12 Optimization-Based and Machine-Learning Methods for Conjoint Analysis: Estimation and Question Design

Olivier Toubia, Theodoros Evgeniou and John Hauser

12.1 Introduction to optimization and machine-learning conjoint analysis

Soon after the introduction of conjoint analysis into marketing by Green and Rao (1972), Srinivasan and Shocker (1973a, 1973b) introduced a conjoint analysis estimation method, Linmap, based on linear programming. Linmap has been applied successfully in many situations and has proven to be a viable alternative to statistical estimation (Jain, et. al. 1979, Wittink and Cattin 1981). Recent modification to deal with "strict pairs" has improved the estimation accuracy with the result that, on occasion, the modified Linmap predicts holdout data better than statistical estimation based on hierarchical Bayes methods (Srinivasan 1998, Hauser, et. al. 2006).

The last few years have seen a Renaissance of mathematical programming approaches to the design of questions for conjoint analysis and to the estimation of conjoint partworths. These methods have been made possible due to faster computers, web-based questionnaires, and new tools in both mathematical programming and machine learning. Empirical applications and Monte Carlo simulations with these methods show promise. While the development and philosophy of such approaches is nascent, the approaches show tremendous promise for predictive accuracy, efficient question design, and ease of computation.

This chapter provides a unified exposition for the reader interested in exploring these new methods. We focus on six papers: Toubia, Simester, Hauser and Dahan (TSHD), 2003; Toubia, Simester and Hauser (TSH), 2004; Evgeniou, Boussios and Zacharia (EBZ), 2005; Toubia, Hauser and Garcia (THG), 2006; Abernethy, Evgeniou, Toubia and Vert (AETV), 2006; Evgeniou, Pontil and Toubia (EPT), 2006. To avoid redundancy, we refer to each of the six reviewed papers by the initials of their authors after the first mention in each section.

We use a framework that clarifies the strengths and limitations of these methods as applied in today's online environment. Online conjoint analysis is often characterized by a lower number of observations per respondent, noisier data, and impatient respondents who have the power to terminate the questionnaire at any time. Such an environment favors methods that allow adaptive and interactive questionnaires, and that produce partworth estimates that are robust to response error even with few observations per respondent.

The framework is that of statistical machine learning (e.g., Vapnik 1998). Within this framework, we interpret recent attempts to improve robustness to response error and to decrease the number of observations required for estimation as an application of "complexity control." We complement this framework to review recent adaptive question design methods, by including experimental design principles which select questions to minimize the expected uncertainty in the estimates.

In the interest of brevity we focus on the conceptual aspects of the methods, and refer the reader to the published papers for implementation details.

12.1.1 Notation and Definitions

We assume I consumers indexed by i (i=1,,.. I) answering J_i conjoint questions each, indexed by j (j=1,..,J_i). Let w_i denote a p-dimensional partworths vector for each consumer i. For ease of exposition, we assume binary features and a main effects specification. Neither of these assumptions are critical to the theory – the reviewed papers address multi-level features and interactions among features. Indeed, an important benefit of complexity control is that feature interactions of any degree may be estimated in an accurate and computationally efficient manner (EBZ; EPT).

The methods we review can be used for most conjoint data-collection formats. For simplicity we focus on the three most common: full-profile analysis, metric paired comparisons, and stated-choice questions.

For full profile rating conjoint data, we assume that the j^{th} question to respondent i consists in rating a profile, x_{ij}. The respondent's answer by y_{ij}. The underlying model is $y_{ij} = x_{ij}.w_i + \varepsilon_{ij}$, where ε_{ij} is a response error term.

For metric paired-comparison conjoint data, we assume that the j^{th} question asks respondent i to compare two profiles, x_{ij1} and x_{ij2}. We denote the respondent's answer by y_{ij}. The sign of y_{ij} determines which profile the respondent prefers; the magnitude of y_{ij} determines the strength of the preference. The underlying model is hence $y_{ij} = (x_{ij1} - x_{ij2}).w_i + \varepsilon_{ij}$ where ε_{ij} is a response error term.

For stated-preference (choice-based) conjoint data, each respondent is asked to choose among a set of profiles. For ease of exposition, we assume that the j^{th} question asked the respondent to choose among two profiles, x_{ij1} and x_{ij2}. Without loss of generality, we code the data such that profile 1 is the chosen profile. (Binary choice simplifies exposition. Empirical applications and simulations use choices among more than two profiles.). The underlying model is that relative true utility, u_{ij}, is given by $u_{ij} = (x_{ij1} - x_{ij2}).w_i + \varepsilon_{ij}$ where ε_{ij} is a response error term. The respondent chooses profile 1 if $u_{ij} \geq 0$. The distribution of ε_{ij} implies alternative probabilistic models.

12.2 Using complexity control in conjoint analysis

In statistical estimation of partworths, researchers often worry about over-fitting the data. For example, if one were to use regression to estimate almost as many partworths as there are data points, then the conjoint model would fit the (calibration) data well, but we might expect that the partworths would be based, in part, on measurement error and would not be able to predict holdout data. Classical statistics address over-fitting by accounting for degrees of freedom and Bayesian statistics address over-fitting with hyper-parameters and the implied shrinkage toward the population mean. In statistical learning methods, over-fitting is addressed with the concept of complexity control. The conceptual idea is that if the model is too complex, it is too susceptible to over-fitting. To avoid this unwanted effect, we limit the complexity of the model by defining a measure of fit, a measure of complexity, and a method for determining the trade off between fit and complexity. Because the concept is important to understanding the philosophy of the new methods, we begin with a brief review of complexity control.

12.2.1 Ridge regression is an example of complexity control

There is a long history in models of consumer behavior that, in the presence of measurement error, unit partworths often predict well (e.g., Einhorn 1971, Dawes and Corrigan 1974.). One way to incorporate this concept in conjoint analysis is with ridge regression (e.g., Wahba 1990; Vapnik 1998; Hastie et al., 2003). Consider a simple ordinary least square regression resulting from a full-profile conjoint questionnaire. Such estimation involves minimizing the following loss function with respect to wi:

$$L(w_i) = \sum_{j=1}^{J} (y_{ij} - x_{ij}.w_i)^2 \quad (1)$$

Minimizing loss function (1) results in the OLS estimate:

$$\hat{w}_i^{OLS} = (X_i^T.X_i)^{-1}.X_i^T Y_i \quad (2)$$

where X_i and Y_i are obtained by stacking all J observations for consumer i. If the number of profiles J is relatively small compared to the number of parameters to estimate p, this simple approach may suffer from over-fitting and the estimates may be very sensitive to small variations in the dependent variable. Mathematically, this instability comes from the poor conditioning of the matrix ($X_i^T X_i$).

Ridge regression addresses instability and over-fitting by replacing $\hat{w}_i^{OLS}$ with:

$$\hat{w}_i^{Ridge} = (X_i^T.X_i + \gamma.I)^{-1}.X_i^T Y_i \quad (3)$$

where I is the identity matrix and the parameter γ may be selected using various methods, such as cross-validation (which we will review later). Note that the matrix (XiT.Xi + γ.I) is better conditioned than (XiT.Xi): all its eigenvalues are greater than or equal to γ. It is easy to show that (3) is the solution to the following modification of the OLS problem (1), where the minimization is done over wi, given γ:

$$L(w_i \mid \gamma) = \frac{1}{\gamma}\sum_{j=1}^{J}(y_{ij} - x_{ij}.w_i)^2 + \mid w_i \mid^2 \quad (4)$$

where $|w_i|^2$ is the Euclidean norm of the vector w_i.

One interpretation of the term, $|w_i|^2$, is as a means to control the *complexity* of the estimate w_i. Complexity control may be viewed as an exogenous constraint imposed on w_i to effectively limit the set of possible estimates. The parameter γ in (4) dictates the relative weight on complexity versus fit. As $\gamma \rightarrow 0$, Equation 4 becomes equivalent to OLS regression; as $\gamma \rightarrow +\infty$, Equation 4 simply minimizes complexity. If we had an additional constraint that the w_i's sum to a constant, the solution would be equal weights. Typically we observe a U-curve relationship between the parameter γ and holdout accuracy (e.g., Evgeniou, Pontil, Toubia [EPT] 2006). Accuracy is poor when γ is too small because of over-fitting. Similarly, accuracy is often poor when γ is too large because the data are virtually ignored. Bootstrapping methods like cross-validation (reviewed in a later section), for example, offer a practical and effective way of searching for this optimal value of γ , which is an issue extensively studied within statistical learning theory.

