# Challenges and Opportunities in High Dimensional Choice Data Analyses

| | |
|---|---|
| Prasad Naik* | University of California Davis |
| Michel Wedel* | University of Maryland |
| Lynd Bacon | Polimetrix Inc. |
| Anand Bodapati | University of California Los Angeles |
| Eric Bradlow | University of Pennsylvania |
| Wagner Kamakura | Duke University |
| Jeffrey Kreulen | IBM Almaden Research Center |
| Peter Lenk | University of Michigan |
| David Madigan | Rutgers University |
| Alan Montgomery | Carnegie Mellon University |

* Co-chairs

## Abstract

Modern businesses routinely capture data on millions of observations across subjects, brand SKUs, time periods, predictor variables, and store locations, thereby generating massive high dimensional datasets. For example, Netflix has choice data on billions of movies selected, user ratings, and geo-demographic characteristics. Similar datasets emerge in retailing with potential use of RFIDs, online auctions (e.g., eBay), social networking sites (e.g., mySpace), product reviews (e.g., ePinion), customer relationship marketing, internet commerce and mobile marketing. We envision massive databases as four-way VAST matrix arrays of Variables × Alternatives × Subjects × Time, where at least one dimension is very large. Predictive choice modeling of such massive databases poses novel computational and modeling issues, and the negligence of academic research to address them will result in a disconnect from the marketing practice and an impoverishment of marketing theory. To address these issues, we discuss and identify challenges and opportunities for both practicing and academic marketers. Thus we offer an impetus for advancing research in this nascent area and fostering collaboration across scientific disciplines to improve the practice of marketing in information-rich environment.

## 1. Introduction

Advances in computing power have revolutionized marketing practice. Companies not only possess massive database on billions of consumer choices, user ratings and geo-demographics, but also recognize that the predictive modeling of massive choice data offers potentially high return on investment. As firms become customer-focused, they amass customer databases by recording various aspects resulting from interactions with customers. A fundamental consequence of building massive databases to marketing is the shift in focus from reaching anonymous consumers (e.g., in mass marketing via television) to targeting identifiable customers (e.g., in direct marketing via online stores or services). In mass marketing, researchers have traditionally relied on data that came from a representative random sample of the population so that inferences could be drawn from that small sample to the population at large. In direct marketing, on the other hand, firms want specific insights at the customer level to enable them to develop marketing plans for each of its customers. To this end, we need models and methods that are scalable up to the entire customer base (e.g. Balasubramanian et al. 1998).

The size of choice datasets has been increasing due to the scanning devices in supermarkets since the nineteen seventies. In the late eighties, with the advent of loyalty cards, marketing companies expanded the details of customer-level information, which enabled a better targeting of marketing activities. Since the late nineties, due to the Internet-based technologies, marketers in diverse sectors such as retailing, direct marketing, banking, and telecom began to compile large databases on consumers' past transactions. More recently, due to the Internet becoming a marketing channel, companies augmented their databases with click-stream data and web-based transaction data. The basic nature of all these data sets, however, is numerical.

Besides numerical data, a qualitatively different *kind* of data known as “unstructured data” is increasingly compiled and used. It includes textual information from customer complaints, product reviews by other customers or expert opinions (ePinion, Netflix), social networking sites (e.g., mySpace), blogs, or patents (Google Patents). In the 21st century, due to advances in software engineering, marketers can use this unstructured information to “listen” to what customers have to say about one’s own existing products (or competitors’ products), and to identify gaps in the product space, trends in innovation or emerging technologies. In *Mining the Talk*, Spangler and Kreulen (2007) describe how to cull insights from textual data using illustrative business cases. In addition, new high-dimensional data sources continue to emerge due to RFID devices in stores and on products; eye-tracking data (e.g. Tobii) from natural settings such as PCs, TV, and billboards; gaming, video and metadata from semantic Web 2.0.

To summarize, marketing data are expanding in several modes; first, the number of Variables to explain choice behavior has greatly increased (large V); second, the number of choice Alternatives in packaged goods exploded to hundreds per product category (large A); third, automated data collection led to the recording of choice decisions from large samples of Subjects (large S); finally, the frequency of data collection on weekly, daily, or even hourly basis led to long Time series (large T). These VAST data matrices offer opportunities for improving the marketing analysis, planning, and implementation. But, they raise novel computational and statistical challenges. These challenges include the need to incorporate substantive issues such as consumer heterogeneity, cross-category preferences, spatial covariances, and dynamic and forward-looking aspects of choice behavior. To manage this enhanced complexity, practitioners resort to *simpler* models that are scalable to real-life size of the databases. On the other hand, to provide greater rigor and better explanation, academic researchers tend to *enrich* the models by focusing on a subset of variables, alternatives, subjects, or times (e.g., by sampling from or aggregating across one or more of the VAST modes). But the enriched models have limited applicability because they lack scalability required by marketing practitioners. This dilemma becomes even more salient in (i) real-time applications that technology-intense marketing environment facilitates (e.g. Google or Amazon), and (ii) decision-theoretic applications that require not only model estimation but also decision optimization. It points to the urgent need for faster numerical methods, efficient statistical estimators, and simpler models.

