Market Lett
DOI 10.1007/s11002-007-9026-x

Hierarchical Bayesian conjoint models incorporating measurement uncertainty

John C. Liechty · Duncan K. H. Fong · Eelko K. R. E. Huizingh · Arnaud De Bruyn

Received: 16 June 2006 / Accepted: 9 October 2007

Abstract The authors explore situations where consumers supplement their judgments with a measurement of uncertainty about their own preferences, either implicitly or explicitly, and develop two sets of hierarchical Bayesian conjoint models incorporating such measurements. The first set of models uses the relative location of a rating to determine the importance or weight given to the rating, in a regression setting. The second set uses interval judgment as a dependent variable in a regression setting. After specifying the models, the authors perform a theoretical comparison with a basic Bayesian regression model. They show that, under different conditions, the proposed models will yield more precise individual-level partworth estimates. Two simulated data examples and data from a conjoint study are used to illustrate the gains that could be obtained from modeling uncertainty. In the empirical application, the authors show that model fit improves when ratings for items that respondents do *not* like are given more weight compared to ratings for items that they do like.

Keywords Conjoint analysis · Weighted regression models · Measurement uncertainty · Confidence · Interval data · Hierarchical Bayesian models

Electronic Supplementary Material The online version of this article (doi:10.1007/s11002-007-9026-x) contains supplementary material, which is available to authorized users.

J. C. Liechty (✉) · D. K. H. Fong
The Pennsylvania State University, 409 BB, University Park, PA 16803, USA
e-mail: jcl12@psu.edu

E. K. R. E. Huizingh
University of Groningen, Groningen, The Netherlands

A. De Bruyn
ESSEC Business School, Cergy-Pontoise, France

Conjoint analysis is probably the most widely used marketing research method to measure consumer trade-offs between multiattribute products and services. (An extensive review of the method is provided by Green et al. 2001.) No matter how data is collected (ratings, rankings, or choice), conjoint analysis assumes that consumers are capable of assessing and expressing their preferences. In addition we assume that it is possible for consumers to have different degrees of uncertainty about their preference statements. In this paper we explore whether more precise utility estimates can be made by including uncertainty in a hierarchical Bayesian model. The introduction of hierarchical Bayesian models by Allenby et al. (1995) and Lenk et al. (1996) demonstrated the value of Bayesian models with respect to conjoint problems. We follow in the spirit of their approach and propose a Bayesian approach, which can incorporate uncertainty in both measurements and utilities into the analysis.

1 Basic and proposed Bayesian models

When consumers are asked to give a preference judgment for a multiattribute object, the traditional approach for inferring their utility for each component is to use a regression model and call the estimated slope parameters the partworth utilities for each associated feature. Stated formally, let

$$y_{ip} = \beta_i^T x_{ip} + \widetilde{\varepsilon}_{ip}, \tag{1}$$

where y_{ip} represents the stated preference for the p^{th} ($p=1,..., n_i$) object seen by the i^{th} ($i=1,..., n_T$) consumer for which the attributes of the object are given by x_{ip}, β_i is the vector of partworth utilities or the utility associated with each attribute, and $\widetilde{\varepsilon}_{ip}$ is an error term that reflects the uncertainty that the i^{th} consumer has regarding his or her understanding and ability to state his or her preference for the p^{th} object accurately.

In our basic hierarchical Bayesian model specification, we assume that

$$\beta_i =_d N\left(\overline{\beta}, \Lambda\right) \text{ and } \widetilde{\varepsilon}_{ip} =_d N\left(0, \widetilde{\sigma}_i^2\right),^1 \tag{2}$$

where $=_d$ means equal in density and N represents a normal density. Then, unless stated otherwise, we assume conjugate prior densities with appropriate prespecified hyperparameters for $\overline{\beta}$, Λ, and $\widetilde{\sigma}_i^2$, or

$$\overline{\beta} =_d N\left(\overline{\overline{\beta}}, \overline{\Lambda}\right), \Lambda^{-1} =_d W\left(n_{\Lambda p}, P_\Lambda\right), \text{and } \widetilde{\sigma}_i^2 =_d IG(shape, scale), \tag{3}$$

where W denotes the Wishart distribution and IG represents an inverse Gamma density. We use vague priors and, in the case of β; we set $\overline{\overline{\beta}} = 0$ to simplify the

[1]We make the assumption that essentially all of the relevant probability mass for each y_{ip} is within a fixed interval (e.g., between 1 and 100). This assumption also holds for the scores given by the two different interval models. Clearly, there may be settings in which this assumption is not valid; in these cases, it is fairly straightforward to replace the normal density for the error term with a truncated normal density. The resulting model requires more sophisticated sampling strategies for the Markov chain Monte Carlo (MCMC) algorithms. One sampling strategy that would work well for relaxing this error assumption is the slice sampler, which we discuss in the Appendix.

notation. This choice is not restrictive because the corresponding variances of β are assumed to be large. The basic hierarchical Bayesian formulation (Eqs. 1, 2, and 3) provides a framework for modeling heterogeneity in the partworths (for a detailed discussion, see Lenk et al. 1996).

