Marketing Science

Successive Sample Selection and Its Relevance for Management Decisions

Stephan Wachtel, Thomas Otter,

Please scroll down for article—it is on subsequent pages

Successive Sample Selection and Its Relevance for Management Decisions

Stephan Wachtel, Thomas Otter

Goethe University Frankfurt, 60323 Frankfurt, Germany
{swachtel@marketing.uni-frankfurt.de, otter@marketing.uni-frankfurt.de}

We reanalyze endogenous sample selection in the context of customer scoring, targeting, and influencing decisions. Scoring relies on ordered lists of probabilities that customers act in a way that contributes revenues, e.g., purchase something from the firm. Targeting identifies constrained sets of covariate patterns associated with high probabilities of these acts. Influencing aims at changing the probabilities that individual customers act accordingly through marketing activities. We show that successful targeting and influencing decisions require inference that controls for endogenous selection, whereas scoring can proceed relatively successfully based on simpler models that provide (local) approximations, capitalizing on spurious effects of observed covariates. To facilitate the type of inference required for targeting and influencing, we develop a prior that frees the analyst from having to specify (often arbitrary) exclusion restrictions for model identification a priori or to explicitly compare all possible models. We cover exclusions of observed as well as unobserved covariates that may cause the successive selections to be dependent. We automatically infer the dependence structure among selection stages using Markov chain Monte Carlo-based variable selection, before identifying the scale of latent variables. The adaptive parsimony achieved through our prior is particularly helpful in applications where the number of successive selections exceeds two, a relevant but underresearched situation.

Key words: Bayesian estimation; targeting; cross-sectional analysis; scoring; causal reasoning; variable selection; sample selection
History: Received: September 17, 2010; accepted: September 14, 2012; Eric Bradlow and then Preyas Desai served as the editor-in-chief and Fred Feinberg served as associate editor for this article. Published online in *Articles in Advance* December 17, 2012.

1. Introduction

Many business processes of interest to marketing are characterized by multiple stages that involve a sequence of consumer decision outcomes. For example, some customers who receive direct mail respond, asking for additional information, and others do not react. Only a subset of the responsive customers end up making a purchase, and again, only a subset of purchasing customers repurchase. Another example for successive, outcome-based sample selection is a survey that maps respondents onto one of a set of ordered stages such as brand awareness, brand consideration, brand purchase, and repeat purchase. A third example is marketing on the Internet, where usually a very large initial base of prospects is winnowed down to a relatively small number of successes, such as in the case of search engine marketing. Here, the stages of the selection process are searching using a particular keyword (or a member of a large set of keywords), clicking on the sponsored link, and deciding to purchase from the sponsor.

1.1. Scoring, Targeting, and Influencing Customers
Organizations invest in monitoring such multiple-stage processes or conducting surveys with built-in

stages with the goal to answer the following questions: (a) How likely are individual customers to eventually act in a way that positively contributes to revenues such as, e.g., purchasing from the firm? (b) How are promising (new) customers selected, based on constrained sets of covariate patterns? (c) What actions are likely to nudge individual customers toward acting in a way that positively contributes revenues to the firm? We will refer to (a) as *scoring*, to (b) as *targeting*, and to (c) as *influencing*.

The goal of influencing is to change the response probabilities of individual consumers or customers through changes in marketing mix variables such as the number of direct mails sent, the characteristics of quotes, or variations in the level of a service. In contrast, both scoring and targeting condition on values of variables related to the probability of passing a particular stage or all stages; i.e., they do not attempt to change individual response probabilities.

The goal of customer scoring is to rank the vectors of x variables describing individual customers according to their probability of eventually acting in a particular, economically relevant, way (e.g., Cao and Gruca 2005). In practical applications, the predictions often refer to primary demand states such as,

e.g., the likelihood of needing consumer credit. The firm action then aims at translating this likely primary demand by, e.g., current account holders into secondary demand for the firm's offerings, implicitly acknowledging that alerting customers to the firm's offer has no impact on the primary demand state.[1] The goal of targeting is to select a subset of (new) consumers based on covariate patterns found to be associated with (usually) high probability of success.

The difference between scoring and targeting is subtle but important. Scoring eventually selects based on probabilities, with no particular regard to the resulting pattern in the covariates. For useful targeting rules, in contrast, the set of different covariate patterns to target has to be suitably small and frequently reflects trade-offs between estimated probabilities for particular covariate levels and the practical implementation of targeting rules, e.g., when mailings are sent to a particular demographic or when media decisions are made based on demographic readership, listenership, or viewership information. In the extreme, particular levels of covariates are completely excluded, such as in the decision to exclude a particular geography from targeting actions.