Some solutions to the challenges in massive data analyses are offered, singularly or in combination, by 1) increasing computational power, for example, through parallelization and grid computing; 2) developing scalable estimation methods via closed-form expressions for posterior distributions of parameters of marketing models or one-pass particle filters and other sequential MCMC methods and Variational inference; 3) creating alternative approaches for data analyses, for example, factor models, regularization methods, and inverse regression methods. The rest of the paper describes the sources of emerging data, then reviews the current computational solutions for analyzing massive data, and finally presents the recent advances in model estimation.

## 2. Sources of High Dimensional Data

The main phenomena that drive the emergence of novel massive databases are Web 2.0, virtual social networks, predictive markets, recommendation systems, and unstructured data, which we discuss below.

**Web 2.0.** Recent considerations for web standards focus on creating the "Semantic Web," an environment that supports reasoning (World Wide Web Consortium, 2005) and enables non-technical users to create their own content. As a result, web sites are moving away from a publishing (i.e., pushing of content) paradigm to a participatory functionality, where a large numbers of consumers can post product evaluations and comments (O'Reilly, 2005, Trendwatching.com, 2007). To extract insights from such consumer-generated data, companies like Umbria and Wize attempt to solve the challenges of locating the data and getting it into analyzable form. Alternatively, when consumers participate in web-based activities, they can produce more usable data and metadata by structuring differently what they do and earning rewards appropriately for their efforts. An example is in the generation of "usefulness" ratings for product reviews on Amazon.com. Such user- and site-related interactions create large databases that contain potentially valuable behavioral information which as yet not much research in marketing has addressed.

**Virtual Social Networks.** Social networks are a centerpiece phenomenon in Web 2.0. Online users generate content and create online activity for other users. Hence the audience base in social networks tends to be self-propelling or self-deflating. As for the social networking web site, it generates revenues via advertising by exposing advertisements to those who visit the site. From a managerial perspective, understanding who the users are and who keeps the social network site running is important for marketing strategies, in particular viral marketing. Who are the key players? What (or who) drives site visitation and therefore advertising revenues? How should the social network site segment its customers? How to advertise on social networking sites? To manage the site, to make it a better place for consumers, and to obtain better information for advertisers, the site's management needs to know exactly what role is being played by each individual user within the network and whose actions have an impact on whom. To shed light on these questions, massive databases on user activity, connectivity and referrals are available. Managers need modeling approaches to cull insights and identify site users who are influential in affecting a firm's primary revenue driver: other members' site usage. Recent research in marketing has begun to address these issues (e.g., Trusov, Bucklin, Pauwels 2007, Trusov, Bodapati, Bucklin  2007).

**Predictive Markets.** Predictive markets, sometimes called virtual stock markets, are gaining popularity (e.g. Iowa Electronic Markets or Hollywood Stock Exchange). Carnegie Mellon University offers an ESP game that mobilizes collective efforts to label images available on the Web. Such applications are "games" in the sense that there are rules and desired outcomes. Other game-like methods have been used for ideation and product development. Prelec (2001) describes a guessing game played over a network that can be used to develop customer descriptions of products and product concepts. Toubia (2006) describes a system for generating ideas in which incentives are aligned with performance. Bacon and Sridhar (2006) present online "innovation tools" that can be used to identify preferences and solve hard problems, including a game-oriented discussion system based on Kunz and Rittel's (1970) concept of an issues-based planning system that aligns rewards with incentives. Similarly, Ding, Park and Bradlow (2007) propose a bartering mechanism to collect data in conjoint settings. Foutz and Jank (2007) show how to use virtual stock market to make early forecasts of new product sales. The European Union plans to build an online environment for collaborative innovation (www.laboranova.com). Such purposeful and structured activities will produce rich data on preferences that can be modeled effectively if participant contributions are constrained appropriately, metadata are defined adequately, and scalable methods are developed rapidly.