1.1 Parabolic weighted regression models

In this paper we assume that the consumers' preference judgments are supplemented with a measurement of uncertainty. Our first approach is to use a weighted regression model, where the weight attached is a function of the corresponding uncertainty measurement, or

$$y_{ip} = \beta_i^T x_{ip} + w_{ip}\varepsilon_{ip}, \tag{4}$$

where w_{ip} modifies the ith consumer's uncertainty about his or her stated preference for the pth object. Note that a smaller weight implies less error for the corresponding judgment. The weighted regression model allows judgments associated with a low level of uncertainty to be more "informative" than judgments associated with a high level of uncertainty.[2] Weighted regression models have been used in the statistical literature to allow for the different variances for each observation (e.g., Draper and Smith 1981). If a weight is available, we analyze data from a weighted regression model by rewriting Eq. 4 as follows:

$$\widetilde{y}_{ip} = \beta_i^T \widetilde{x}_{ip} + \varepsilon_{ip}, \tag{5}$$

where $\widetilde{y}_{ip} = y_{ip}/w_{ip}$ and $\widetilde{x}_{ip} = \left(1/w_{ip}\right)x_{ip}$.

We model the weight for each object as a function of a basic uncertainty measurement (denoted by δ_{ip}). Three commonly used uncertainty measurements in the literature are: (1) a confidence measurement accompanying a single preference judgment (e.g., Laroche et al. 1996), (2) a response latency measurement (e.g., Bassili and Fletcher 1991), and (3) the location of a single preference judgment (e.g., Raden 1985). For the confidence measurement approach, consumers report their preference judgment y_{ip} and a confidence score c_{ip}, which becomes the basic measurement of uncertainty, or $\delta_{ip}=c_{ip}$. For the response latency approach, the length of time between when an object is presented for judgment and when it is judged, T_{ip}, is used as the basic measurement of uncertainty, or $\delta_{ip}=T_{ip}$. Note that it may be appropriate to use a ratio of actual response time to average response time, instead of the response time, to adjust for any systematic trends in the response time data (Haaijer et al. 2000). For the location approach, the location of a single preference judgment with respect to the range of all judgments made by the consumer is used as the basic measurement of uncertainty, or $\delta_{ip} = \frac{y_{ip}-\ell_i}{u_i-\ell_i}$, where $u_i = \max_p\{y_{ip}\}$ and $\ell_i = \min_p\{y_{ip}\}$. (Note, similar to a data dependent prior, we use the actual

[2]One limitation of our modeling approaches is that the models cannot distinguish between uncertainty in the judgment scores and uncertainty in the utilities; stated differently, the covariance matrix of β_i does not depend on the uncertainty measurements. Extending the models so that Λ is a function of uncertainty could distinguish between uncertainty in preference statements and uncertainty about the utility vector, which represents a fruitful area for further research.

observations to compute the basic measurement of uncertainty trying to obtain more precise parameter estimates.)

For the confidence score, we anticipate that higher values will represent a judgment where a consumer is more certain and for which the error variance should be small. For the response latency measurement, we anticipate that lower values represent a judgment where the consumer is more certain and the variance should be small. A parabolic weight function, $w(\delta)$, of the respective basic measurement of uncertainty can accommodate all these weighting schemes, and we assume that the weight used in the weighted regression is given by

$$w_{ip} = \max\left(w_i\left(\delta_{ip}\right), 0\right) \text{and } w_i\left(\delta_{ip}\right) = a_{0i} + a_{1i}\delta_{ip} + a_{2i}\delta_{ip}^2, \tag{6}$$

where a_{0i}=1 to ensure that the model is likelihood identified and

$$a_{ji} =_d N\left(\overline{a}_j, v_j^2\right); \text{ for } j = 1,\ 2, \tag{7}$$

We also assume conjugate prior densities for $\overline{a}_j$ and v_j^2 with the appropriate prespecified hyperparameters, or

$$\overline{a}_j =_d N\left(\overline{\overline{a}}_j, \tau_j^2\right),\ v_j^2 =_d IG\left(shape_j,\ scale_j\right). \tag{8}$$

Again, we set $\overline{\overline{a}}_j = 0$ to simplify the notation, which is possible because we use vague priors. Note that if we set $v_j^2 = 0$, for j=1,2 and $\tau_i^2 = 0$, for all i, then w_{ip}= $w_i(\delta_{ip})$=1, and the model simplifies to the basic hierarchical Bayesian model. The full conditional distributions for this model are given in a Technical Appendix, which is available upon request from the authors.

1.2 Interval models

Classical ratings based conjoint models use point estimates of judgments, e.g., scores on a 1–100 scale. An alternative method of uncertainty measurement is asking consumers for an interval on a scale, where the range of the interval represents their uncertainty (the larger the range, the larger the uncertainty). When interval judgments are available, we develop corresponding Bayesian models for those data. We present two interval models, one for upper and lower interval judgments $\left(y_{\ell ip}, y_{uip}\right)$ and one for symmetric, midpoint interval judgments $\left(y_{\ell ip}, y_{mip}, y_{uip}\right)$, where $y_{uip} - y_{mip} = y_{mip} - y_{\ell ip}$. These intervals can be viewed as imprecise statements of a pair of quantiles from the distribution of a consumer's true preference. Exactly which quantiles a consumer states depends on how he or she has internalized the idea that he or she is "very certain" that the interval contains the true preference. (In this modeling approach, we do not claim to be able to nor do we need to be able to make inferences about the quantiles used by a consumer.)