12.2.2 A Bayesian Interpretation of complexity control

We can use Bayes Theorem to provide another interpretation of complexity control. We augment the data likelihood with a Bayesian prior as follows:

(5) Likelihood: $y_{ij} = x_{ij}.w_i + \varepsilon_{ij}$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

Prior: $$w_i \sim N(0, \beta.I)$$

We compute the posterior distribution on wi conditioned on the data and a specific value of the parameters β and σ:

(6)

$$P(w_i \mid \{y_{ij}\}, \sigma, \beta) \propto P(\{y_{ij}\} \mid w_i, \sigma, \beta).P(w_i \mid \sigma, \beta)$$

$$\propto \exp\left(\sum_{j=1}^{J} -\frac{(y_{ij} - x_{ij}.w_i)^2}{2\sigma^2}\right).\exp\left(-\frac{\mid w_i \mid^2}{2.\beta^2}\right)$$

$$= \exp\left(-\frac{1}{2.\beta^2}\left[\frac{1}{\sigma^2/\beta^2}\sum_{j=1}^{J}(y_{ij} - x_{ij}.w_i)^2 + \mid w_i \mid^2\right]\right)$$

The posterior likelihood in Equation 6 is now in the same form as the loss function in Equation 4 if $\gamma = \frac{\sigma^2}{\beta^2}$. Equation 6 provides a useful interpretation of the trade off parameter γ as the ratio of the uncertainty in the data (σ^2) relative to the uncertainty in the prior (β^2). We place more weight on the data when they are less noisy (small σ^2). We shrink our estimates more toward the prior when the data are noisy (large σ^2) or when we have a stronger belief in the prior (small β^2).

While there is a mathematical equivalence, the two approaches differ in philosophy and, in particular, in how γ is selected. In the Bayesian interpretation, γ is set by the prior beliefs – exogenously. In statistical machine learning γ is estimated endogenously from the calibration data. This also makes any interpretation of the methods as "maximum likelihood or a posteriori estimation" (i.e., estimation of the mode in Equation 6) not straight forward. This difference is a fundamental philosophical interpretation that leads to differences in estimation accuracy between statistical machine learning and Bayesian methods, as shown by EPT and discussed below.

12.2.3 General framework

Equations 4 and 6 are illustrative. The loss function in its general form, for a given γ, may be written as:

(7) $$L(w_i \mid \gamma) = \frac{1}{\gamma}.\sum_{j=1}^{J} V(w_i, data) + J(w_i)$$

The first term, $V(w_i, data)$, measures the fit between a candidate partworth estimate, w_i, and the observed data. The second term, $J(w_i)$, measures the complexity of w_i. The quadratic complexity function, $|w_i|^2$, is common, but any function may be used. In general the choice of $J(w_i)$ in Equation 7 may also depend on the data.

12.2.4 Minimization of the loss function

A potential (but not necessary) restriction is that both functions V and J in (7) should be convex. In that case, the large literature on convex optimization provides efficient methods to minimize the loss function given γ. Otherwise, non-convex (even combinatorial, for discrete decision variables) optimization methods can be used, leading to solutions that may be only locally optimal. Most of the reviewed papers use some variation of Newton's method and achieve convergence after few (e.g., 20) iterations. In some cases (Abernethy, Evgeniou, Toubia and Vert (AETV), 2006; EPT) the loss function is minimized using closed-form expressions. Computation time is rarely a limitation, and often an advantage compared to other methods such as hierarchical Bayes.

12.2.5 Trade of between fit and complexity

The tradeoff, γ, between maximizing fit and minimizing complexity can be set exogenously by the modeler or the Bayesian prior or endogenously, for example by means of cross-validation. For example, the polyhedral methods of Toubia, Simester, Hauser and Dahan (TSHD, 2003) and Toubia, Hauser and Simester (THS, 2004) implicitly assume an infinite weight on fit by maximizing fit first and then minimizing complexity among the set of estimates that maximize fit. The probabilistic polyhedral methods of Toubia, Hauser and Garcia (THG, 2006) use pretest information to select the tradeoff between fit and complexity, captured by a response error parameter α'. AETV set γ to the inverse of the number of questions, to ensure that the weight on fit increases as the amount of data increases (Vapnik 1998).

Of the conjoint analysis papers reviewed in this chapter, EBZ and EPT select γ using cross-validation – a typical approach in statistical machine learning. (See for example Wahba 1990; Efron and Tibshirani 1993; Shao 1993; Vapnik 1998; Hastie et al., 2003, and references therein). It is important to stress that *cross-validation does not require any data beyond the calibration data*.

The parameter γ is set to the value that minimizes the cross-validation error, typically estimated as follows:

- Set Cross-Validation(γ)= 0.
- For $k = 1$ to J:
  - Consider the subset of the calibration data that consists of all questions except the k^{th} one for each of the I respondents.[1]
  - Using only this subset of the calibration data, estimate the individual partworths $\{w_i^{-k}\}$ for the given γ.

[1] Variations exist. For example one can remove only one question in total from all I respondents and iterate I x J times instead of J times.

- Using the estimated partworths $\{w_i^{-k}\}$, predict the responses to the I questions (one per respondent) left out from the calibration data and let $CV(k)$ be the predictive performance achieved on these questions (e.g., root mean square error between observed and predicted responses for metric questions, logistic error for choice questions).
- Set Cross-Validation(γ) = Cross-Validation(γ) + $CV(k)$.

The parameter γ is set to the value that minimizes the cross-validation error, and is typically identified by using a line search. The cross-validation error is, effectively, a "simulation" of the out-of-sample error *without* using any out-of-sample data.

12.3 Recent optimization-based and machine-learning estimation methods

Five of the reviewed papers propose and test new estimation methods (Abernethy, Evgeniou, Toubia and Vert [AETV], 2006 is the only reviewed paper that focuses exclusively on questionnaire design and not on estimation). We examine these methods in light of the general framework outlined above. Each method may be viewed as a combination of a specific fit function, a specific complexity function, and a method for selecting the amount of trade off between fit and complexity.

12.3.1 Support vector machine estimation for choice-based conjoint analysis

Evgeniou, Boussios, and Zacharia (EBZ, 2005) focus on choice-based conjoint analysis and use a standard formulation known as the Support Vector Machine (SVM, Vapnik 1998). This has been arguably the most popular statistical machine learning method over the past 10 years, with numerous applications in various fields outside of marketing such as text mining, computational biology, speech recognition, or computer vision. An SVM uses the following loss function:

$$L(w_i \mid \gamma) = \frac{1}{\gamma} \sum_{j=1}^{J} \theta\left(1 - (x_{ij1} - x_{ij2}).w_i\right)\left[1 - (x_{ij1} - x_{ij2}).w_i\right] + \mid w_i \mid^2 \quad (8)$$

where the function θ is chosen such that $\theta(a) = 1$ if $a > 0$ and 0 otherwise. Equation 8 combines quadratic complexity control with a fit function that is slightly different from that normally used in conjoint analysis.

Recall that we assume, without loss of generality, that x_{ij1} is chosen over x_{ij2}. Hence a partworth vector w_i is consistent with choice j if $(x_{ij1} - x_{ij2}).w_i \geq 0$.