**Recommendation Systems.** Product recommendation system based on collaborative filtering is the backbone of major Internet-based companies. For example, Netflix and

Blockbuster ask their customers to rate the movies they have viewed as a basis for generating recommendations; Amazon and Barnes & Noble implement the recommendation systems to suggest books to read; CDNow and iTunes use it to recommend music recordings; Findory and Tivo recommend news items and TV shows, respectively. A recent trend, however, is to shift this collaborative filtering from the companies, who do not utilize all the product ratings they collect to generate the recommendations, to stand-alone services. Such open source collaborative filtering systems enable consumers to update their own information about actual listening, reading, or viewing behaviors.

Following this industry practice, marketing literature has refined the methods for making recommendations; for example, Bayesian network models (Breese, Heckerman, and Kadie 1998), mixture models (Chien and George 1999, Bodapati 2007), hierarchical Bayes' models (Ansari, Essegaier, and Kohli 2000), and hierarchical Bayes selection models (Ying, Feinberg and Wedel 2004). These models show substantial improvements in the quality of recommendations on test datasets. However, these methods are not yet applicable to the entire customer database available with the companies because the estimation methods are computationally intensive. Besides scalability, recommendation systems need to incorporate vast choice alternatives, large-scale missing data, scale-usage heterogeneity, endogeneity of recommendations, and sensitivity to shilling attacks; these topics will be a fruitful area of future research (e.g., Chung, Rust and Wedel 2007).

**Unstructured Data.** Using database management systems, companies build numerical databases and apply data mining techniques to extract customer insights and gain competitive advantage. However, extracting information from text is under-utilized in corporate marketing and under-researched in academic marketing. For example, companies can learn how consumers view brands in a certain product category (i.e., brand positioning) by analyzing textual conversations obtained from chat rooms or blogs or social networking sites.

To analyze textual information, a first step is to create structure from this inherently unstructured textual data by classifying words into meaningful categories or taxonomies, i.e., natural hierarchy of information consistent with the business goals and processes (see, e.g., Kreulen, Spangler, Lessler 2003; Kreulen and Spangler 2005). Then, the resulting datasets, which are much larger than traditional (numerical) databases, can be analyzed by adapting statistical techniques for classification, prediction, and visualization to extract business intelligence and knowledge management (e.g., Kreulen, Cody, Spangler, and Krishna 2002). IBM has explored the application of these technologies in several domains including customer relationship management, enterprise content analysis, and World Wide Web analytics. For further details, see Spangler and Kreulen (2007).

## 3. Computational Issues

High-dimensional data leads to computational challenges because many algorithms do not scale linearly; they scale exponentially. A few select examples of computationally intensive research include the use of store transaction data to set optimal pricing strategies (Montgomery 1997), clickstream data for adaptive web design (Montgomery et al. 2004), or consumer choices for shopbot design (Brynjolfson, Smith, and Montgomery 2007). Although MCMC based procedures facilitated the estimation of intractable models, these techniques are time-consuming: the estimation of a hierarchical choice model could take hours to converge (Allenby, Rossi, and McCulloch 1996). In practice, companies such as Amazon or Google typically require its system to respond in 2 seconds or less. Hence fast computational methods are necessary. We next discuss two such methods: grid computing and specialized electronic devices.

**Grid computing** environments (Bryant 2007) harnesses the power of thousands of low-cost personal computers by splitting up the overall computational task into many smaller tasks that can be executed in parallel. Even techniques like MCMC that are inherently sequential can be parallelized. Specifically, Brockwell and Kadane (2005) propose a method of parallelizing these algorithms that relies on regeneration; Brockwell (2006) suggest a pre-fetching scheme for large-scale problems.

**Application Specific Integrated Circuit** (ASIC) and field-programmable gate array (FPGA) are specialized electronic devices that can increase computational speed (Brown and Rose 1996). Traditional microprocessors execute a set of instructions to perform a computation. By changing the software instructions, the functionality of the system is altered without changing the hardware. The downside of this flexibility is that the performance can suffer. The processor must read each instruction from memory, decode its meaning, and only then execute it. This results in a high execution overhead for each individual operation. The second method of algorithm execution is to use hardwired technology like ASICs, which are designed specifically to perform a given computation and so they are fast and efficient when executing the exact computation for which they were designed. However, the circuit cannot be altered after fabrication. FPGAs lie between these two extremes, being cheaper and more flexible than ASICs, while also being much faster than software programmable microprocessors. FPGAs contain an array of computational elements whose functionality is determined through multiple programmable configuration bits. For integer or fixed point arithmetic, FPGAs can improve speed by two orders of magnitude compared to a conventional cluster CPU.