For upper and lower judgments $\left(y_{\ell ip}, y_{uip}\right)$ we assume the following likelihood:

$$\begin{aligned} y_{uip} &= \beta_i^T x_{ip} + \alpha_i + \xi_{uip} \\ y_{\ell ip} &= \beta_i^T x_{ip} - \alpha_i + \xi_{\ell ip}, \end{aligned} \tag{9}$$

where $\xi_{ip} = \begin{pmatrix} \xi_{uip} \\ \xi_{\ell ip} \end{pmatrix} =_d N(0, S_i)$, S_i is a 2×2 covariance matrix, and α_i represents

the preference uncertainty. Furthermore, the following prior distributions are considered:

$$\begin{aligned} S_i^{-1} &=_d W\left(n_{pr}, P\right),\ \alpha_i =_d N\left(\alpha, \tau_\alpha^2\right) I\{\alpha_i > 0\}, \\ \alpha &=_d N\left(0, \tau_\alpha^2\right),\ \tau_\alpha^2 =_d IG\left(Shape_{\tau_\alpha^2}, Scale_{\tau_\alpha^2}\right). \end{aligned} \tag{10}$$

For midpoint interval data $\left(y_{\ell ip}, y_{mip}\right)$, we assume the following likelihood:

$$\begin{aligned} y_{mip} &= \beta_i^T x_{ip} + \xi_{mip} \\ y_{\ell ip} &= \beta_i^T x_{ip} - \alpha_i + \xi_{\ell ip}, \end{aligned} \tag{11}$$

where $\xi_{ip} = \begin{pmatrix} \xi_{mip} \\ \xi_{\ell ip} \end{pmatrix} =_d N(0, S_i)$. The same prior distributions as described for upper and lower interval model are assumed. To the best of our knowledge, this is the first time that this class of models has appeared in the literature. The corresponding full conditional distributions are given in a Technical Appendix, which is available upon request from the authors.

1.3 The test–retest model

Another strategy to understand a consumer's level of uncertainty with regard to preference for an object is to have each consumer provide multiple judgments about each object. Although this statistical model has been widely studied in the literature—for example, the test–retest approach has been proposed for conjoint reliability studies (e.g., Bateson et al. 1987; Green and Srinivasan 1978)—we include it in this section for comparison.

The repeated observation model is given as follows:

$$y_{ipt} = u_{ip} + \xi_{ipt} \text{ and } u_{ip} = \beta_i^T x_{ip} + \varepsilon_{ip}, \tag{12}$$

where $\xi_{ipt} =_d N\left(0, s_i^2\right)$, $\varepsilon_{ip} =_d N\left(0, \sigma_i^2\right)$, and the errors are independent. Using conjugate priors for the parameters in Eq. 12, we can derive the full conditional densities for the model parameters, which are given in the Technical Appendix.

2 Theoretical comparison of models

Several observations of interest can be made about the models proposed in the previous section. For example, a careful review of the full conditional density of β_i can demonstrate the relationship between each of the proposed models and the basic hierarchical Bayesian regression model and thus may help determine when each of these competing models will, in theory at least, perform better. For the purposes of this discussion, we first state the full-conditional means to allow the reader to contrast the differences in how the uncertainty impacts these means and then in the Technical Appendix, which is available from the authors upon request, we view performance in terms of the full conditional precision (inverse of the variance–covariance matrix) of β_i.

The full conditional density for β_i of the basic hierarchical Bayesian model is given by,

$$\beta_i|- =_d N\left(B_i^{-1}b_i, B_i^{-1}\right), \tag{13}$$

where $B_i = \frac{1}{\tilde{\sigma}_i^2}\sum_p \left(x_{ip}x_{ip}^T\right) + \Lambda^{-1}$, and $b_i = \frac{1}{\tilde{\sigma}_i^2}\sum_p x_{ip}y_{ip} + \Lambda^{-1}\overline{\beta}$. Thus the full conditional mean is

$$B_i^{-1}b_i = A_i^{-1}\left(\sum_p y_{ip}x_{ip}\right) + B_i^{-1}\Lambda^{-1}\overline{\beta}, \tag{14}$$

where $A_i = \sum_p \left(x_{ip}x_{ip}^T\right) + \tilde{\sigma}_i^2\Lambda^{-1}$. For the upper and lower interval model, the full conditional mean is (cf. Equation (A12) from the Technical Appendix)

$$B_i^{-1}b_i = A_{ui}^{-1}\left(\sum_p \left(\frac{s_{\ell i}^2}{s_{Ti}^2}\left(y_{uip}-\alpha_i\right) + \frac{s_{ui}^2}{s_{Ti}^2}\left(y_{\ell ip}+\alpha_i\right) - \frac{2\rho s_{\ell i}s_{ui}}{s_{Ti}^2}\left(\frac{\left(y_{uip}-\alpha_i\right)+\left(y_{\ell ip}+\alpha_i\right)}{2}\right)\right)x_{ip}\right) + B_i^{-1}\Lambda^{-1}\overline{\beta} \tag{15}$$

