents' ratings across all companies. If the relative ratings have a more symmetric distribution, it may also simplify the statistical analysis.* The discussants' divisor would have a quite small variance compared with the variance of a rating; thus the skewness would be altered very little. We believe it makes more sense to model uncorrected data and then compute, at the end, whatever measures might seem sensible.

While we recognize that any choice of model at a hierarchical stage is typically done out of convenience, and is mostly arbitrary, we must always avoid the desire to overfit the data. There are two dangers. One is to overfit. The second is to choose specifications that do not fit the data. If we carefully study the data, mixing the study with reason and exogenous information, we can discipline the inclination to overfit. But if in the name of protecting ourselves from overfitting we avoid model building, we run great risks of choosing specifications that do not fit, that is, specifications that are manifestly disproved by the data. Our observation is that the latter is the serious problem in the practice of statistics, both among declared frequentists and declared Bayesians.

The complex modeling approach employed by CCDL leads to a nice summary of the data and an approach to derive inferences based on posterior distributions of model parameters. The underlying question that remains is: which of these inferences could I have learned though simple ordering of means, and looking at relative sizes of variances. The model provides a full description of the structure and variability in the data. It becomes a fundamental tool that can then be used for many purposes, not just those that can be satisfied by looking at means and rough estimates of variability. For example, we can use it to redesign our survey instrument or to design a new one; in this case, the description of the person effects is critical.

In Bradlow (1994) and Bradlow and Zaslavsky (1997) they derive a nonignorable model for the missing data process for customer satisfaction survey data. This work is quite interesting. While 20% of our data is missing, only 9% is missing not by design (randomly). Some of our checking of the Bayesian model (not reported in the paper) showed that this random missingness was informative, but the magnitude of the effect was quite small.

Rossi

I would encourage them to correlate the satisfaction data with market performance. Within corporations there is enormous pressure already to link company process data with customer ratings of attributes from surveys such as ours. The goal is to find important drivers: company actions that can have major influences on customer opinion. The company processes and customer perceptions are governed by complex causal links. It is not possible to undergo extensive experimentation, so one must rely for the most part on the natural variation that occurs across companies and across time. But this natural variation cannot necessarily support the isolation of the causal links. There is a tendency to take at face value results from regression analyses and factor analyses. In some cases these analyses are carried out without accounting for case (person) effects, which invalidates the results. The case effects, when ignored, distort the estimation, quite substantially if the variances of the case distributions are large. Often, an outcome of the distortion, one that results from the case location effects, is positive, significant estimates of parameters that are in theory positive. This creates a false sense of well being (Cleveland, Denby, and Liu 1998) even though the isolation of causal links can be utter nonsense.

It would be interesting to investigate whether or not the tighter posterior intervals [tighter than intervals for sample means] are a result of the prior or as the result of more efficient use of the data in making inferences about company-attribute means. In Section 7 the nine-quarter average performance for supplier company s on attribute a,

$$\kappa_{as} = \frac{\sum_q \kappa_{asq}}{9},$$

is discussed and posterior means and 95% posterior intervals are plotted. Let $\gamma(\kappa_{as})$ be the posterior standard deviation of κ_{as} and let $\gamma(\bar{r}_{as})$ be the standard deviation of the sample mean of r_{ap} across all p for supplier company s and attribute a, that is, the square root of the posterior variance of $\bar{r}_{as}$. Figure 17 graphs $\log_2(\gamma(\bar{r}_{as})) - \log_2(\gamma(\kappa_{as}))$ against $\log_2(n_{as})$ where n_{as} is the number of ratings of attribute a for company s. We see that the differences decrease substantially with $\log_2(n_{as})$. This suggests to us that the shrinkage resulting from the prior distributions accounts for the reduction. Were we to study the quarterly means in a similar way we would find a considerably bigger difference due to both shrinkage and the smoothness imposed by the time series specification for the quarterly effects.

References for the Rejoinder

Cleveland, W. S., Denby, L., and Liu, C. (1998). Models and Model Building Tools for Case Data, cm.bell-labs.com/stat/doc/case.ps, Bell Labs Technical Report.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *J. Amer. Statist. Assoc.*, **91**, 883-904.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statist. Sci.*, **7**, 457-511.

Liu, C. and Rubin, D. B. (1996). Markov-normal analysis of iterative simulations before their convergence, *J. Econometric*, **75**, 69-78.