Thus, when scoring, the decision maker is neutral between different covariate patterns predicting the same probabilities. When targeting, the decision maker will prefer some covariate patterns over others at the same predicted probability for reasons beyond the consumer response model. The decision maker's interest in particular restricted sets of covariates requires that model-based predictions extrapolate well into the tails of the distribution of observed covariates. Any restricted multivariate definition of targets in a high-dimensional space of covariates almost invariably refers to some tail in the observed multivariate distribution of covariates. Interesting populations for targeting, i.e., consumers with relatively extreme response probabilities, are, by definition, found in these tails.[2] Alternatively, specialized media to reach different targets are characterized by an audience that is in the tail of the observed covariate distribution.

To continue the consumer credit example from above, the firm may want to go beyond its current database comprising, e.g., current account holders and select media to advertise its consumer credit offerings. The optimal choice of media depends, among other things, on the firm's ability to correctly predict probabilities associated with restricted sets of covariates, such as when advertising in magazines read by younger, well-to-do male consumers that live in medium-sized or large cities.

1.2. Models for Successively Selected Samples

The analysis required to inform scoring, targeting, and influencing decisions invariably involves some type of regression with the success or failure of consumers on the stages as dependent variables. And even though it is well known that ignoring endogenous sample selection in this step can result in biased inferences (e.g., Maddala 1983), popular approaches in practice analyze stages independently (e.g., Riesenbeck and Perrey 2009, Williams 2010) or collapse the multistaged process to one stage when measuring probabilities to pass all stages.[3] Although common in practice, such naïve analysis can result in substantial problems for targeting and influencing, but, as we will show, it may be sufficient for scoring.

A stream of literature in econometrics discusses inference from such endogenously selected samples. Comprehensive surveys are provided by Maddala (1983), Vella (1998), and Li and Prabhala (2007). Recent marketing papers build on these results (e.g., Cao and Gruca 2005, Chib et al. 2004, Ghose and Yang 2009, Lee et al. 2004, Zhao et al. 2009).

A limitation of existing models is that they provide limited guidance for the likely case where prior knowledge about what covariates are effective on what stage is lacking (Li and Prabhala 2007). Papers in marketing that include all covariates on all stages (e.g., Chib et al. 2004, Ghose and Yang 2009) and papers that rely on exclusion restrictions with little discussion about their specification (e.g., Cao and Gruca 2005, Lee et al. 2004) testify that prior knowledge about what covariates are effective on what stage (and not effective on another stage) is often scarce.

The empirically apparent difficulty of specifying exclusion restrictions a priori can be explained from a basic feature of selection models: in the absence of detailed knowledge about the concrete (micro)mechanisms operating at the various selection stages, prior knowledge is limited to ruling out covariates that originate at later stages (and thus did not even exist at prior stages) as causes for previous selections, e.g., when direct mails are only sent

---

[1] Scoring is a purely intellectual exercise if it pertains to the (stable) probability of the focal outcome, i.e., adopting the firm's offering in this example. However, ideally, firms eventually also learn about any differences in secondary demand among customers that are predicted to exhibit primary demand, i.e., are "in the market" by analyzing the heterogeneous response to its specific offering.

[2] Scoring is interested in extreme probabilities too. However, scoring by definition does not commit to a particular tail of the covariate distribution as the source of the probabilities of interest because scoring relies on lists of ordered probabilities with no regard to patterns of associated covariates.

[3] "Collapsing" therefore treats all failures as if they had occurred in the last stage of the process.

to customers that have bought a particular product. Obviously, the direct mail could not have caused the condition that triggered itself, i.e., the purchase of that product or earlier selections.[4]

Consider the example of advertising, in contrast. Exposure to an advertising campaign increases brand awareness through attentional and memory mechanisms. However, exposure to the campaign may or may not increase brand consideration and purchase likelihood (conditional on awareness). The mechanisms underlying brand consideration and purchase intention are relatively more complex. It is possible that a specific ad campaign not only increases brand awareness but also consideration and purchase likelihood, whereas another ad campaign only increases awareness or even decreases consideration. Without a detailed micro model of the interaction between consumers and specific ads, the presence and the direction of effects beyond increasing awareness are not determined a priori.[5]

These limitations on prior knowledge complicate inference from selected samples more than inference from standard samples because selection models that enter all covariates on all stages while controlling for selection effects are poorly identified, even conditional on parametric assumptions (e.g., Greene 2002, Li and Prabhala 2007). In the following we refer to models that enter all covariates on all stages while controlling for selection effects and therefore proceed without exclusion restrictions, as saturated models.[6] We illustrate how estimating such a saturated model results in biased inferences about what covariates are effective on what stage and the degree of dependence across stages in Web Appendix 1 and in our illustrative case study. Therefore, fitting a saturated model is unlikely to provide useful insights into what covariates are effective on what stage. In contrast, unknown parameters in models that a priori assume independence between selection stages or even collapse the

multistaged process to one stage are relatively well identified from data. This may have contributed to the prevalence of naïve analytical approaches in practice.