and for the midpoint interval model, it is

$$B_i^{-1}b_i = A_{mi}^{-1}\left(\sum_p \left(\frac{s_{\ell i}^2}{s_{Tmi}^2}y_{mip} + \frac{s_{mi}^2}{s_{Tmi}^2}\left(y_{\ell ip}+\alpha_i\right) - \frac{2\rho_i s_{mi}s_{ui}}{s_{Tmi}^2}\left(\frac{y_{mip}+\left(y_{\ell ip}+\alpha_i\right)}{2}\right)\right)x_{ip}\right) + B_i^{-1}\Lambda^{-1}\overline{\beta}, \tag{16}$$

where $A_{ui} = \sum_p \left(x_{ip}x_{ip}^T\right) + \left(\frac{1}{\left(1-\rho_i^2\right)}\left(\frac{1}{s_{ui}^2} + \frac{1}{s_{\ell i}^2} - \frac{2\rho_i}{s_{\ell i}s_{ui}}\right)\right)^{-1}\Lambda^{-1}, s_{Ti}^2 = s_{ui}^2 + s_{\ell i}^2 - 2\rho_i s_{\ell i}s_{ui}$, and A_{mi} and s_{Tmi}^2 are obtained by replacing s_{ui}^2 with s_{mi}^2 in A_{ui} and s_{Ti}^2, respectively.

By comparing Eqs. 15 and 16 with Eq. 14, we find that the observed interval values combine in a natural way that offers a type of estimate of the central preference judgment, or that y_{ip} is replaced by

$$\frac{s_{\ell i}^2}{s_{Ti}^2}\left(y_{uip}-\alpha_i\right) + \frac{s_{ui}^2}{s_{Ti}^2}\left(y_{\ell ip}+\alpha_i\right) - \frac{2\rho_i s_{\ell i}s_{ui}}{s_{Ti}^2}\left(\frac{\left(y_{uip}-\alpha_i\right)+\left(y_{\ell ip}+\alpha_i\right)}{2}\right) \tag{17}$$

for the upper and lower interval model and by

$$\frac{s_{\ell i}^2}{s_{Tmi}^2}y_{mip} + \frac{s_{mi}^2}{s_{Tmi}^2}\left(y_{\ell ip}+\alpha_i\right) - \frac{2\rho_i s_{mi}s_{ui}}{s_{Tmi}^2}\left(\frac{y_{mip}+\left(y_{\ell ip}+\alpha_i\right)}{2}\right) \tag{18}$$

for the midpoint interval model. These estimates of the central judgment are calibrated in terms of the relative size of the variance for each part of the interval. For example, when the upper limit is more uncertain than the lower limit, or $s_{ui}^2 > s_{\ell i}^2$, $\left(y_{uip}-\alpha_i\right)$ will have less of an impact compared with $\left(y_{\ell ip}+\alpha_i\right)$. A similar statement can be made for the midpoint interval model.

For the test–retest model, which can be rewritten as

$$y_{ipt} = \beta_i^T x_{ip} + \varepsilon_{ip} + \xi_{ipt}, \tag{19}$$

the full conditional density of β_i is:

$$\beta_i|- =_d N\left(B_i^{-1}b_i, B_i^{-1}\right). \tag{20}$$

where $B_i = \left(\frac{n_{i0}}{s_i^2+\sigma_i^2 n_{i0}}\right)\sum_p \left(x_{ip}x_{ip}^T\right) + \Lambda^{-1}$, $b_i = \left(\frac{1}{s_i^2+\sigma_i^2 n_{i0}}\right)\sum_{p,t}\left(y_{ipt}x_{ip}\right) + \Lambda^{-1}\overline{\beta}$ and n_{io} is the assumed common value of n_{ip}. Thus y_{ip} is replaced by $\sum_t y_{ipt}/n_{i0}$ here.

The full conditional densities, in Eqs. 13 to 20, can give insights regarding the performance of an interval model, when the interval model is compared with a regression model. For this comparison, we assume that researchers are interested in understanding individual partworths β_i, that a high precision (small variance) of β_i is better than a low precision, and that the variance terms for both models are equal or $s_{ui} = s_{\ell i} = s_{mi} = \widetilde{\sigma}_i = s_i$. Under these assumptions, it is easy to show that the interval models will have higher precision than the basic regression model as long as the correlation between the upper and lower interval ratings are less than 1; see the Technical Appendix for details. In addition, the precision increases to infinity as the correlation goes towards −1.

The intuition behind this result is straightforward; if the correlation is high, then each pair of interval ratings will tend to be either above or below their average values. This makes it harder to identify the center of the actual interval and hence the actual partworths, as the observed pairs keep forming intervals that tend to be offset from the true interval, even though the average center of these intervals is at the center of the actual interval. Alternatively, if the correlation is negative, each pair of interval ratings tends to be centered at the center of the true interval, although the width of these intervals can vary dramatically; this happens because, when one rating tends to be above its average value, the other one tends to be below its average value.

This comparison can be extended to the test-retest regression model—the point at which the interval model outperforms the test–retest regression model is dependent on the specific assumptions of the variance parameters; again for a more detailed discussion see the Technical Appendix.