In sum, future research needs to apply these computational methods to not only perform statistical tasks such as estimation, inference, prediction, but also for decision-theoretic optimization.

## 4. Models and Estimation Methods

Although massive databases pose computational challenges, is there a conceptual difference between small versus large datasets? The answer is affirmative. In a multivariate probability density function, as the number of variables increases, any local neighborhood is almost surely empty (i.e., sparseness of data cloud), whereas a non-empty neighborhood is almost surely not local (Simonoff 1996, p. 101). To understand this so-called empty space phenomenon, consider the standard normal density whose 68% of its total mass lies within ± 1 standard deviation from the origin. In a bi-variate normal, the mass within a unit square centered at the origin is about $(0.68)^2$, which is less than one-half of its total mass. As the number of variables $p$ increases, the mass within the unit hypercube, which equals $(0.68)^p$, tends to zero rapidly. With just ten variables (i.e., $p = 10$), the center of the density function —where the mode with the largest frequency is located— contains merely 2% of the sample.

Consequently, as Silverman (1986, p. 92) cautions, "large regions of high [dimensional] density may be completely devoid of observations in a sample of moderate size." In contrast to small datasets, most observations in large databases are outliers rather than representative. Thus, empty space phenomenon induces different conceptual issues when analyzing massive data; nonparametric models cannot be estimated accurately for 10 or more regressors and dimension-reduction methods are necessary to project the data into a low-dimensional subspace.

In many application areas, the move from small datasets to large datasets requires researchers to re-think, in a qualitatively different manner, how one approaches the problem of data analysis. In several cases, the data analysis is initially tackled using combinatorial approaches because that is how the problem is first conceptualized; with larger datasets the combinatorial approach often becomes computationally impracticable and one is forced to adopt model-based approaches where the researcher needs to postulate, in the abstract, the stochastic processes that could generate the dataset. As an example, for many decades, cluster analysis was approached as a combinatorial NP-hard problem, but researchers now approach the problem from the perspective of statistical mixture distributions (Wedel and Kamakura 2000). While the problem continues to be hard, the computational burden increases much more slowly with the size of the dataset. A similar story has played out with social network analysis: the early approaches in Wasserman and Faust (1994) are combinatorial in nature, whereas Handcock et al. (2007) offer model-based approaches that are computationally feasible for gigantic networks. Below we describe methods for analyzing high-dimensional data using factor-analysis, inverse regression, Bayesian methods, and regularization methods.

**Factor Analytic Approach.** Consider a retailer with data on sales and pricing for each of 30 brands across 300 customers for 36 months who wants to understand the (cross) effects of price on brand sales, while allowing the price elasticities to vary across stores and over time. An aggregate model would require the estimation of 900 price coefficients, which is feasible given the available data. But a random-coefficients model that accounts for unobserved heterogeneity would additionally require the estimation of 405,450 covariance terms. To transform this random-coefficients model into a feasible one while allowing for the possibility that the (cross)-elasticities might be correlated across stores and over time, Kamakura and Kang (2007) propose a factor-regression model where the covariance of the random regression coefficients is decomposed into its principal components across stores and over time. With a two-factor solution, the model would require 1,800 estimates for the covariance of the price coefficients, still large but more manageable than a full-covariance model.

A more traditional approach for reducing a high-dimensional problem to a more manageable one is by factor-analyzing the dependent variables (Wedel and Kamakura 2001). Du and Kamakura (2007), for example, propose a stochastic frontier factor model that takes advantage of the covariance structure among the dependent variables to define efficiency frontiers for each of multiple product categories based on the observed sales in all categories, as well as exogenous variables. But even if the "number of variables problem" is solved with some form of dimension-reduction model, it is still critical that the calibrated model can be implemented in the entire database. Simulation-based methods requiring many iterations per case are not as scalable as the simpler data-mining models based on singular-value decomposition. To this end, a pragmatic solution that has been employed is to calibrate the model on a random sample from the database, and then implement the estimated model on all the other cases. This approach, however, does not efficiently use all of the available data.

**Inverse Regression Methods.** A broad class of methods for dimension-reduction (see Naik, Hagerty and Tsai 2000) is based on the work by Li (1991) on Sliced Inverse Regression. Recent advances allow, for example, the extraction of confirmatory factors via Constrained Inverse Regression (Naik and Tsai 2005) or the estimation of models even if the number of regressors exceeds the number of observations via Partial Inverse Regression (Li, Cook and Tsai 2007). Next, we describe an application of inverse regression for nonparametric estimation of binary choice models for data with many variables.