If $a \leq 1-(x_{ij1}-x_{ij2}).w_i$, then $\theta(a) = 0$ if $(x_{ij1}-x_{ij2}).w_i \geq 1$, hence, the product $\theta(1-(x_{ij1}-x_{ij2}).w_i)[1-(x_{ij1}-x_{ij2}).w_i]$ equals 0 whenever $(x_{ij1}-x_{ij2}).w_i \geq 1$. This will happen whenever the observed choice j is predicted by wi with a margin of at least 1. If choice j is not predicted by a margin of at least 1, the loss function introduces a penalty equal to the distance between $(x_{ij1}-x_{ij2}).w_i$ and 1. Fit is measured by the sum of these penalties across choices. Setting the margin to 1 plays the role of scaling the magnitudes of the partworths; any other scaling number could be used. EBZ select the parameter γ using cross-validation.

This loss function may be related to the analytic center criterion reviewed below. In particular, if each choice is interpreted as a constraint $(x_{ij1}-x_{ij2}).w_i \geq 1$, then the set of points wi that satisfy all the constraints forms a polyhedron, and for each point wi in this polyhedron, the complexity term |wi|² becomes the inverse of the radius of the largest sphere centered at wi inscribed in this polyhedron (Vapnik 1998). As a result, the value of wi that minimizes complexity is the center of the largest sphere inscribed in the polyhedron.

12.3.2 Analytic center estimation for metric paired-comparison conjoint analysis

Polyhedral methods introduced by Toubia, Dahan, Simester and Hauser (TDSH, 2003) and Toubia, Hauser and Simester (THS, 2004), and extended by Toubia, Hauser and Garcia (THG, 2006) were developed explicitly to improve adaptive question design. The primary application of these methods is to web-based conjoint analysis where respondents are free to leave the questionnaire at any time. Polyhedral methods provide means to gather the most efficient information from each question.

However, each of the three polyhedral methods provides an estimation method as a byproduct of question design. This estimation method is based on the analytic center of the set of feasible partworths – possibly probabilistic. We provide here an interpretation of analytic-center estimation within the framework of statistical machine learning.

We begin with TDSH, who assume a metric paired-comparison conjoint format. TDSH first consider the case in which there is no response error (ε_{ij}=0), and observe that the answer to each question may be interpreted as a constraint on w_i: $y_{ij} = (x_{ij1}-x_{ij2}).w_i$. The set of "feasible" estimates that satisfy all the constraints associated with all the questions is a *polyhedron*, defined as:

$$(9) \qquad \boldsymbol{\Phi}_{\{1,\ldots J\}}=\{w_i,\ 0 \leq w_i \leq 100,\ (x_{ij1} - x_{ij2}).w_i = y_{ij} \text{ for } j=1,\ldots J\}$$

where the constraint, $0 \leq w_i \leq 100$, is chosen without loss of generality to establish the scale of the partworths. Out of all feasible estimates defined by this polyhedron, TDSH select the *analytic center* of the polyhedron as their working estimate, defined as:

$$\hat{w}_i^{AC} = \arg\max_{w_i} \sum_{k=1}^{p} \log(w_{ik}) + \log(100 - w_{ik}) \tag{10}$$

subject to: $(x_{ij1} - x_{ij2}).w_i = y_{ij}$ for $j=1,\ldots J$

where w_{ik} is the k^{th} element of w_i. The analytic center is the point that maximizes the geometric mean of the slack variables associated with the inequality constraints. The logarithmic function is called a "barrier function" in interior point programming. It prevents the expression inside the logarithm from being non-positive.

For small number of questions the feasible polyhedron will be non-empty, however, as the number of questions grows in the presence of response error, it will no longer be possible to find partworths that are consistent with all of the questions and the feasible polyhedron $\boldsymbol{\Phi}_{\{1,\ldots J\}}$ will become empty. Toubia et al. (2003) follow a two-step estimation procedure: (1) find the minimum amount of response error δ^* necessary for the polyhedron to become non-empty, (2) find the analytic center of the resulting polyhedron. In particular, they first find the minimum δ^* such that the polyhedron defined as:

$$\boldsymbol{\Phi}_{\{1,\ldots J\}} = \{w_i,\ 0 \leq w_i \leq 100,\ y_{ij} - \delta^* \leq (x_{ij1} - x_{ij2}).w_i \leq y_{ij} + \delta^* \text{ for } j = 1,\ldots J\}$$

is non empty, and then estimate the partworths using the analytic center of this new polyhedron:

$$\hat{w}_i^{AC} = \arg\max_{w_i} \sum_{k=1}^{p} \log(w_{ik}) + \log(100 - w_{ik}) + \sum_{j=1}^{J} \log(y_{ij} + \delta^* - (x_{ij1} - x_{ij2}).w_i) + \log((x_{ij1} - x_{ij2}).w_i - y_{ij} + \delta^*) \tag{11}$$

We now reformulate Equation 7 within the general framework. We begin by rewriting the polyhedron $\boldsymbol{\Phi}_{\{1,\ldots J\}}$ in standard form:

$$\{\tilde{w}_i = (w_i, a_i, b_i, c_i);\ \tilde{w}_i \geq 0;\ w_i + a_i = 100;\ (x_{ij1} - x_{ij2}).w_i - b_{ij} = y_{ij} - \delta^*;\ (x_{ij1} - x_{ij2}).w_i + c_{ij} = y_{ij} + \delta^*\}$$

TDSH's fit measure becomes $V^{AC}(\tilde{w}_i, data) = \delta^*$ and their complexity control becomes:

$$J^{AC}(\tilde{w}_i) = \sum_{k=1}^{p} -\log(w_{ik}) - \log(a_{ik}) + \sum_{j=1}^{J} -\log(b_{ij}) - \log(c_{ij}) \quad (12)$$

Their two-step procedure becomes the limiting case (when $\gamma \to 0$) of the following loss function:

$$L(\tilde{w}_i \mid \gamma) = \frac{1}{\gamma} V^{AC}(\tilde{w}_i, data) + J^{AC}(\tilde{w}_i) \quad (13)$$

If one wishes, one can generalize TDSH's analytic-center estimation by choosing a non-limiting parameter γ to balance fit and complexity.

12.3.3 Analytic center estimation for choice-based conjoint analysis

THS developed a choice-based polyhedral conjoint method. Each stated-choice question is interpreted as an *inequality* constraint of the form $(x_{ij1} - x_{ij2}).w_i \geq -\delta^*$, where δ^* is a non-negative parameter that captures response error. The polyhedron of feasible partworths becomes:

$$\boldsymbol{\Phi}_{\{1,\ldots J\}} = \{w_i,\ w_i \geq 0,\ \mathbf{1}.w_i = 100,\ (x_{ij1} - x_{ij2}).w_i \geq -\delta^* \text{ for } j = 1,\ldots J\} \quad (14)$$

where $0 \leq w_i$ and $\mathbf{1}.w_i = 100$ are scaling constraints chosen without loss of generality. ($\mathbf{1}$ is a vector of one's, such that $\mathbf{1}.w_i$ is equal to the sum of the elements of w_i).

Like the metric version, the primary goal of polyhedral choice-based conjoint analysis is to select questions efficiently. Using the proposed method, THS select questions such that each choice by a respondent selects a subset of the feasible polyhedron of partworths. With this method, the feasible polyhedron never becomes empty. Ideally, with no measurement error the polyhedron will shrink toward the true value of a respondent's partworths. Intermediate estimates are the analytic center of the feasible polyhedron.

When choice-based polyhedral methods are not used to select questions, it is possible that the feasible polyhedron will become empty.