3 Simulated data examples

We investigate the properties of the weighted regression model and the upper and lower interval model, via two different simulation studies. In the first study we generate weighted regression data, where we simulate a ‘measurement of uncertainty’ or a δ and then use a parabolic map to convert δ into a standard deviation which is then used to generate a synthetic response, based on individual specific part-worth and random levels of covariates (see Fig. 1 for the average of the parabolic map between δ and the standard deviation.) We have four conditions driven by two factors, low and high information levels and low and high δ variation. In the high information condition, we have 200 individuals and 25 repetitions per individual, and in the low information condition, we have 50 individuals and 10 repetitions per individual. In the low δ variation condition most of the levels of δ are

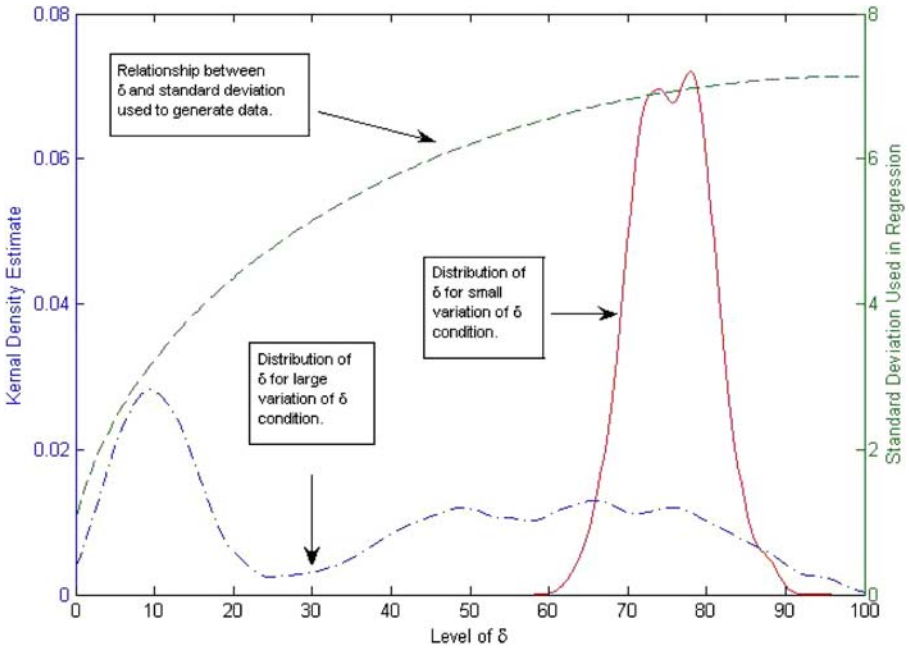

Fig. 1 Summary of δ dispersion and a parabolic mapping between δ and standard deviation used to generate the simulated data. The *solid line* summarizes the kernel density estimate of the simulated values of δ in the "small variance" condition, which uses a narrow distribution of δ. (A Kernel Density Estimates can be viewed as smoothed histograms obtained from the values of δ that were generated for the simulations.) The *dash-dot line* summarizes the kernel density estimate of the simulated values of δ in the "large variance" condition, which uses a wide, multi-modal distribution of δ. The *dashed line* gives the relationship between the δ and their associated variances, with high scores having the highest variance or dispersion, hence being less informative. The *vertical axis* on the *left side* of the graph gives the values for the two kernel density estimates and the *vertical axis* on the *right side* of the graph gives the values that correspond to the relationship between δ and the standard deviation

centered around 75 and in the high variation condition the levels of δ are distributed between 0 and 100 (also, see Fig. 1 for a graphical summary of the low and high δ variation conditions.) As a result the low δ variation condition results in fairly similar variance, while the high δ variations condition results in a wide range of variances for each individual. In the second study we generate upper and lower interval data and here we have six conditions driven by two factors, low and high information levels (which are the same as for the weighted regression study) and three different levels of correlation (ρ=−0.7, 0 and 0.7) between the upper and lower ratings.

For both simulation studies, we contrasted the performance of the appropriate model (i.e. weighted regression or interval model) with the standard hierarchical regression model and we measured performance in terms of the ability to recover the true part-worth parameter values. This was accomplished by finding the root mean square error (RMSE) and mean absolute deviation (MAD) between the true parameters and the posterior mean from the resulting Monte Carlo Markov chain (MCMC) analysis. For the interval data, we analyzed the upper and lower ratings

Table 1 Results for several simulation studies which compare the ability to recover part-worth parameter values: standard regression model versus competing models

Simulation studies				Parameter RMSE	Parameter MAD
Weighted regression data	High information	Large δ variation	Regression model	0.427	0.292
			Weighted regression model	0.384	0.255
		Small δ variation	Regression model	0.500	0.345
			Weighted regression model	0.501	0.345
	Low information	Large δ variation	Regression model	0.655	0.474
			Weighted regression model	0.576	0.418
		Small δ variation	Regression model	0.645	0.465
			Weighted regression model	0.641	0.464
Interval data	High information	ρ=0.7	Regression model	0.258	0.154
			Interval model	0.222	0.139
		ρ=0	Regression model	0.267	0.157
			Interval model	0.167	0.104
		ρ=−0.7	Regression model	0.326	0.179
			Interval model	0.098	0.059
	Low information	ρ=0.7	Regression model	0.444	0.267
			Interval model	0.418	0.262
		ρ=0	Regression model	0.488	0.282
			Interval model	0.358	0.224
		ρ=−0.7	Regression model	0.447	0.267
			Interval model	0.202	0.126