For parametric (choice) models the sensitivity to the specification of the link function may be limited for small to moderate sized data, but it is likely much larger for massive data where if the pre-specified link departs from the true unknown link, the model cannot accurately predict rare events with low probability of occurrence. Nonparametric binary models overcome this drawback but break down when datasets contain ten or more regressors. Hence semi-parametric models have been proposed that relate an index (or a factor) —a linear combination of regressors— to the expected response via a nonparametric link function (Naik and Tsai 2004). Multi-Index Binary Response (MBR) model extend this idea to include more than one index and can be estimated using a non-iterative two-step method, thus allowing for flexibility of the link function and scalability to the large datasets (Naik, Wedel and Kamakura 2007). The two-step estimation method involves *projection* and *calibration*. In the *projection* step, information available in high-dimensional predictor space is combined and projected into a low-dimensional index space. Dimension reduction is achieved by estimating an index structure (i.e., the composition of indexes in terms of original regressors and the number of indexes to retain) without specifying the link function. In the *calibration* step, the unknown link function is estimated via local polynomial regression (or a parametric model). One of the main advantages of the MBR model is that, by allowing for multiple indexes, it facilitates a more refined understanding of consumer behavior.

In an application to predict customer defection for a telecom company, Naik, Wedel and Kamakura (2007) formulated the Multi-factor Binary Choice model, where they reduced dimensionality from 124 predictors to 4 factors using a new method for high-dimensional data analysis known as sliced average variance estimation (Cook and Weisberg 1991). Having reduced dimensionality markedly, in the resulting low-dimensional subspace they deploy local polynomial regression —which would be practically impossible in the original predictor space —to visualize the regions where customers least or most likely to defect reside and to discover that attractive regions of prospective customers need not be contiguous or connected.

**Bayesian Methods.** As previously mentioned, VAST datasets pose computational problems in Bayesian inference due to its computationally intensive procedures. Specifically, MCMC algorithms recursively generate the model's parameters from the full conditional distributions given the data. Standard implementation of MCMC requires holding all the data in memory or repeatedly reading from the data file in every iteration of the MCMC chain. Because these approaches are not feasible with massive datasets, DuMouchel (1999) uses empirical Bayes (EB) to analyze very large frequency tables. EB can be viewed as a large sample approximation to hierarchical Bayes models where the parameters for the population-level distributions are estimated by non-Bayes methods, such as maximum likelihood, often after marginalizing over individual-level parameters. Ridgeway and Madigan (2002) proposed a fully Bayesian method that first performs traditional MCMC on a subset of the data, then applies importance sampling and re-sampling to the remainder of the data. Balakrishnan and Madigan (2006) modified this method by applying particle filters (Gordon, Salmond, and Smith 1993; Liu and Chen 1998) to obtain a method that requires only a single pass through the data.

One approach to reduce computational costs relies on the property of conditional independence of observations. One may divide the data into manageable blocks and recursively analyze each block by using the posterior distribution from the previous block as the "updated prior" for the next one. The recursion continues until all the data has been processed. Whereas this idea can be directly implemented if the all prior and posterior distributions are conjugate, for many models this is not the case and here approximations need to be used. This idea needs further work, but holds promise for estimating mixtures of Dirichlet processes (Escobar and West 1996) to describe heterogeneity distribution, which is of interest in marketing (Wedel and Zhang 2005). Huang and Gelman (2006) have also pursued related ideas.

Another approach to reduce computational cost is to derive analytical approximations. Recently, Bradlow, Hardie and Fader (2002), Everson and Bradlow (2002), and Miller, Bradlow, and Dayartna (2006) offer closed-form Bayesian inference to models previously thought to be non-conjugate problems. Bradlow and colleagues demonstrate this approach for commonly used models in marketing such as the Negative Binomial, Beta-Binomial, and heterogeneous Logit model. The general approach is to (i) approximate the likelihood function by a polynomial with a user-specified number of terms in the summation (which could be a very large number, even thousands to get the desired accuracy), (ii) find a conjugate prior to the approximated likelihood function, and (iii) integrate term-by-term the "posterior polynomial" since the approximated likelihood and prior are now conjugate.