In this case, THS again follow a two-step estimation procedure: (1) find the minimum value of δ^* such that the polyhedron $\boldsymbol{\Phi}_{\{1,\ldots J\}}$ is non-empty, (2) find the analytic center of the resulting polyhedron, defined as:

$$\hat{w}_i^{AC} = \arg\max_{w_i} \sum_{k=1}^{p} \log(w_{ik}) + \sum_{j=1}^{J} \log((x_{ij1} - x_{ij2}).w_i + \delta^*) \tag{15}$$

subject to: $\mathbf{1}.w_i=100$

The polyhedron may again be written in standard form as:

$$\boldsymbol{\Phi}_{\{1,\ldots J\}}=\{\tilde{w}_i = (w_i, a_i); \tilde{w}_i \geq 0; 1.w_i = 100; (x_{ij1} - x_{ij2}).w_i - a_{ij} = -\delta^*\}.$$

The two-step estimation procedure becomes the limiting case ($\gamma \to 0$) of the following loss function:

$$L(\tilde{w}_i \mid \gamma) = \frac{1}{\gamma}.\delta^* + J^{AC}(\tilde{w}_i) \tag{16}$$

subject to: $\mathbf{1}.w_i=100$

where:

$$J^{AC}(\tilde{w}_i) = \sum_{k=1}^{p} -\log(w_{ik}) + \sum_{j=1}^{J} -\log(a_{ij}) \tag{17}$$

12.3.4 Probabilistic analytic center estimation

THG offer a Bayesian interpretation of the method proposed by THS, which enables a richer treatment of response error and which allows capturing informative priors on the partworths. They consider prior distributions represented by mixtures of uniform distributions supported by polyhedra. (In one dimension, a uniform distribution supported by a polyhedron simply becomes a uniform distribution on an interval; in two dimensions, it is a uniform distribution on a rectangle, etc.) Mixtures of such distributions may be used to approximate any prior distribution. THG also provide a method by which prior beliefs are not directly captured by probability distributions, but rather by probabilistic constraints on some combinations of the parameters (e.g., the importance of feature A is greater than m with probability q). The general expression for this class of distributions is as follows:

$$P(w_i) = \sum_{m=1}^{M} \omega_m P_{\Psi_m}(w_i)$$

where M is any positive integer, $\{\omega_1,...,\omega_M\}$ is a set of positive weights such that $\sum_{m=1}^{M}\omega_m = 1$, $\{\Psi_1,...,\Psi_M\}$ is a set of polyhedra, and $P_{\Psi_m}(w_i)$ is the uniform probability distribution with support Ψ_m. The previous methods of TDSH and THS implicitly assume a uniform prior on the polyhedron defined by the scaling constraints.

THG combine this class of prior distributions with a conjugate class of likelihood function such that in each question, the profile with the highest deterministic utility is chosen with probability α', and the (J-1) other profiles are chosen with probability (1- α')/(J-1) each. Such likelihood functions are step functions with two values, one taken by all points that are consistent with the choice, and the other taken by all points that are inconsistent with the choice. The specific values are driven by the parameter α'. This class of likelihood functions is attractive because the posterior distribution on the partworths is also equal to a mixture of uniform distributions supported by polyhedra. After J questions, the posterior distribution on w_i may be written as follows:

$$P(w_i) = \sum_{m=1}^{M}\sum_{s\in S_J}\omega_{ms}P_{\Phi_s\cap\Psi_m}(w_i),$$

where S_J is the set of all subsets of the questions {1, 2, ,… J}, and for a subset s of S_J, Φ_s is the polyhedron corresponding to the questions in s. The parameter ω_{ms} is the mixture weight on the polyhedron defined by the intersection between Φ_s and the prior polyhedron Ψ_m (see THG for a method to approximate these weights).

Although other techniques could be used as well, THG select the parameter α' from a pretest sample of respondents, following the tradition of aggregate customization (Arora and Huber 2001, Huber and Zwerina 1996).

Given the posterior distribution written as a mixture of uniform distributions supported by polyhedra, several methods could be used to produce point estimates of the partworths. For example, an algorithm could be developed that allows sampling from this posterior distribution and estimating the partworths as the mean of this distribution. For simplicity, THG estimate the partworths as the mixture of the analytic centers of the polyhedra involved in the mixture.

THG essentially shift the focus from the minimization of a loss function to the exploration of a posterior distribution. However their approach may be still be framed within statistical machine learning. In particular, complexity control is achieved by the prior distribution. The parameter α' effectively controls the trade off between fit and complexity. For example, α'=1 implies no response error and the estimates fit the data perfectly; α'=1/J implies non-informative choices and all inference will be based only on the prior.

12.3.5 Using complexity control to model heterogeneity

Hierarchical Bayes has been one of the most successful developments in the estimation of conjoint-analysis partworths (Lenk et al., 1996; Allenby and Rossi 1999; Rossi and Allenby 2003; Rossi, Allenby and McCulloch., 2005).[2] Liu, Otter, and Allenby (2006) suggest that one reason for this accuracy is the likelihood principle which states that the likelihood best summarizes the information in the data. Another, less formal hypothesis is that Bayesian methods are accurate because they robustly shrink individual-level estimates toward the mean of the population. As motivated by the analogy of ridge regression and Bayesian priors, the shrinkage in hierarchical Bayes can be seen as analogous to complexity control.

EPT explore this interpretation for both metric and choice data. In the metric case, the loss function can be formulated as follows:

(18)

$$L(\{w_i\}, w_0, D \mid \gamma) = \frac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - x_{ij}.w_i)^2 + \sum_{i=1}^{I}(w_i - w_0)^T D^{-1}(w_i - w_0)$$

subject to D being a positive semi-definite matrix scaled to have a trace of 1.

We note that this formulation is not identical to hierarchical Bayes methods. It differs in both philosophy and computation. Nonetheless, it is an interesting analogy.

There are a number of interesting characteristics associated with this loss function: (1) estimates are obtained simultaneously for all respondents, (2) estimates are shrunk toward a common partworth vector that may differ from the unit vector, and (3) the parameter γ dictates the trade off between fit and shrinkage.

Although the population vector w_0 is not defined to be the population means, EPT show that the value of w_0 that minimizes the loss function must equal the population mean. The matrix D is analogous to the covariance matrix of the partworths; the shrinkage penalty is greater for partworths that are distant from the mean w_0 along directions in which there is less variation across respondents. By scaling D with its trace, the authors assure that the optimization problem is convex. Although the actual minimization is beyond the scope of this chapter, we note that, for a given γ, the optimal solution is in closed form and hence computationally efficient (see paper for more details).

---

[2] Technically, Bayesian methods sample from the posterior distribution of the parameters rather than provide estimates in the classical sense. For simplicity, we refer to the mean of the posterior distribution as the partworth estimates.

For choice data, EPT substitute the logit log-likelihood as the fit measure. The loss function becomes:

(19)

$$L(\{w_i\}, w_0, D \mid \gamma) = \frac{1}{\gamma}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(-\log\frac{\exp(x_{ij1}.w_i)}{\exp(x_{ij1}.w_i)+\exp(x_{ij2}.w_i)}\right) + \sum_{i=1}^{I}(w_i - w_0)^T D^{-1}(w_i - w_0)$$

Because closed-form expressions are not available with this formulation, Newton's method is used (any other convex optimization method could be used) to minimize the loss function for a given γ.

To assess the impact of the differing philosophies, EPT compare their approach to hierarchical Bayes. In particular, they consider the following two HB models for metric and choice data respectively (in both cases a diffuse prior is assumed on w_0):

	Metric data:
Likelihood:	$y_{ij}=x_{ij}.w_i+\varepsilon_{ij}$, $\varepsilon_{ij}\sim N(0, \sigma^2)$
First-stage prior:	$w_i\sim N(w_0,D)$
Second-stage priors:	$\sigma^2\sim IG(r_0/2,s_0/2)$
	$D^{-1}\sim W(\eta_0, \eta_0.\Delta_0)$
	Choice data:
Likelihood:	$\text{Prob}(x_{ij1} \text{ chosen})= \dfrac{\exp(x_{ij1}.w_i)}{\exp(x_{ij1}.w_i)+\exp(x_{ij2}.w_i)}$
First-stage prior:	$w_i\sim N(w_0,D)$
Second-stage prior:	$D^{-1}\sim W(\eta_0, \eta_0.\Delta_0)$,

where *IG* denotes the inverse gamma distribution and *W* the Wishart distribution.

Both the machine learning and hierarchical Bayes approaches shrink estimates toward the population mean. Moreover, in the case of metric data, EPT are able to show that the individual-level estimates *conditional* on D and w_0 are given by the exact same mathematical expressions.