Part-worth parameter recovery is measured by RMSE and MAD. Weighted regression model vs. standard regression model on synthetic weighted regression data. Interval model vs. standard regression model on synthetic interval data. High information is 200 individuals with 25 repetitions each and low information is 50 individuals with 10 repetitions each

separately, each with their own standard regression analysis and then calculated the RMSE and MAD using both sets of posterior means and true parameter values. The results, as reported in Table 1, are as we expected. In general as the amount of information decreases, the ability to recover the part-worth parameters becomes worse. In addition, using the correct model results in an improved part-worth estimate when compared to the standard hierarchical regression model. (Notice one exception with weighted regression data: for the case of high information condition and small δ variation, the performance is equivalent.) For the weighted regression study, as δ becomes more dispersed (and hence the standard deviation becomes more varied) the weighted regression model does even better than when δ is not as dispersed. This is not surprising, as the standard regression model tends to overweight observations with a large δ and underweight observations with a small δ, resulting in an inappropriate weighting of random errors which leads to an increase in estimation bias.

Not only is the weighted regression model better at recovering the true part worth parameter values, it is better at estimating the level of uncertainty with respect to

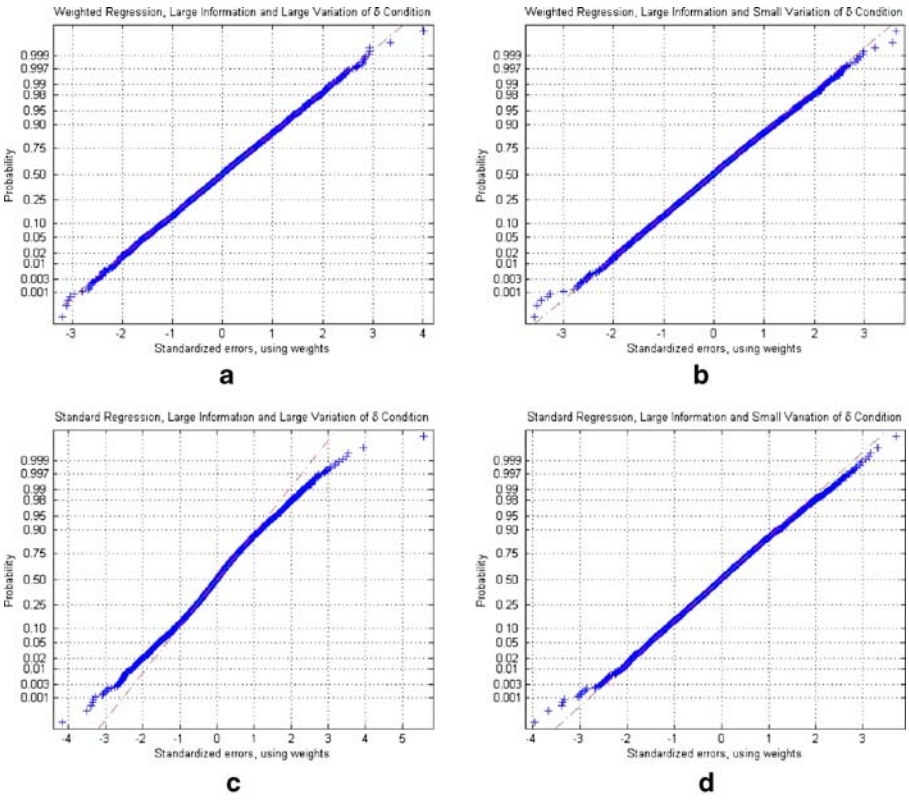

Fig. 2 Normal probability plots of the standardized residuals. The residuals are standardized using the estimated weights in the Weighted Regression models (**a**, **b**) and using a constant variance in the Standard Regression models (**c**, **d**). These *graphs* are for the high information case of the weighted regression simulation study. **a** Weighted Regression, High Variance. **b** Weighted Regression, Small Variance. **c** Standard Regression, High Variance. **d** Standard Regression, Small Variance.

each observation. This can be seen by calculating the *z*-score for each error, using the parameter estimates and the weighted standard deviation. If the model can correctly recover the varying standard deviations, then the resulting *z*-scores should act as if they came from a standard normal distribution. By comparing the normal probability plots for *z*-scores from the weighted regression analysis and the standard regression analysis (cf. Fig. 2), it is clear that the standard regression analysis systematically gets the wrong standard deviation for a large set of data points and this systematic flaw is more pronounced as the variance of δ increases.

Perhaps the most striking feature in the recovery of the parameters is with respect to the interval model simulation. For each of the different levels of correlation, the standard regression model performs roughly the same based on the level of information; however as we predicted in the theoretical comparison (see Section 2 and the Technical Appendix), as the correlation decreases, the ability of the interval model to estimate the part-worth parameters increases. This decrease can be seen by comparing the parameter RMSE for the regression model and interval model as reported in Table 1.

4 Conjoint study

To gain some empirical understanding of the performance of the proposed models, we collected data in an experimental setting. During the study, we randomly divided respondents (undergraduate and graduate students from a large northeastern US university) into several groups according to the way their preference was measured (test–retest, upper–lower interval, and midpoint interval) and exposed each group to one measurement method. We did not find any significant sociodemographic differences among groups. In both studies, respondents rated a set of web pages that represented the customized front page of a university news site. We administered the survey in computer labs using a computer program specifically developed for this experiment. In the final design, the web pages included five attributes: two with three levels and three with two levels. For a description of the attributes and their corresponding levels, see Table 2.