Finally, as an alternative to MCMC methods, variational approximations have recently attracted considerable attention, for example, see Wainwright and Jordan (2003) and the references therein. Like MCMC, variational inference methods have their roots in statistical physics; unlike MCMC, they are deterministic. The basic idea of variational inference is to formulate the computation of marginal or conditional probability as an optimization problem, perturb the optimization problems (i.e., induce "variations"), and find solutions to the perturbed problems (see Jordan et al. 1998). The application of variational inference to VAST marketing databases is as yet unexplored.

**Regularization Methods.** Regularization methods offer a promising route to tackle the problem of many variables. For example, Tibshirani (1996) developed the least absolute shrinkage and selection operator (LASSO) regression, which estimates the coefficients and selects the variables simultaneously. It minimizes the usual sum of squared errors in the linear regression model subject to constraining the sum of the absolute values of the coefficients. Fan and Li (2001) developed SCAD (smoothly clipped absolute deviation) regression, which correctly eliminates with probability one all the variables whose true effects are negligible (i.e., the so-called oracle property). Bayesian variant of LASSO regression places a Laplace prior on the coefficients of the regression model (Genkin, Lewis and Madigan 2007; Balakrishnan and Madigan 2007).

As an example of "many" variables, consider the monitoring of the safety of licensed drugs (Hauben et al. 2005). To detect novel drug-adverse event associations, patients, physicians, and drug companies submit reports containing drug(s), adverse event(s), and basic demographic information. Because most reports contain multiple drugs and multiple adverse events, complex multi-drug interactions can exist. About 15,000 drugs are available in the US market; coincidentally, the dictionary of adverse events employed by the FDA also contains about 15,000 unique adverse events. Thus, the challenge reduces to performing a multivariate regression with a 15,000-dimensional response variable and 15,000 input variables. By including just two-way interactions, one obtains over 100,000,000 input variables, which is well beyond the capabilities of current algorithms (see Hauben et al. 2005 for more details). Regularization methods, dimension-reduction methods (i.e., factor analysis, inverse regression), and their Bayesian variants hold the promise to provide effective solutions for VAST data.

## 5. Conclusion

In methodological research, mathematics used to reign supreme. Historically, data existed primarily on paper, and useful inferences owed their existence to deft, computation-avoiding mathematics. In the mid-1960's, diagonalizing a 7 x 7 matrix represented a significant computational challenge. Stunning progress has occurred since then leading some, for example, Benzécri (2005), to claim that in data analysis "there is no longer any problem of computation." But two interrelated events have utterly changed the computational landscape: ubiquitous computing and ubiquitous databases. In terms of bringing models to data, extraordinarily exciting times await us. Massive data present challenging marketing applications and will alter our perspectives on the computational scale of things to come. In addition, failure to address these problems posed by VAST data will potentially lead to an impoverished marketing theory and will disconnect us from the marketing practice. Because a single solution to all computational challenges does not exist, a combination of computational, statistical and algorithmic approaches are needed to solve specific problems. The VAST dimensions of massive data will require different approaches as well: solutions to handling massive samples (vaSt) will differ from those dealing with high frequency data (vasT), which will differ from those addressing massive numbers of variables (Vast). To address these challenges expeditiously, marketing scholars may benefit from collaborations with different disciplines such as statistics and computer science. We hope this article sets an agenda and stimulates progress in marketing with VAST databases.

## 6. References

Allenby, Greg M., Robert McCulloch, and Peter E. Rossi (1996), "The Value of Purchase History Data in Target Marketing", *Marketing Science*, 15, 321-340.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), "363-75.

Bacon, L. & Sridhar, A. (2006), "Interactive Innovation Tools and Methods." *Annual Convention of the Marketing Research Association*, Washington D.C., June 2006.

Balakrishnan, S. and Madigan, D. (2006), "A One-Pass Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets," *Bayesian Analysis*, 1 (2), 345-362.

Balakrishnan, S. and Madigan, D. (2007). LAPS: Lasso with Partition Search. *Manuscript*.

Balasubramanian, Sridhar, Sunil Gupta, Wagner A. Kamakura, and Michel Wedel (1998), "Modeling large datasets in marketing," *Statistica Neerlandica*, 52 (3), 303-324, S

Benzécri, Jean-Paul (2005), "Foreword," In *Correspondence Analysis and data coding with JAVA and R* by Fionn Murtaugh, Chapman & Hall/CRC Press, London, UK.

Bodapati, Anand (2007), "Recommendation Systems with purchase Data," Journal of Marketing Research, forthcoming.

Bradlow, E.T., Hardie, B.G.S., and Fader, P.S. (2002), "Bayesian Inference for the Negative Binomial Distribution Via Polynomial Expansions," *Journal of Computational and Graphical Statistics*, 11 (1), 189-201.