Liu, C. and Rubin, D. B. (1998). Markov-Normal analysis of iterative simulations before their convergence: reconsideration and application, Technical Report, Bell-Labs, Lucent Technologies and Department of Statistics, Harvard Univ.

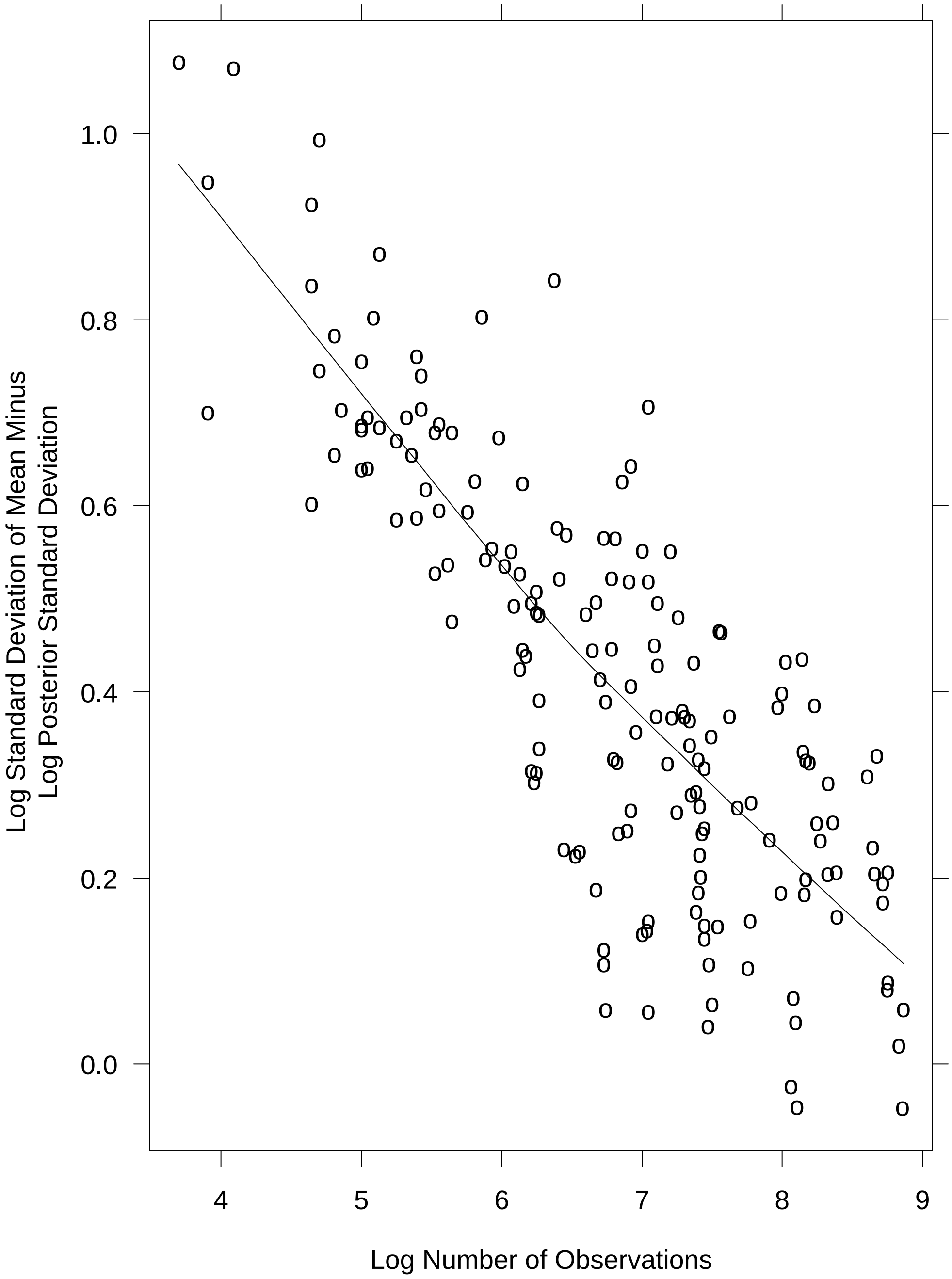

FIGURE 17. Log Base 2 Standard Deviation of Posterior Means Minus Log Base 2 Standard Deviation of Sample Means vs. Log Base 2 Sample Size