The test–retest group completed a self-explicated task and a dummy task (a demographic survey) and then rated a set of 12 web pages. Next, they completed a second dummy task (a university trivia quiz) and rated the same set of 12 web pages. The upper–lower and midpoint groups performed a similar set of tasks except that they used an appropriate interval tool to make their judgments and only rated the set of 12 web pages once. A fractional factorial design was used to create the different sets of web pages.

A total of 115, 128, and 130 subjects completed the study using the test–retest method, the upper–lower interval method, and the midpoint interval method, respectively. Each data set was analyzed using the MCMC algorithm and assuming vague priors. Convergence diagnostics were calculated to ensure the models exhibited reasonable mixing properties and converged in distribution. We assessed model performance by calculating the log marginal probabilities, which were estimated using a method proposed by Newton and Raftery (1994). The test–retest data were analyzed using both a standard regression model and the weighted regression model in which the weight is a function of the location. The interval data sets were analyzed using their respective interval models.

Clearly, there are challenges with regard to comparing these competing models across all the competing data sets. The easiest and most direct comparisons can be made between the regression and weighted regression models using the test–retest

Table 2 Attributes used to construct the web pages

Attribute	Levels		
Weather forecast	One week ahead general forecast	One-day extensive forecast	
University news	University sports	General university news	
General news	US news (six headlines)	World news (six headlines)	Mixed (three headlines of US and world news)
Business news	General business news (six headlines)	Stock market news (six headlines)	Mixed (three headlines of general and stock news)
Online coupon	$2	$4	

Table 3 Log marginal probabilities for regression and weighted regression (WR) models

Model	Data set	Log marginal probability
Regression	Test–retest	−9,815.87
WR (location)	Test–retest	−9,272.91
Upper–lower	Upper–lower	−10,896.68
Midpoint	Midpoint	−10,630.16

data. Using the log marginal probability, this comparison shows that the weighted regression model fits the data better than the standard regression model (Table 3).

Comparisons among all of the models is challenging given that we have three different sets of data, two of which are bi-variate observations intended for analysis using interval models. Each data sets has the same number of ratings or scalar data points per participant (two sets of 12 ratings for the test-retest data and 12 pairs of ratings for the interval data), but the size of the data sets (e.g. the number of participants) are different for each data set. Given that each data set has the same number of ratings per participant, one basis of comparison among the four analyses is the log marginal probability per person; on this basis, the weighted regression has the largest score at −80.63 (−9,272.91/115), followed by the midpoint analysis at −81.77(−10,630.16/130) and finally the upper–lower analysis at −85.13 (−10,896.68/128). Although not a definitive basis of comparison, further empirical research (e.g., a modified research design, more elaborate model comparison approaches) could focus on developing a more satisfying comparison across the different models.

For the weighted regression analysis, as we show in Fig. 3, the aggregate weight function ($\overline{w}(\delta) = 1 + \overline{a}_1\delta + \overline{a}_2\delta^2$, with posterior means of $\overline{a}_1$ and $\overline{a}_2$) increases with respect to the location percentage (δ) with a slightly positive rate of increase. The location percentage is defined as (*score*−min)/(max−min), and is equal to 0 (resp. 1) for the minimum (resp. maximum) score given by a respondent during a conjoint task. This scaling is made necessary by the fact that respondents may use different portions of the rating scale to express their preferences.

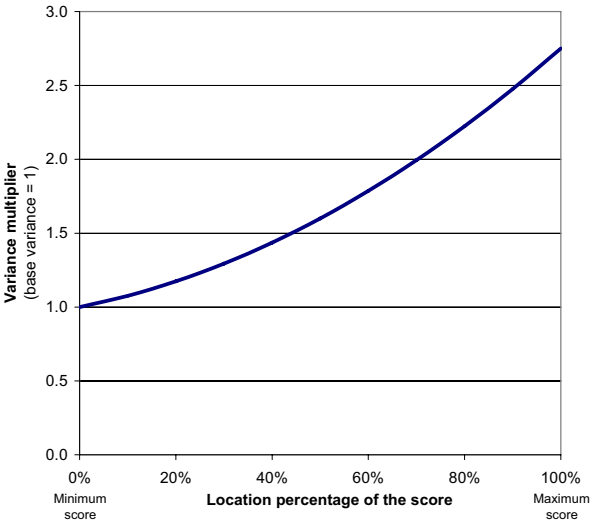

Fig. 3 Variance multiplier (parabolic function) of the location score used as a weight in the weighted regression model. The lowest scores have a lower variance, hence are more predictive

Table 4 Aggregate partworth ($\bar{\beta}$) parameter estimates, posterior mean (Std)