Breese, Jack, David Heckerman, and Carl Kadie (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publisher.

Brockwell, A.E. (2006), "Parallel Markov Chain Monte Carlo Simulation by Pre-Fetching", *Journal of Computational and Graphical Statistics*, 15 (1), 246-261.

Brockwell, A.E. and J.B. Kadane (2005), "Identification of Regeneration Times in MCMC Simulation, With Application to Adaptive Schemes", *Journal of Computational and Graphical Statistics, 14 (2), 436-458.*

Brown, Stephen and Jonathan Rose (1996), "Architecture of FPGAs and CPLDs: A Tutorial", *IEEE Design and Test of Computers*, 13 (2), 42-57.

Bryant, Randal E. (2007), "Data-Intensive Supercomputing: The case for DISC", Carnegie Mellon University, School of Computer Science, Working Paper CMU-CS-07-128.

Brynjolfson, Erik, Michael Smith, and Alan Montgomery (2007), "The Great Equalizer: An Empirical Study of Choice in Shopbots", Carnegie Mellon University, Tepper School of Business, Working Paper.

Chien, Yung-Hsin and Edward I. George (1999), "A Bayesian Model for Collaborative Filtering," *Working Paper*, University of Texas at Austin.

Chung, Tuck, Siong, Roland Rust and Michel Wedel (2007), "My Mobile Music: Automatic Adaptive Play-list Personalization," *Working Paper*, Robert H. Smith School of Business, University of Maryland.

Cook, R. D. and Weisberg, S. (1991), "Discussion of Li (1991)," *Journal of the American Statistical Association*, 86, 328-332.

Ding, Min, Young-Hoon Park, Eric Bradlow (2007), "Barter Markets", *Working Paper*, The Wharton School.

Du, Rex and Wagner A. Kamakura (2007) "How efficient is your category management? A stochastic-frontier factor model for internal benchmarking", *Working Paper*

DuMouchel, W. (1999), "Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System," *The American Statistician*, 53 (3), 177-190.

Escobar, M. D. and West, M. (1996), "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 1996, 90, pp. 577-588.

Everson, P.J. and Bradlow, E.T. (2002), “Bayesian Inference for the Beta-Binomial Distribution via Polynomial Expansions,” *Journal of Computational and Graphical Statistics*. 11 (1), 202-207.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Foutz, N. Z. and W. Jank (2007), “Forecasting New Product Revenues via Online Virtual Stock Market,” MSI Report.

Genkin, A., Lewis, D.D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, to appear.

Gordon, N., Salmond, D., and Smith, A. F. M. (1993), “Novel approach to nonlinear/non-gaussian Bayesian state estimation,” *IEE Proc. -F Radar, Sonar Navig*., vol. 140, 107–113.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model-based clustering for social Networks. *Journal of the Royal Statistical Society. Series A*, 170(2), 301-352.

Hauben, M., Madigan, D., Gerrits, C., and Meyboom, R. (2005). The role of data mining in pharmacovigilance. *Expert Opinion in Drug Safety*, 4(5), 929-948.

Huang, Z. and Gelman, A. (2006). Sampling for Bayesian computation with large datasets. http://www.stat.columbia.edu/~gelman/research/unpublished/comp7.pdf

Jordan, M.I., Z. Ghahramani, T.S. Jaakkola, and L.K. Saul (1998), “An Introduction to variational methods for graphical models”, in *Learning Graphical Models*, Vol 89 of *Series D: Behavioural and Social Sciences,* Dordrecht, The Netherlands, Kluwer. pp. 105-162.

Kamakura, Wagner A and Wooseong Kang (2007) “Chain-wide and Store-level Analysis for Cross-Category Management,” *Journal of Retailing* 83(2) 159-70.

Kreulen, Jeffrey and W. Spangler. (200x), “Interactive Methods for Taxonomy Editing and Validation*,” Next Generation of Data-Mining Applications*, ISBN 0-471-65605-4, Chapter 20, pp. 495-522.

Kreulen, Jeffrey, W. Cody, W. Spangler and V. Krishna (2002), “The Integration of Business Intelligence and Knowledge Management,” *IBM Systems Journal*, Vol. 41, No. 4, 2002.

Kreulen, Jeffrey, W. Spangler and J. Lessler, (2003), “Generating and Browsing Multiple Taxonomies over a Document Collection,” *Journal of Management Information Systems*, 19 (4), 191-212.