Another limitation of existing models for selected samples is that virtually all published models are developed for only two successive outcomes, whereas many applications include more than two stages. Finally, the existing literature is explicit about the inferential problems posed by selected samples but leaves the correspondence between model requirements and decision support in the context of scoring, targeting, and influencing decisions implicit.

In this paper, we provide a model that generalizes existing models of selection across two stages to a potentially large number of successive selections. In our model, we address the practically relevant situation where prior knowledge about exclusion restrictions is limited. The generalization to potentially large numbers of successive selections is facilitated by Wold's decomposition (Wold 1954) of the unobserved dependence structure across selection stages.[7] Based on this decomposition, we generalize exclusion restrictions to include the unobserved dependence structure.

The technique presented promises to be useful in the general context of potentially correlated systems of equations, beyond the successive selection problem discussed in this paper. The ability to learn automatically about a potentially parsimonious structure of dependence relationships improves inference when, as usual, prior knowledge about unobserved dependence relations is limited, and estimating full, unstructured dependence results in a weakly identified model. Examples are random coefficient regression models (e.g., Frühwirth-Schnatter and Tüchler 2008), (latent) correlated factor models (e.g., Erdem 1996), and the multivariate probit model (e.g., Keane 1992).

Finally, we clarify the relation between the different management goals of scoring, targeting, and influencing and modeling requirements relating the concept of sample selection to unobserved heterogeneity. We show that successful targeting and influencing actions require (posterior) knowledge about the pattern of active and excluded variables on each selection stage, whereas scoring can work well based on (simple) independent analyses that capitalize on spurious relationships. In our illustrative application we find that ignoring endogenous selection reverses the relative attractiveness of targets living in West and East Germany and biases inference about how age,

---

[4] Moreover, we show in Web Appendix 2 (available at http://dx.doi.org/10.1287/mksc.1120.0754) that for strong identification of unobserved dependence, we need covariates to selection that can be excluded from the equations for subsequent outcomes. Unobserved dependence is identified from variation in selection probabilities holding current-stage probabilities constant. Selection probabilities can only be varied independent from current-stage probabilities when the previous selection is based on at least one covariate that is excluded from the current stage. Additional covariates originating at later stages, although excluded from previous selections, are insufficient for strong identification of unobserved dependence.

[5] Li and Prabhala (2007) describe the common situation in two-stage models where no observed covariate can be reasonably excluded from the selection equation a priori, and therefore no prior exclusion restrictions can be formulated in a two-stage model.

[6] The list of observed covariates could include nonlinear transformations of covariates to capture, e.g., potential interaction effects. In our examples we, however, consider main effects only.

[7] Wold's decomposition originated as the (infinite) moving average representation of a covariance-stationary time series (Wold 1954). The time ordering in our application is given by the ordering of the selection stages.

household income, and gender are associated with brand consideration and purchase.

The remainder of this paper is organized as follows. Section 2 provides basic notation and reanalyzes the selection problem using directed acyclic graphs (DAGs) that explain the link between different management decisions and modeling requirements. Section 3 develops our framework and Bayesian inference using Markov chain Monte Carlo (MCMC). Section 4 illustrates our method analyzing the data from a large-scale survey investigating brand awareness, consideration, purchase, and repeat purchase for fashion brands. We offer a concluding discussion and suggestions for future research in §5.

2. Reanalysis of the Selection Problem

Equation (1) introduces general notation for a selection model across K successive stages for consumers $i = 1, \ldots, N$. The observed K-dimensional outcome $\mathbf{y}_i$ is functionally related to underlying continuous latent variables $\mathbf{z}_i$ to accommodate limited dependent variables. In anticipation of our case study, and without loss of generality, we discuss the case where individual elements of $\mathbf{y}_i$ are Bernoulli variables taking values of 1 when the selection stage is passed and 0 when it is failed. The relationship to latent variables $\mathbf{z}_i$ is such that an element of $\mathbf{y}$ is equal to 1 only if the corresponding element of $\mathbf{z}$ *and all elements of $\mathbf{z}$ corresponding to prior selection stages* are larger than 0 and equal to 0 otherwise. This characteristic of selection models implies that failures, i.e., values of 0 after the first failure in an ordered sequence of selections, contribute no additional information. For example, the observation that someone excludes a brand from his consideration set contributes no additional information about why, if he is not aware of the brand:

$$\mathbf{y}_i = f(\mathbf{z}_i),$$

$$\begin{pmatrix} z_{1,i} \\ z_{2,i} \\ \vdots \\ z_{K,i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1,i} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2,i} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_{K,i} \end{pmatrix}' \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix}$$

$$+ \begin{pmatrix} \lambda_{11} & 0 & \ldots & 0 \\ \lambda_{21} & \lambda_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{K1} & \lambda_{K2} & \ldots & \lambda_{KK} \end{pmatrix} \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ \vdots \\ \eta_{Ki} \end{pmatrix}, \quad (1)$$

$$\mathbf{z}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\Delta} \boldsymbol{\eta}_i,$$

$$\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I}_K).$$

The goal is to infer, from a sample of consumers with varying response patterns, what drives the probabilities of advancing to the next stages. Ordered lists

of inferred probabilities are the basis for scoring, and both targeting and influencing build on the inferred relationships between observed covariates and success on the stages of the selection process. Observed predictors $\mathbf{X}_i$ could include relatively stable consumer characteristics such as demographics or marketing mix variables experienced at various stages. When, as is often the case, only limited prior knowledge about what observed variables drive success on what stage (and not on other stages) is available, the configuration of stage-specific design matrices $\mathbf{x}_{1i}, \ldots, \mathbf{x}_{Ki}$ becomes a variable in the inferential process.

The scalars η_{1i} through η_{Ki} represent unobserved causes to successes and failures in the selection process. Coefficients $\{\lambda\}$ on the diagonal of the lower-triangular matrix $\boldsymbol{\Delta}$ measure their scale.[8] Nonzero coefficients $\{\lambda\}$ below the diagonal introduce unobserved causes emerging on previous stages as causes on subsequent stages, creating dependence among selections. Another way to see this is to derive the covariance of $\mathbf{z}_i$ integrating out $\boldsymbol{\eta}_i$; i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\mathbf{I}\boldsymbol{\Delta}'$.[9] When all lower-triangular elements of $\boldsymbol{\Delta}$ are different from 0, outcomes on stages 1 through K constitute $K, K-1, \ldots$, and two repeated measurements for $\eta_{1i}, \eta_{2i}, \ldots$, and $\eta_{K-1,i}$, respectively.[10] Thus, the scalars η_{1i} through η_{Ki} coupled with nonzero $\boldsymbol{\Delta}$ introduce correlated heterogeneity in intercepts.

In general, prior knowledge about the structure of the covariance or equivalently that of coefficients $\{\lambda\}$ is absent. The potential structures range from independence where all off-diagonal elements $\{\lambda\}$ are equal to 0, to full unstructured dependence where all elements in the lower-triangular matrix $\boldsymbol{\Delta}$ are different from 0. Next we reanalyze the selection problem using DAGs. Our reanalysis explains the problems from a lack of prior exclusion restrictions in detail and links the analysis of selected samples to managerial goals of scoring, targeting, and influencing.

It is well known that independent analyses of outcomes in successively selected samples generally result in inconsistent inferences, unless $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}'$ is diagonal, which is the case if λ_{21} in Equation (1) is 0 in a model with two stages; see, e.g., Heckman's (1979) seminal paper. The nature and the precise cause of the inconsistency is illustrated by the DAGs in Figures 1(a) and 1(b) for a model with two successive

---

[8] In the case of Bernoulli outcomes, these scales and regression coefficients $\{\boldsymbol{\beta}\}$ are not jointly identified. We discuss identification in §3.

[9] $\boldsymbol{\Delta}$ is the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}$. The restriction that $\boldsymbol{\Delta}$ is a lower-triangular can be motivated from the technical requirement that the implied $\boldsymbol{\Sigma}$ needs to be positive definite or from the substantive argument that unobservables that only emerge on later selection stages should be excluded as causes from previous selections, a priori.

[10] η_{Ki} only affects the stage K and thus cannot be distinguished from "pure" stage-specific error.

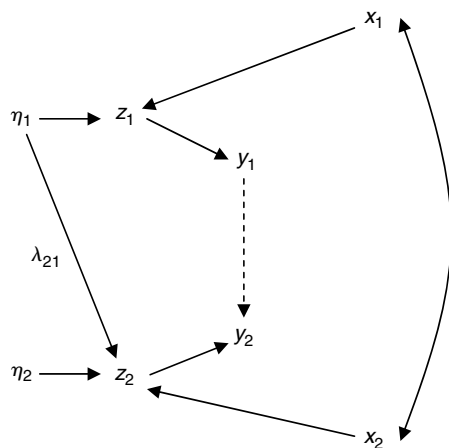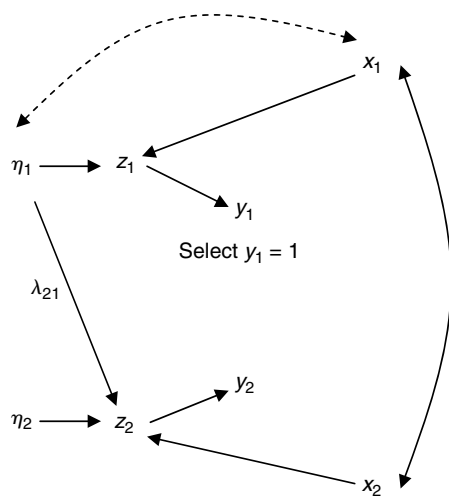**Figure 1(a)   DAG for Two Successive Selections**