However they identify two major and fundamental differences between their approach and HB. First, while the former involves the minimization of a loss function, the latter involves sampling from a posterior distribution. Hence in HB point estimates are only one of the many ways to summarize and describe the posterior distribution. Other important statistics include the standard deviation of this distribution. EPT illustrate that standard deviations and confidence intervals may also be obtained in their framework, using for example bootstrapping (Efron and Tibshirani, 1993).

Second, the two methods differ on how they select the amount of shrinkage. In HB the amount of shrinkage is selected, in part, by prior judgment embodied in the second-stage prior parameters (η_0, Δ_0, r_0, and s_0 in the metric case; η_0 and Δ_0 in the choice case); in machine-learning it is determined from the calibration data (γ). By selecting γ through cross-validation, it may not be surprising that the machine-learning approach can outperform HB unless, of course, the second-stage priors are chosen with prescience. See EPT for detailed results.

12.3.6 Summary of optimization and machine-learning estimation

Table 1 describes and contrasts the estimation methods reviewed in this section.

Table 1: Characteristics of the reviewed estimation methods

Paper(s)	Fit measured by	Complexity measured by	Trade off fit / complexity
Evgeniou,Boussios, Zacharia (2005)	Support vector machine	Quadratic norm on the partworths	Determined by cross-validation
Toubia, Simester, Hauser, Dahan (2003)	Response error to obtain feasble polyhedron	Analytic center	Maximize fit first, then minimize complexity
Toubia, Simester, Hauser (2004)	Response error to obtain feasible polyhedron	Analytic center	Maximize fit first,then minimize complexity
Toubia, Hauser, Garcia (2006)	Polyhedral mixture	Informative prior	Based on pretest
Evgeniou, Pontil, Toubia (2006)	Sum of squared errors / logistic likelihood	Difference from population means	Determined by cross-validation

12.4 Recent optimization-based and machine-learning adaptive questionnaire design methods

One of the breakthroughs in the 1980s was the ability to adapt conjoint analysis questions to the observed responses of consumers. Algorithms developed by Johnson (1987, 1991) for Adaptive Conjoint Analysis (ACA) enabled researchers using computer-aided interviews to ask more efficient questions. For almost 20 years ACA was one of the most commonly applied methods, only recently surpassed by choice-based conjoint analysis. It is only in the past few years that we have seen a resurgence in adaptive questionnaire design. This resurgence has been

made possible by the development of new efficient computational algorithms and the continued growth in computing power. It is now feasible to adapt questions in an on-line environment using sophisticated background computations that run in the time it takes to download the code for the next page display – the respondent notices little or no delay due to this computation. While the methods are still being perfected, the results to date suggest that in many applications these adaptive questioning methods enable researchers to design methods that ask fewer questions yet still provide estimates that are sufficiently accurate for important managerial decisions. The methods work for a variety of conjoint analysis formats, including both metric paired-comparison data and choice-based data.

In this chapter we review four newly proposed methods that enable researchers to adapt questions at the level of the individual respondent.

12.4.1 Experimental design principles

Non-adaptive questionnaire design builds primarily on the field of experimental design (Chaloner and Verdinelli 1995; Ford, Kitsos and Titterington 1989; Kuhfeld, Tobias and Garratt 1994; Pukelsheim 1993; Steinberg 1984). The approach can be summarized as minimizing a norm of the asymptotic covariance matrix of the parameter estimate $\hat{w}_i$. Under mild assumptions (Newey and McFadden 1994), it can be shown that the maximum likelihood estimate of w_i is asymptotically normal with covariance matrix equal to the inverse of the information matrix $\mathbf{\Omega}$, given by the Hessian (second-derivative matrix) of the loss function minimized in estimation.

Non-adaptive efficient designs maximize a norm of the information matrix $\mathbf{\Omega}$, the inverse of the covariance matrix. The most widely used norm is the determinant, giving rise to so-called D-efficient designs (Arora and Huber 2001; Huber and Zwerina 1996; Kuhfeld, Tobias and Garratt 1994; Kuhfeld 2005). D-efficiency minimizes the volume of the confidence ellipsoid around the maximum likelihood estimate $\hat{w}_i$, defined by $\{w : (w - \hat{w}_i)^T \Omega (w - \hat{w}_i) \leq 1\}$, and makes this ellipsoid as spherical as possible (Greene 2000). For example, the well-known orthogonal and balanced designs (Addelman 1962, Kuhfeld, Tobias and Garratt 1994), when they exist, maximize efficiency.

For stated-choice data, the information matrix depends on the true partworths w_i. In most cases, efficiency can be improved by attempting to achieve utility (or choice) balance such that the alternatives in each choice set are close in utility (close in probability of choice) where utility is often calculated based on prior beliefs about the partworths. There are many algorithms to increase efficiency: Arora and Huber (2001), Huber and Zwerina (1996), Kanninen (2002), Sandor and Wedel (2001), and Hauser and Toubia (2005). Abernethy, Evgeniou, Toubia and Vert (AETV, 2006) note that similar principles have been used in other fields such as active learning (Tong and Koller 2000).

The *adaptive* question design methods use similar fundamental principles. For example, Toubia, Dahan, Simester and Hauser (TDSH, 2004), Toubia, Hauser and

Garcia (THG, 2006), and AETV select the next questions to achieve utility balance based on estimates from the answers to previous questions. These methods attempt to minimize the amount of "uncertainty" around the estimate and to make uncertainty similar in all directions. Questions are chosen to reduce the uncertainty along the most uncertain dimension.

In polyhedral methods, uncertainty is characterized by the polyhedron of feasible estimates (which may conceptually be related to the confidence ellipsoid in maximum likelihood estimation), and questions are selected to maximally reduce the volume of this polyhedron and minimize the length of its longest axis (making it more spherical). In a similar vein, AETV characterize uncertainty by the inverse of the Hessian of the loss function (equal to the information matrix), and select questions to maximally increase the smallest positive eigenvalue of the Hessian. We now review these methods in greater detail.

12.4.2 Polyhedral question design

For ease of exposition, we describe the intuition for polyhedral methods when the feasible polyhedron is non-empty. The same intuition applies to the expanded polyhedron.

In polyhedral question design, the constraints imposed by the answers to previous questions form a polyhedron. All points in the polyhedron are consistent with prior answers. A smaller polyhedron implies a smaller set of feasible estimates and, hence, less uncertainty about the partworths. For example, the gray region in Figure 1 is the feasible polyhedron after a set of questions and all points (partworths) in that gray area are consistent with the answers to prior questions. Our goal is to select the next question such that when the question is answered, the resulting polyhedron is as small as possible.

Formally, let $\Phi\{1,J\}$ denote the polyhedron defined by the answers to the first J questions asked of a given respondent. Let $\Phi\{1,J+1\}$ denote the new polyhedron formed when the J + 1st answer constrains $\Phi\{1,J\}$. Consider first metric paired-comparison questions. The new constraint will be of the form $(x_{i(J+1)1}-x_{i(J+1)2}).w_i = y_{i(J+1)}$. The set of points, w_i that satisfy this new constraint is a hyperplane perpendicular to the vector $(x_{i(J+1)1}-x_{i(J+1)2})$. This is shown as the green surface in Figure 1. The new polyhedron, $\Phi\{1,J+1\}$, is the intersection between the current polyhedron, $\Phi\{1,J\}$, and this hyperplane.

We must now select a question that, when answered, minimizes the volume of the new polyhedron, $\Phi\{1,J+1\}$. In addition we want to make it more spherical. Intuitively, we satisfy these criteria if we select the hyperplane to be orthogonal to the longest axis of the current polyhedron. Mathematically, this means that we select the two profiles in the next question such that the line, $(x_{i(J+1)1}-x_{i(J+1)2})$, is as close as possible to the longest axis of the polyhedron. (At minimum, by intersecting the current polyhedron with a hyperplane perpendicular to the current longest axis will ensures that the longest axis of the next polyhedron will be strictly smaller than the longest axis of the current polyhedron.)

The mathematics are complex, but the basic idea is to find the analytic center of the polyhedron and choose the smallest ellipsoid such that the polyhedron is surrounded by the ellipsoid with its center at the polyhedron's analytical center. Then, by solving an eigenvalues problem, TDSH select the longest axis of the ellipsoid as representing the longest axis of the polyhedron.