Kunz, W. and Rittel, H. (1970), “Issues as Elements of Information Systems." Berkeley: Institute of Urban and Regional Development, Working Paper 131.

Li, K.-C. (1991), “Sliced Inverse Regression for Dimension Reduction,” *Journal of the American Statistical Association*, 86, 316-342.

Li, Lexin, R. D. Cook and C.-L. Tsai (2007), “Partial Inverse Regression,” *Biometrika*, forthcoming .

Liu, J. S. and R. Chen, R. (1998), “Sequential Monte Carlo methods for dynamical systems,” *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044.

Miller, S.J., Bradlow, E.T., and Dayartna, K. (2006) “Closed-Form Bayesian Inferences for the Logit Model via Polynomial Expansions”, *Quantitative Marketing and Economics,* 4 (2), 173-206.

Montgomery, Alan L, Shibo Li, Kannan Srinivasan, and John Liechty (2004), “Modeling Online Browsing and Path Analysis Using Clickstream Data”, *Marketing Science*, 23 (4), 579-595.

Montgomery, Alan L. (1997), “Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data”, *Marketing Science*, 16 (4), 315-337.

Naik, P. A. and C.-L. Tsai. 2004. “Isotonic Single-Index Model for High-Dimensional Database Marketing,” *Computational Statistics and Data Analysis*, 47 (4), 775-790.

Naik, P. A. and C.-L. Tsai. 2005. Constrained Inverse Regression for Incorporating Prior Information,” *Journal of the American Statistical Association*, 100 (469), 204-211.

Naik, P. A., M. Hagerty, and C.-L. Tsai. 2000, “A New Dimension Reduction Approach for Data-Rich Marketing Environments: Sliced Inverse Regression,” *Journal of Marketing Research*, 2000, 37 (1), 88-101.

Naik, P. A., M. Wedel and W. Kamakura (2007), “Multi-Index Binary Response Model for Analysis of Large Datasets,” *working paper,* UC Davis.

O’Reilly, T. (2005), “What is Web 2.0: Design patterns and business models for the next generation of software.” http://www.oreillynet.com/lpt/a/6228.

Prelec, D. (2001), “Readings Packet on the Information Pump,” MIT Sloan School of Management, Boston, MA.

Rhodes, James, Stephen Boyer, J. Kreulen, Ying Chen, and Patricia Ordonez, “Mining Patents Using Molecular Similarity Search,” *Pacific Symposium on Biocomputing* 2007, pp. 304-315.

Ridgeway, G. and Madigan, D. (2002), “A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets.” *Journal of Knowledge Discovery and Data Mining*, 7, 301-319.

Silverman, B. W. (1986), *Density Estimation*, London, U. K.: Chapman & Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics,* New York, N. Y.: Springer.

Spangler, S. and J. Kreulen (2007). *Mining the Talk: Unlocking the Business Value in Unstructured Information*, IBM Press.

Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*,58 (1), 267-288.

Toubia, O. (2006), “Idea generation, creativity, and incentives.” *Marketing Science*, 25(5), 411-425.

Trendwatching.com (2007), “Transparency,” http://trendwatching.com/briefing/.

Trusov, Michael, Anand Bodapati and Randolph E. Bucklin (2007), “Determining Influential Users in Internet Social Networks,” *Working Paper*, Robert H. Smith School of Business, University of Maryland.

Trusov, Michael, Randolph E. Bucklin and Koen Pauwels (2007), “Estimating the Dynamic Effects of Online Word-of-Mouth on Member Growth of a Social Network Site,” *Working Paper*, Robert H. Smith School of Business, University of Maryland.

Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference. *Technical Report 649*, Department of Statistics, UC Berkeley.

Wasserman, S. and Faust, K, (1994). *Social Network Analysis*, Cambridge University Press.

Wedel, Michel and Jie Zhang (2003), “Analyzing Brand Competition Across Subcategories,” *Journal of Marketing Research***,** 41 (4), 2004, 448-456.

Wedel, Michel and Wagner A. Kamakura (2001), “Factor Analysis with Mixed Observed and Latent Variables in the Exponential Family,” *Psychometrika*, 66 (4), pp. 515-30.

Wedel, Michel and Wagner Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, Dordrecht: Kluwer, 2nd ed.

World Wide Web Consortium (2005), “The Semantic Web.” www.w3.org/2005/Talks/1111-Delhi-IH/.

Ying, Y., F.Feinberg., and M.Wedel (2006), “Improving Online Product Recommendations by Including Nonrated Items,” *Journal of Marketing Research***,** 43 (August), 355-365.