Figure 1(b)   Independent Analysis of Second Stage

selections and response patterns $[0, 0]'$, $[1, 0]'$, and $[1, 1]'$. Curved double arrows represent covariances. The dashed arrow connecting y_1 and y_2 in Figure 1(a) represents the selection through $p(y_2 = 0 \mid y_1 = 0) = 1$.

Figure 1(b) illustrates the independent analysis of y_2, i.e., the second stage, when selecting is based on $y_1 = 1$. The conditioning on the outcome $y_1 = 1$ introduces covariance between x_1 and η_1 in the selected sample as indicated by the curved dashed double arrows. The covariance is due to restricting the sum of these variables, i.e., $z_1 = x_1\beta_1 + \eta_1$, to the range $z_1 > 0$ by selecting only first-stage successes.

The induced covariance between x_1 and η_1 connects x_1 with z_2 through λ_{21}. It also opens a second connection between x_2 and z_2 through the covariance between x_2 and x_1, x_1 and η_1, and finally, the effect from η_1 on z_2. In turn, this causes inconsistent variable selection and inconsistent estimates of regression coefficients in independent analysis of the selected sample on the second stage. Inconsistent variable selection causes x_1 to be included as a significant

covariate on the second stage and is due to the link between x_1 and z_2 in the selected sample.[11] Inconsistent estimates of regression coefficients are due to the additional link between x_2 and z_2 in the selected sample.[12]

At this point, the choice between dependent and independent analysis seems obvious in the sense that dependence should always be considered as a default. A question is then how to practically accomplish this default. Consider the situation where we lack any prior (causal) knowledge about observed covariates, i.e., that x_2 only affects y_2 but not y_1, and unobserved covariates, i.e., that λ_{21} is different from 0. In this case we face three potential mean structures for y_1, i.e., $\{x_1\}$, $\{x_2\}$, and $\{x_1, x_2\}$, times the same three mean structures for y_2, times the possibility that λ_{21} is equal or different from 0 for a total of 18 different models.[13] Remembering that the saturated model is unlikely to be informative about the data-generating structure because of weak empirical identification in all but extremely large samples[14] (Greene 2002; see also Web Appendices 1 and 2), we need to explicitly compare these models, which is tedious but still manageable. However, with, e.g., $K = 3$ selection stages and $p = 8$ observed covariates, there are already $2^{K \cdot p} = 16{,}777{,}216$ different models for the means and $2^{(K-1) \cdot K \cdot 0.5} = 8$ different dependence patterns for a total of $134{,}217{,}728$ different models. A direct comparison of these models, each with potentially different implications for management decisions is no longer practically possible (see the example in Web Appendix 1).

---

[11] Conditioning on some $\lambda_{21} \neq 0$ when the data-generating mechanism has $\lambda_{21} = 0$ also leads to inconsistent variable selection. Consider, for example, positive λ_{21} and a positive effect of x_1 on the first stage. Observations low on x_1 need more positive error draws to pass the first stage. These then erroneously imply relatively larger success on the second stage because of positive λ_{21}. To correct for this spurious imbalance between "low x_1" and "high x_1" first-stage successes, x_1 will be entered with a positive effect on the second stage. This problem is regularly encountered when estimating saturated models that include all covariates on all selection stages while controlling for dependence. The "conditioning" on $\lambda_{21} \neq 0$ comes through the multivariate likelihood that is unbounded for increasing error correlations.

[12] Interestingly, removing the link between x_1 and x_2 enables consistent inferences about the slope coefficients connecting x_2 and y_2 from a selected sample, even if $\lambda_{21} \neq 0$. This is because the additional link between x_2 and z_2 comes through the covariance between x_1 and x_2. However, the practical value of this result is limited to the case where exclusion restrictions are known a priori.

[13] We implicitly condition on functional form assumptions that rule out higher-order effects or interactions in this example.

[14] The complete argument for empirical identification only based on functional form, and without exclusion restrictions, requires in addition that stage-specific regression coefficients are linearly independent and that the data are informative enough to rule out linear dependence (see Web Appendix 2).