Model	Data set	Intercept	Weather	University sports	Coupon	Mix gen. news	World news	Mix bus. news	Gen bus. news
Regression	Test–retest	63.85 (1.49)	−5.67 (0.94)	−3.29 (1.12)	1.55 (0.40)	7.90 (0.88)	−5.12 (0.97)	7.29 (0.91)	2.99 (0.95)
Weighted regression	Test–retest	61.79 (1.17)	−5.82 (0.67)	−3.72 (0.47)	1.50 (0.22)	7.21 (0.63)	−5.51 (1.27)	7.21 (0.90)	2.87 (0.85)
Upper–lower interval	Upper–lower	57.71 (2.36)	−5.16 (1.08)	−0.64 (0.70)	1.12 (0.32)	4.61 (0.94)	−1.87 (0.70)	4.72 (0.75)	2.04 (0.67)
Midpoint interval	Midpoint	59.59 (2.18)	−4.25 (1.02)	−1.45 (0.91)	1.68 (0.40)	3.38 (0.50)	−4.21 (0.67)	6.13 (1.00)	4.15 (0.98)

This increase suggests that less weight should be granted to judgments that have a relatively high score; stated differently, subjects' judgments about what they do not like are more informative or perhaps more stable than their judgments about what they do like. This result contrasts with the widely held assumption that the uncertainty–location relationship follows an inverted-U shape. It also illustrates the benefits of using parabolic weighting rather than relying on fold-over intensity to gauge the weight of responses (Raden 1985). More broadly, the results of this study suggest that incorporating measurements of uncertainty can lead to a better understanding of individual utility functions, but we readily acknowledge that these results and models must be investigated in more detail in further research.

The aggregate parameter estimates for the four different modeling approaches are relatively similar, as we show in Table 4.

The main difference between the regression and weighted regression models is in the intercept estimates, which suggests that the baseline web page has a lower overall utility than is estimated by the standard regression model. The interval models give parameter estimates that differ from the regression estimates, which may be because of differences among the groups of participants. Both the interval models have high correlation values between the endpoint errors: the average correlation from S_i, according to the posterior means, is a relatively high 0.996 for the upper–lower interval model and 0.970 for the midpoint interval model. In the Technical Appendix, we argued that the farther these correlations are from 1, the better the partworth estimates are (i.e., they have smaller variances). It appears that this insight translates into better model fit in practice, in that the upper–lower interval has the larger correlation but the midpoint interval model has the larger average log marginal probability (or the better fit). This finding suggests that the upper–lower interval judgments may result in more biased estimates than the weighted regression and midpoint estimates (see the relevant estimates in Table 4).

5 Conclusions

We have proposed several new approaches to incorporate preference uncertainty into a statistical model using a hierarchal Bayesian framework. The first set of proposed models incorporates a parabolic function of uncertainty as a weight in a weighted

regression model. The second set uses interval judgments as the dependent variables. As a basis of comparison, we also consider a standard hierarchical Bayesian regression model and a regression model in which subjects make repeated judgments of the same item.

The contributions of this article include a presentation of the full model specifications for the interval models and the parabolic, weighted regression model. To the best of our knowledge, models such as the interval models have not been discussed in the literature previously. In addition, the weighted regression models that we present build on existing models in the literature by introducing a shrinkage model that provides a means to determine both the individual and the aggregate impact of a covariate on variance.

In addition to introducing a new class of models, we have conducted a theoretical comparison of their performance. Using rather mild assumptions, we demonstrate that models that include preference uncertainty measurements perform better than a standard regression model that does not include such measurements. In addition, we find that the correlation between stated intervals affects the variance of the utility estimates. In particular, as the correlation tends toward −1, the variance of the utility estimates moves toward 0. Finally, by introducing this class of models, we provide a framework for investigating when and how different measurements of uncertainty can lead to a better understanding of individual utilities. If individual subjects truly have different levels of uncertainty about the judgments they make, these models should help improve the accuracy of utility parameter estimates compared with the standard hierarchical Bayesian regression model.

To investigate the performance of this class of models, we give a short report of a simulation study and of an empirical study in which we find clear evidence that the weighted regression model performs better than the standard regression model, as well as some evidence that the interval measurement methods and models may perform better than the standard measurement methods when used in conjunction with the weighted regression model. In addition, the aggregate weight function suggests that more weight should be given to judgments with lower scores compared with judgments with higher scores, which indicates that judgments about what subjects do not like may be more stable and informative than judgments about what they do like. These empirical results offer some guidelines for future empirical investigations. The models that we have introduced, along with their natural extensions, should enable researchers to investigate and gain new insights into the cognitive processes that underlie judgment formation and decision making.

References

Allenby, G., Arora, N., & Ginter, J. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, *32*, 152–162.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A method for Cati and a new look at nonattitudes. *Public Opinion Quarterly*, *55*, 331–346.

Bateson, J. E. G., Reibstein, D. J., & Boulding, W. (1987). Conjoint analysis reliability and validity: A framework for future research. In M. J. Houston (Ed.) *Review of marketing* (pp. 451–481). Chicago: American Marketing Association.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis*. New York: Wiley.
Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, *31*, S56–S73.
Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*, 103–123.
Haaijer, R., Kamakura, W., & Wedel, M. (2000). Response latencies in the analysis of conjoint choice experiments. *Journal of Marketing Research*, *37*, 376–382.
Laroche, M., Kim, C., & Zhou, L. (1996). Brand familiarity and confidence as determinants of purchase intention: An empirical test in a multiple brand context. *Journal of Business Research*, *37*, 115–120.
Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, *15*, 173–191.
Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B, 56*, 3–48.
Raden, D. (1985). Strength-related attitude dimensions. *Social Psychology Quarterly*, *48*, 312–330.