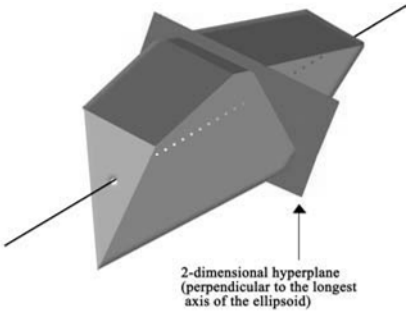

Figure 1: Cut perpendicular to the longest axis – metric data case (From Toubia, Simester, Hauser, and Dahan 2003)

The methods and philosophy for choice-based data follow the same intuition, with a few modifications (THS, THG). For binary stated-choice data, new constraints are inequality constraints of the form (xi(J+1)1-xi(J+1)2).wi ≥ 0 (The method is extended easily to multiple alternatives in the choice set.). The set of points that satisfy the constraint implied by the J+1st answer is a half-space. If the boundary of his half-space intersects the feasible polyhedron, Φ{1,J}, it will divide the polyhedron into two sub-polyhedra. One sub-polyhedron corresponds to the choice of xi(J+1)1 and the other to the choice of xi(J+1)2. In other words, the respondent's choice in the J+1st question identifies one or the other sub-polyhedron. All points in the chosen sub-polyhedron are consistent with the answers to all J+1 questions.

THS again seek to choose questions such that the resulting sub-polyhedron will be as small and spherical as feasible. THS show that the expected volume of Φ{1,J+1} is reduced efficien tly if the separating hyperplane is chosen so that it goes through the center of the feasible polyhedron, such that each choice alternative is as equally likely as possible. Such choice balance assures that the resulting polyhedra are of approximately equal volume. This is illustrated in Figure 2a.

Of the many hyperplanes that split the feasible polyhedron, the hyperplane that will make the resulting sub-polyhedra as spherical as possible is the hyperplane that is perpendicular to the longest axis of the polyhedron. This is illustrated in Figure 2b.

The two points at which the longest axis intersects the boundary of the polyhedron provide two target partworth vectors. The final step is to construct one

profile associated with each of them. Each profile is obtained by simply solving a budget constraint problem. That is, for each target partworth vector, THS construct a choice alternative that maximizes utility subject to a budget constraint.

A strength of adaptive polyhedral question design is that questions are chosen such that the resulting polyhedra are always feasible and non-empty. However, this strength is also a weakness. When there is response error, early errors propagate. A choice made in error forever assures that the true partworths are not in any subsequent polyhedra. As a result, early tests indicated that adaptive choice-based questions improved accuracy and efficiency when response error was small, but not when it was large.

THG address response error with a probabilistic generalization of polyhedral methods. They model potential response error by assuming that each constraint applies with probability α^3, where α is based on pretest data. They then show that polyhedral methods can be given a Bayesian interpretation such that the posterior distribution of the partworths is a mixture of polyhedra, each defined by a subset of the constraints imposed by the respondent's answers to the chosen questions (see details above). With this interpretation, it is simple conceptually to extend adaptive polyhedral choice-based question design. New questions are chosen based on the longest axis of the appropriate mixture of polyhedra.

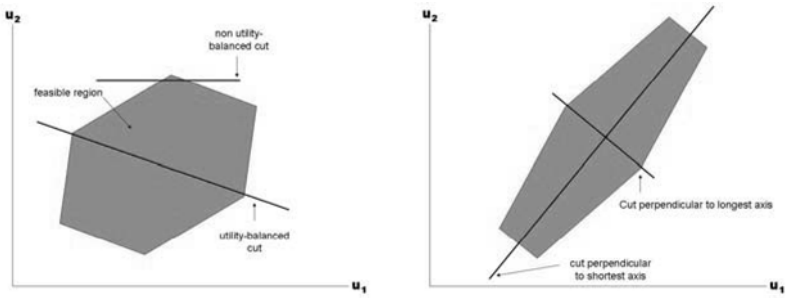

Figure 2: Utility balance cuts and cuts perpendicular to the longest axis–choice data case (from Toubia, Hauser, Simester 2004)

12.4.3 Hessian-based adaptive question design for choice-based conjoint analysis

Statistical learning methods also provide a means to adapt conjoint questions. AETV define loss functions that are convex and twice differentiable. For such loss functions, uncertainty with respect to the partworths is captured by the inverse of the Hessian of the loss function. Their goal is then to design a question by selecting a direction in parameter space that maximally decreases this matrix subject to enforcing utility balance.

In particular, AETV propose solving the following quadratic optimization problem in order to find a direction of maximal uncertainty:

(20)

$$\min_{z} z\nabla^2 L(\hat{w}_i \mid \gamma) z^T$$

subject to: $z.\hat{w}_i = 0, \quad z.z^T = 1$

where $\hat{w}_i$ is the value of the partworths that minimize the loss function, L, $\nabla^2 L(\hat{w}_i \mid \gamma)$ is the Hessian of the loss function at $\hat{w}_i$, and $z.z^T = 1$ is a scaling constraint. The optimal solution to Equation 20 is the eigenvector associated with the smallest positive eigenvalue of the matrix: $B_i = (I - \frac{\hat{w}_i.\hat{w}_i^T}{\hat{w}_i^T.\hat{w}_i}).\nabla^2 L(\hat{w}_i \mid \gamma)$ where I is the identity matrix.

The question-design algorithm is implemented as follows:

1. Find $\hat{w}_i$ such that $\hat{w}_i$ minimizes the loss function $L(w_i \mid \gamma)$. Because L is convex, there are many convex optimization methods that are efficient.
2. Find the (normalized) eigenvector z associated with the smallest positive eigenvalue of the matrix $B_i = (I - \frac{\hat{w}_i.\hat{w}_i^T}{\hat{w}_i^T.\hat{w}_i}).\nabla^2 L(\hat{w}_i \mid \gamma)$.
3. Find a pair of profiles such that $(x_{i(J+1)1}\text{-}x_{i(J+1)2})$ is as close as possible to being proportional to z and such that utility balance is preserved: $(x_{i(J+1)1}\text{-}x_{i(J+1)2}).\hat{w}_i \approx 0$.[3]

[3] AETV use the Knapsack approach of THS.

AETV illustrate the Hessian approach with a Ridge Regression loss function – similar to that used in support vector machines (where the constant, 1, scales the partworths):

$$L(w_i \mid \gamma) = \frac{1}{\gamma}\sum_{j=1}^{J}(1-(x_{ij1}-x_{ij2}).w_i)^2 + \mid w_i \mid^2 \tag{21}$$

With this loss function, the minimum $\hat{w}_i$ and the Hessian are given in closed form. To avoid the computational delays of cross-validation, γ is set equal to the inverse of the number of questions so that the data are weighed more heavily as more data become available. This specification is motivated by Vapnik (1998).

12.4.4 Summary of adaptive question design

Machine-learning and fast polyhedral algorithms have made it feasible to adapt both metric paired-comparison and choice-based conjoint questions to each respondent. Such questions promise to be more accurate and customized to focus precision where it is most needed. The basic concept is that each conjoint question constrains the set of feasible partworths. A researcher's goal is to find the questions that impose the most efficient constraints, where efficiency is defined as maximally decreasing the uncertainty in the estimated partworths.

To date, all question-design algorithms use information from a single respondent to select questions for that respondent. However, one of the lessons of both hierarchical Bayes and the machine learning approaches of EPT is that population-level information can improve accuracy at the individual level. We predict that such pooling methods will be feasible in the near future and make promising areas for research. For example one could adapt the Hessian method of AETV to a loss function like the ones in Equations 18 or 19 used by EPT.

12.5 Applications, simulations, and empirical tests

Conjoint analysis has a long history of validation and application. See, for example, Green (2004). Methods such as ACA, logit analysis of choice-based conjoint analysis, and hierarchical Bayes estimation have been improved through hundreds of applications. Such use and its related research have led to incremental improvement of these standard methods. By contrast, the methods reviewed in this paper are relatively new, each with only a few applications. On one hand, such tests usually involve only one or a few applications and, thus, must be considered experimental. On the other hand, we expect the performance on these tests to be lower bounds on eventual performance which is likely to improve with experience.