We now turn to the implications of the DAG analysis for managerial decision making. First, it is apparent that actions aimed at influencing customers that have passed the first stage may be completely misguided unless successive selection and the possibility of unobserved dependence is taken into account when selecting active variables on the second stage and, generally, after any number of previous selections. After selection has occurred, causal manipulations of variable x_1, spuriously included as the predictor on the second stage, translate into a null effect. An example is advertising that increases awareness only but may spuriously appear to also affect consideration. In this case, increasing the level of advertising to consumers that are aware of the brand already has no effect on demand.

It is well known that the success of causal interventions is especially sensitive to inference about the model structure. However, in the case of selected samples, this inference needs to extend to the effects $\{\lambda\}$ of unobservables $\boldsymbol{\eta}_i$ that a priori, i.e., before selection, are orthogonal to observed covariates. Moreover, in substantive research, usually little if anything is known a priori about the dependence structure among successive selections. Dependence between successive selections is generally viewed as a phenomenon that complicates inference and has to be controlled for without much substantive interest in dependence itself.

Second, spuriously including x_1 on the second stage effectively captures part of the unobserved dependence between the successive selections and therefore actually helps when the goal is to only score consumers that make it to the second stage in a stable environment. The spurious variable x_1 then functions as a proxy for the unobserved random effect η_1 in the independent analysis of the second stage. We confirm a surprisingly good performance of independent analyses in the context of scoring numerically in Web Appendix 3. Spuriously included variables on later stages of the selection process proxy well for unobserved dependence, as evidenced by high overlap between estimated lists of high-scoring consumers and those from data-generating models with various degrees of unobserved dependence.[15]

Third and finally, the implications for targeting decisions are least obvious but given the prevalence of targeting, in practice, are probably most relevant. Targeting decisions aim at bringing new consumers to the first stage that are likely to pass the first

and second stages. Independent analysis of the first stage, i.e., analysis before selection, results in consistent inferences about covariates to success on that stage. The direction of the bias incurred when independently analyzing the second stage depends on the sign of the unobserved dependence between stages.

We first investigate the case where the unobserved dependence is positive such that "lucky" passes on the first stage, i.e., passes that are unlikely based on observed covariates, are predictive of lucky passes on the second stage. In this case, x_1 values associated with a low likelihood of success on the first stage will be coupled with relatively more positive draws of unobservables that link the selection stages, in the selected sample. Consumers that unexpectedly pass the first stage are relatively more likely to also pass the second stage. Conversely, consumers with a high probability of passing the first stage will appear as relatively less successful on the second stage.

Taken at face value, this situation implies an involved analysis of economic trade-offs between the seemingly opposing effects of x_1 on the first and the second stages. However, according to the data-generating mechanism, there is no such trade-off because x_1 is orthogonal to the random effect η_1 before selection.[16] If the choice between a consumer likely to be successful on stage 1 and another consumer less likely to be successful based on the observable x_1 is free, the former should always be chosen. That consumer is more likely to preserve the option of a favorable draw of unobservables that originate on stage 2 by surviving the first selection. Web Appendix 3 provides a numerical illustration. If prior knowledge about effects of observed covariates and unobserved covariates, i.e., dependence, is limited, a viable data-based approach is needed to decide if targets high on x_1 are universally better (at equal costs) or if a there is a trade-off between expected success on the first stage and that on the second stage.

In the case of negative unobserved dependence across selection stages, consumers that pass the first stage with a lucky draw of the random effect are relatively less likely to succeed on the second stage. This translates into a positive dependence between values of x_1, spelling success on the first stage and success on the second stage, in the selected sample. The relative benefits of targeting based on x_1 will be overestimated. We will revisit this scenario in our case study.

Overall, the DAG analysis implies that the more active the approach to marketing, from the description of (arriving) prospects and existing customers in a stable environment, i.e., scoring, to actively targeting subsets of arriving customers based on

---

[15] As an aside, the popular practice of "collapsing" the successive selection stages to one stage, treating all failures as if they occurred on the last stage, even fails at scoring. This is because collapsing ignores the stochastic nonlinearity introduced through selection (see Web Appendix 3).

[16] Note that respondents with relatively large x_1 values are in the tail of the observed covariate distribution.

suitably simple covariate patterns, to finally allocating the marketing mix to influence customers, the more important a careful consideration of different model structures becomes. The next section develops a model that facilitates the consideration of different model structures.