Despite the nascent nature of these methods, they have performed remarkably well in both Monte Carlo simulations and empirical applications. We review here

applications, comparisons of estimation methods, and comparisons of question design methods.

12.5.1 Applications

Metric paired-comparison polyhedral methods. Toubia, Dahan, Simester and Hauser (TDSH, 2003) study preferences for the features of laptop computer bags. In their experiments, respondents were given the choice of real laptop bags worth approximately $00. Predictions were qu ite accurate. In addition, the models appear to have described market shares when the laptop bags were introduced to a real market.

Adaptive choice-based polyhedral conjoint methods. Toubia, Hauser, and Simester (THS, 2004) studied the preferences for the features of executive educational programs. The data were used to design MIT's 12-month executive program, which has since been implemented successfully. Toubia, Hauser and Garcia (THG, 2006) study the diffusion of non-traditional closures, "Stelvin" screw-tops, for premium wines by interviewing over 2,200 leading-edge wine consumers in the US, Australia and New-Zealand. They were able to identify the marketing actions that would be necessary to achieve market penetration in the US to match that in Australia and New Zealand.

Hessian-based adaptive choice-based conjoint analysis. Abernethy, Evgeniou, Toubia and Vert (AETV, 2006) study consumer preferences for digital cameras. They explore how respondents value different levels of price, resolution, battery life, optical zoom, and camera size.

Heterogeneous partworth estimation with complexity control. Evgeniou, Pontil and Toubia (EPT, 2006) test their method with a full-profile ratings study of personal computers collected by Lenk et al. (1996) and apply their method using data from a choice-based conjoint study of carbonated soft drinks collected by a professional market research company.

12.5.2 Comparisons of estimation methods

The basic results from the papers reviewed in this chapter are three-fold. (1) Individual-level optimization methods tend to outperform traditional individual-level methods that use neither complexity control nor shrinkage. (2) Individual-level methods often under-perform methods that use population-based shrinkage (either Bayesian or complexity-control shrinkage). (3) Complexity-control shrinkage often outperforms Bayesian shrinkage.

Metric paired-comparison analytic-center estimation. TDSH test metric analytic-center estimations with both Monte Carlo simulations and an empirical application. In the simulations they find that, for homogeneous populations, HB consistently performs better than analytic center estimation, likely because HB uses population-level data to moderate individual estimates. For heterogeneous populations, analytic-center estimation performs better, especially when paired

with polyhedral question design. They also find that HB is relatively more accurate when response errors are high, but analytic center estimation is more accurate when response errors are low. For external validity tests, they found that HB outperforms analytic-center estimation for fixed, orthogonal questions, but that analytic-center estimation does better when matched with polyhedral questions.[4]

Adaptive choice-based analytic-center estimation. THS compare choice-based analytic-center estimation to HB on four metrics – root mean square error, hit rate, correlation among partworths, and the percent of respondents for whom a method predicts best. Analytic-center estimation performs well when matched with polyhedral question design in domains were there is high heterogeneity. Otherwise, HB does well in all domains. However, if one takes a convex combination of the population mean and the individual-level analytic center estimates, the resulting "shrinkage" estimates outperform HB.[5]

THG test the probabilistic interpretation of adaptive choice-based analytic-center estimation. Based on Evgeniou, Boussios and Zacharia (EBZ, 2005), their HB benchmark includes constraints that all partworths be positive. Such constraints improve predictive ability and are easily implemented with rejection sampling. To distinguish this method from standard HB, we label it HBP (P for positivity). THG find that taking response errors into account and using informative priors improve analytic-center estimation. At least one of the two improvements outperforms deterministic analytic-center estimation in all tests. Informative priors appear to provide the greater improvement. HBP is significantly better in most cases. We suspect that had HBP been applied in the earlier tests, it would have been best in most comparisons.

As a summary, analytic-center estimation is better than HB in some domains, but not as good as HBP. On the other hand, shrinkage-based analytic-center estimation shows considerable promise. We hypothesize that the dominant effect is the ability to use population-level information to improve individual-level estimates. If population-level information is used, analytic-center estimation may ultimately improve to be as accurate or more accurate than HBP.

Support vector machines. EBZ show that their method based on Support Vector Machines is more robust to response error compared to other individual-level methods. While their method does not perform as well as HBP in situations in which there is no interaction between attributes, it consistently outperforms HBP when interactions are present.

Heterogeneous partworth estimation with complexity control (HPECC). EPT show that their methods perform consistently better than HB (with relatively diffuse second-stage priors), both with choice and metric data, and both on

---

[4] We caution the reader that the HB method used as a benchmark in this paper was such that no external constraints were imposed. Subsequent research suggests that HB does much better if the partworths are constrained to be positive (Evgeniou, Boussios and Zacharia 2005). This caveat also applies to the simulation tests in THS.

5 THS do not estimate a γ through cross-validation but rather choose a γ based on out-of-sample performance. Their results are, thus, only suggestive.

simulated as well as field data.[6] In the case of metric data, they report simulations in which they vary heterogeneity, response error, and the number of questions per respondent. They find that their method significantly outperforms standard HB in 7 out of their 8 experimental conditions (2 levels per experimental factor). They further compare these two metric estimation methods using a metric-full-profile data set on the features of computers (from Lenk et al. 1996). Heterogeneous partworth estimation with complexity control (HPECC) significantly outperforms HB on holdout prediction, using both all 16 questions as well as a random subset of 8 questions per respondent (14 parameters are estimated per respondent). For choice data, they find that HPECC outperforms HB in 6 out of 8 experimental conditions. Empirically, HPECC outperforms HB with 16 questions per respondent for data on carbonated soft drinks, and does not perform significantly differently when 8 questions are used per respondent (17 parameters are estimated per respondent).

EPT's simulation and empirical validity tests reinforce the dominating effect of shrinkage/complexity-control. Population means clearly improve predictive performance by making the partworth estimates more robust. Their results also suggest that prediction is improved when γ is chosen endogenously rather than based on prior beliefs. Finally, EPT show that their approach allows modeling and estimating models with large numbers of attribute interactions. Estimates remain robust and significantly better than that of HB even if the total number of parameters becomes substantially larger than the number of observations per respondent. This result confirms earlier findings reported by EBZ for individual-level partworth estimation.

12.5.3 Comparisons of question design methods

The overall summary of the comparisons of adaptive question design methods is that adapting questionnaires at the individual level can improve performance.

Adaptive metric paired-comparison polyhedral question design. TDSH compare polyhedral question design to ACA as well as fixed designs and random designs. Monte Carlo simulations suggest that, when there are a small number of questions, polyhedral question design method outperforms the other three benchmarks. However, the performance may be due, in part, to endogeneity bias in ACA – prior, self-explicated questions are used in question design but standard HB estimation uses these only as constraints in estimation (Hauser and Toubia 2005; Liu, Otter, and Allenby 2006). When more questions are asked such that the questions cover the range of features more completely, fixed designs emerge as viable alternatives for some domains. In empirical tests, adaptive polyhedral questions outperform both fixed and ACA benchmarks.

---

[6] EPT do not consider positivity constraints on the partworths, neither for their methods nor for HB.

Adaptive choice-based polyhedral question design. THS simulations suggest that, when response error is low, choice-based polyhedral questions outperform random questions, fixed orthogonal questions, and questions chosen by aggregate customization (Arora and Huber 2001, Huber and Zwerina 1996). Furthermore, high heterogeneity tends to favor individual-level adaptation. When response error is high, the best method depends on the tradeoffs between response error and heterogeneity. THS apply their method empirically, but were not able to obtain validation data. However, they do show that the method achieves choice balance throughout the questioning sequence.

THG attempt to improve adaptive choice-based polyhedral methods so that they might handle high-response error domains. Their simulations suggest that taking response errors into account and using informative priors improve polyhedral question design. Compared to the THS's deterministic algorithm, random questions, fixed questions, and aggregate customization, at least one of the two probabilistic modifications is best or tied for best in all experimental cells. Their empirical tests (wine consumers) suggest that probabilistic polyhedral question design performs better than aggregate customization question design in three of the four panels and never significantly worse.

Hessian-based adaptive choice-based conjoint analysis. The Monte Carlo simulations and the field test reported by AETV confirm that individual-level adaptation outperforms random and non-adaptive benchmarks when response error is low and/or when respondent heterogeneity is high. Moreover, the use of complexity control in the loss function improves robustness to response error, hence largely overcoming possible endogeneity biases inherent to adaptive questionnaires (Hauser and Toubia 2005).

In summary, optimization-based adaptive question design for conjoint analysis shows considerable promise. In many cases, the tested methods outperform non-adaptive methods. Adaptation shows the most promise when response errors are low, when heterogeneity is high, and/or when relatively few questions are to be asked. However, the potential of individual-level adaptation is not limited to these domains. With application and incremental improvements we expect that the performance of these methods will improve further.

12.6 Conclusions and opportunities for future research

This chapter reviews some recent developments in the application of optimization methods and machine learning in conjoint estimation and question design. Although the many methods are disparate, they can be linked through a statistical learning framework and philosophy. This framework suggests that specific methods may be described by the choice of a measure of fit, a measure of complexity, and an approach for determining the trade off between fit and complexity. Adaptive questionnaire design is achieved by combining optimization and machine learning with principles of experimental design to select questions that minimize the uncertainty around the estimates.

We hope that this chapter will motivate future applications and research in this area. In particular, we hope that researchers will build upon the many successful methods in conjoint analysis that have been developed either to estimate partworths or to design questions. Complexity control, shrinkage, and adaptive optimization of questions all show considerable potential to improve extant methods and to develop new methods.

12.7 References

Abernethy, Jacob, Theodoros Evgeniou, Olivier Toubia, and Jean-Philippe Vert (AETV, 2006), "Eliciting Consumer Preferences using Robust Adaptive Choice Questionnaires," Working Paper, INSEAD, Fontainebleau, France.

Addelman, Sidney (1962), "Symmetrical and Asymmetrical Fractional Factorial Plans", *Technometrics*, 4 (February) 47-58.

Allenby, Greg M., Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, March/April, p. 57 - 78.

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.

Chaloner, Kathryn, and Isabella Verdinelli (1995), Bayesian Expe rimental Design: A Review", *Statistical Science*, 10(3), 273-304.

Dawes, R. M and B. Corrigan (1974), "Linear Models in Decision Making," *Psychological Bulletin*, 81, 95-106.

Efron, Bradley, and Robert Tibshirani (1993), *An Introduction to the Bootstrap*, (New York, NY: Chapman and Hall).

Einhorn, Hillel J. (1970), "The Use of Nonlinear, Noncompensatory Models in Decision Making," *Psychological Bulletin*, 73, 3, 221-230.

Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (EBZ, 2005), "Generalized Robust Conjoint Estimation," *Marketing Science*, 24(3), 415-429.

Evgeniou, Theodoros, Massimiliano Pontil, and Tomaso Poggio (2000), "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, 13, 1-50.

Evgeniou, Theodoros, Massimiliano Pontil, and Olivier Toubia (EPT, 2006), "A Convex Optimization Approach to Modeling Heterogeneity in Conjoint Estimation," Working Paper, INSEAD, Fontainebleau, France.

Ford, I., Kitsos, C.P. and Titterington, D.M. (1989) "Recent Advances in Nonlinear Experimental Designs," *Technometrics*, 31, 49-60.

Green, Paul E. (2004), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Conjoint Analysis, Related Modeling, and Applications: Market Research and Modeling: Progress and Prospects*, Jerry Wind and Paul Green, Eds., (Boston, MA: Kluwer Academic Publishers), 141-168.

Green, Paul E. and Vithala R. Rao (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8, (August), 355-63.

Greene, William (2003), *Econometric Analysis*, (Englewood Cliffs, NJ: Prentice Hall).

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2003), *The Elements of Statistical Learning*, (New York, NY: Springer Series in Statistics).

Hauser, John R., Ely Dahan, Michael Yee, and James Orlin (2006), ""Must Have" Aspects vs. Tradeoff Aspects in Models of Customer Decisions," *Proceedings of the Sawtooth Software Conference in Del Ray Beach, FL*, March 29-31, 2006

Hauser, John R., and Olivier Toubia (2005), "The Impact of Endogeneity and Utility Balance in Conjoint Analysis," *Marketing Science*, Vol. 24, No. 3.

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, (August), 307-317.

Jain, Arun K., Franklin Acito, Naresh K. Malhotra, and Vijay Mahajan (1979), "A Comparison of the Internal Validity of Alternative Parameter Estimation Methods in Decompositional Multiattribute Preference Models," *Journal of Marketing Research*, 16, (August), 313-322.

Johnson, Richard (1987), "Accuracy of Utility Estimation in ACA," Working Paper, Sawtooth Software, Sequim, WA, (April).

Johnson, Richard (1991), "Comment on 'Adaptiv e Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28, (May), 223-225.

Kanninen, Barbara (2002), "Optimal Design for Multinomial Choice Experiments," *Journal of Marketing Research*, 36 (May), 214-227.

Kuhfeld, Warren F. (2005), *Marketing Research Methods in SAS*, SAS Institute Inc., Cary, NC (USA). Available at http://support.sas.com/techsup/technote/ts722.pdf.

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Applications," *Journal of Marketing Research*, 31, 4(November), 545-557.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15(2), p. 173--91.

Liu, Qing, Thomas Otto, and Greg M. Allenby (2006), "Investigating Conjoint Analysis Bias in Conjoint Analysis," Working Paper, Ohio State University, Columbus, OH.

Newey, Whitney K., and Daniel McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, edited by R.F. Engle and D.L. McFadden, Elsevier Science.

Rossi, Peter E., Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22(3), p. 304-328.

Rossi, Peter E., Greg M. Allenby, Robert McCulloch (2005), *Bayesian Statistics and Marketing*. (New York, NY: John Wiley and Sons).

Sandor, Zsolt, and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research*, 38, 4, 430-444.

Shao, Jun (1993), "Linear model selection via cross-validation," *Journal of the American Statistical Association*, 88(422), p. 486--494.

Srinivasan, V., and Allan D. Shocker (1973a), "Linear Programming Techniques for Multidimensional Analysis of Preferences," *Psychometrika*, 38 (3), 337-369.

Srinivasan, V. and Allen D. Shocker (1973b), "Estimating the Weights for Multiple Attributes in a Composite Criterion Using Pairwise Judgments," *Psychometrika*, 38, 4, (December), 473-493.

Steinberg, D.M., and Hunter, W.G. (1984), "Experimental Design: Review and Comment," *Technometrics*, 26, 71-97.

Tikhonov, A., and V. Arsenin (1977), *Solutions of Ill-posed Problems*, W. H. Winston, Washington, D.C. (OLIVIER: add full first names)

Tong, S., and D. Koller (2000), "Support vector machine active learning with applications to text classification," *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford University, CA, USA.

Toubia, Olivier, John R. Hauser, and Rosanna Garcia (THG, 2006), "Probabilistic Polyhedral Methods for Adaptive Conjoint Analysis: Theory and Application," forthcoming, *Marketing Science*.

Toubia, Olivier, John R. Hauser, and Duncan I. Simester (THS, 2004), "Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis," *Journal of Marketing Research*, 41, 116-131.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (TDSH, 2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22(3), 273-303.

Vapnik, Vladimir (1998), *Statistical Learning Theory*, (New York, NY: John Wiley and Sons).

Wahba, Grace (1990), "Splines Models for Observational Data," *Series in Applied Mathematics*, Vol. 59, SIAM, Philadelphia.

Wittink, Dick R. and Philippe Cattin (1981), "Alternative Estimation Methods for Conjoint Analysis: A Monte Carlo Study," *Journal of Marketing Research*, 18, (February), 101-106.