3. Model Development

We propose to alleviate the inferential problem in selection models by formally acknowledging any lack of prior knowledge about exclusion restrictions and the finer structure of unobserved dependence across selection stages, including the possibility of independence. We do so by formulating a prior distribution over all possible exclusion restrictions. This set considers both exclusions of observed ($\mathbf{x}$) and unobserved ($\boldsymbol{\eta}$) covariates and spans from a model of uncorrelated intercepts to full unstructured dependence among selection stages with all observed covariates present on all stages. By doing so, we first shift the burden of formulating dogmatic prior beliefs about specific exclusions to the weaker statement that some exclusions may be supported by the data. Second, by allowing for exclusions of unobserved covariates ($\boldsymbol{\eta}$), we avoid biases in the exclusions of observed variables resulting from an overparameterized dependence structure.[17] An alternative approach to our prior is to estimate all models implied by different patterns of exclusions separately and then compare or average these models. As demonstrated in §2, this approach is impractical in all but the smallest problems. We proceed by first discussing exclusion restrictions for unobserved causes that result in parsimonious models of dependence. We then present the general approach and highlight specifics of Bayesian estimation.

3.1. Parsimonious Models of Dependence

Recall that dependence across selections in Equation (1) is implied by the coefficients $\{\lambda\}$; i.e., the covariance of $\mathbf{z}_i$ integrating out unobserved causes $\boldsymbol{\eta}_i$ is $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}'$. Setting off-diagonal coefficients $\{\lambda\}$ equal to 0, we can create various parsimonious models of dependence, including the special case of independence. Parsimonious dependence structures improve inference about the mean structure, and vice versa. Equations (2a)–(2c) show three illustrative examples for a problem with four selection stages:

$$\begin{pmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ 0 & \lambda_{32} & \lambda_{33} & 0 \\ 0 & 0 & \lambda_{43} & \lambda_{44} \end{pmatrix}, \qquad (2a)$$

$$\begin{pmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & 0 & \lambda_{33} & 0 \\ \lambda_{41} & 0 & 0 & \lambda_{44} \end{pmatrix}, \qquad (2b)$$

$$\begin{pmatrix} \lambda_{11} & 0 & 0 & 0 \\ 0 & \lambda_{22} & 0 & 0 \\ 0 & 0 & \lambda_{33} & 0 \\ 0 & 0 & 0 & \lambda_{44} \end{pmatrix}. \qquad (2c)$$

In (2a), the unobservable η_{1i} impacts the outcome on the first and second stage; i.e., it introduces direct dependence between the first and second stages, but is excluded as a direct cause from the third and fourth stages. Similarly, η_{2i} and η_{3i} are restricted to directly impact only the outcome on the corresponding stage as well as the stage immediately following, respectively. This implies that for unbiased, consistent inferences about covariate effects on the fourth (third) stage, only conditioning on probabilities of passing the third (second) stage is required. The probabilities of passing earlier stages can be ignored. Therefore, no observed covariates need to be excluded from stage 4 relative to stages 1 and 2 for strong identification.

In (2b), all dependence between successive selection stages comes through unobserved causes emerging on stage 1. Unobserved causes η_{2i} through η_{3i} are excluded from all respective following stages.[18] The implication is that, e.g., a bivariate model for stages 1 and 4 is sufficient for unbiased, consistent inferences about effects of observed covariates on these stages. The analysis uses all observations on stage 1 as well as all observations on the fourth stage, acknowledging that successes on the first stage that never made it to the fourth stage are uninformative about what generated the data on stage 4. Because of a lack of connections through unobservables between stages 2, 3, and 4, failures on stages 2 and 3 simply generate a random, exogenous selection of first-stage successes. Analogously, bivariate models for stages 1 and 2 as well as stages 1 and 3 produce unbiased, consistent inferences about effects of observed covariates on these stages in example (2b). Therefore, strong identification only requires exclusions of observed covariates relative to stage 1, conditional on this dependence structure. Finally, example (2c) is the independence model that excludes all unobserved causes from influencing outcomes on other stages.

Under the restriction that $\boldsymbol{\Delta}$ is a Cholesky factor of $\boldsymbol{\Sigma}$, the exemplary finer structures (2a) and (2b) in the coefficients $\{\lambda\}$ and the implied exclusion restrictions for unobserved causes $\boldsymbol{\eta}$ are generally only meaningful when the causal/temporal ordering of successive

---

[17] Relying on experimental design, Zeithammer and Lenk (2009) show how excluding unobserved covariates that create dependence in a multinomial probit model results in improved inference.

[18] Note that example pattern (2b) generally implies a full covariance matrix with all entries different from 0. For this reason, selecting covariances directly is not generally useful.

selection stages is preserved—hence, our reference to the Wold decomposition in time-series analysis. For example, when we rearrange selection stages in example (2b) in the order 3, 4, 1, and 2, the causal structure of this example can only be preserved if nonzero values in $\boldsymbol{\Delta}$ are allowed below and above the diagonal:

$$\begin{pmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & 0 & \lambda_{33} & 0 \\ \lambda_{41} & 0 & 0 & \lambda_{44} \end{pmatrix} \begin{pmatrix} \eta_{1,i} \\ \eta_{2,i} \\ \eta_{3,i} \\ \eta_{4,i} \end{pmatrix}$$

$$\xrightarrow[(1,2,3,4) \xrightarrow{\text{reorder}} (3,4,1,2)]{} \begin{pmatrix} \lambda_{33} & 0 & \lambda_{31} & 0 \\ 0 & \lambda_{44} & \lambda_{41} & 0 \\ 0 & 0 & \lambda_{11} & 0 \\ 0 & 0 & \lambda_{21} & \lambda_{22} \end{pmatrix} \begin{pmatrix} \eta_{3,i} \\ \eta_{4,i} \\ \eta_{1,i} \\ \eta_{2,i} \end{pmatrix} \quad (3)$$

Moreover, the lower-triangular Cholesky factor that corresponds to the reordered $\boldsymbol{\Sigma}$ is, in general, no longer sparse; i.e., it contains no zeros below the diagonal. Overall, an ability to detect parsimonious dependence structures fundamentally helps when making inference from selected samples and more generally in systems of dependent equations, because parsimonious dependence structures are strongly identified with fewer exclusions of observed covariates.

3.2. A Prior Distribution Over Exclusion Restrictions

Rewriting Equation (1) as in Equation (4) shows that the search for exclusion restrictions is a variable selection problem (see George and McCulloch 1997 and Dellaportas et al. 2002 for overviews). In Equation (4) all observed covariates are present in $\mathbf{x}_i'$ on each stage, and exclusion restrictions correspond to zero elements in the vector of regression coefficients:

$$\begin{pmatrix} z_{1,i} \\ z_{2,i} \\ \vdots \\ z_{K,i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i' & \mathbf{0} & \cdots & \mathbf{0} & \eta_{1i} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \mathbf{0} & \mathbf{x}_i' & \cdots & \mathbf{0} & 0 & \eta_{1i} & \eta_{2i} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_i' & 0 & 0 & 0 & \cdots & \eta_{1i} & \cdots & \eta_{Ki} \end{pmatrix}$$
$$\cdot (\boldsymbol{\beta}_1' \; \boldsymbol{\beta}_2' \; \cdots \; \boldsymbol{\beta}_K' \; \lambda_{11} \; \lambda_{21} \; \lambda_{22} \; \cdots \; \lambda_{K1} \; \cdots \; \lambda_{KK})^{\mathrm{T}}. \quad (4)$$

We summarize the model structure implied by a set of exclusion restrictions by the indicator vector $\boldsymbol{\tau}$. It is of length equal to the maximum dimensionality of the model, i.e., the model with all observed variables entered on all selection stages and full, unstructured, unobserved dependence across stages. If a particular element of $\boldsymbol{\tau}$ takes the value 0 (1), the corresponding variable is excluded (included) at this position. Elements of $\boldsymbol{\tau}$ that correspond to diagonal elements of $\boldsymbol{\Delta}$, i.e., $\lambda_{11}, \lambda_{22}, \ldots, \lambda_{KK}$, are always included in the model.

We introduce α as the hierarchical prior probability that a particular element of the indicator vector

$\boldsymbol{\tau}$ is equal to 1, and the corresponding covariate is included in Equation (4). We complete our prior over model structures by a noninformative Beta(1, 1) prior for α. The Beta prior is the conjugate prior to the probability α. Estimating α instead of fixing it to, say, 0.5 is important for two reasons. First, with multiple successive selections and larger sets of covariates, any fixed value α translates into an informative prior for the total number of exclusions. Second, recognizing uncertainty in α improves the numerical performance of the MCMC algorithm described in the appendix. We note that our noninformative prior on the pattern of exclusion restrictions can and should be adapted when (partial) prior knowledge about exclusions is available. In this case the corresponding elements of the indicator vector $\boldsymbol{\tau}$ can be fixed to the values implied by theory. Also, the subjective Beta prior and, consequently, α can be stage-specific or used to differentiate a priori between the probability of excluding observed and unobserved covariates.

3.3. Bayesian Inference

Recall that both $\mathbf{z}_i$ and $\boldsymbol{\eta}_i$ in Equations (1) and (4) are unobserved. Augmenting the data with latent $\mathbf{z}_i$ is desirable because we obtain a conditionally linear model (Albert and Chib 1993). Augmenting unobserved $\boldsymbol{\eta}_i$ allows us to treat $\{\lambda\}$ as regression coefficients. However, conditional on $\mathbf{z}_i$ and $\boldsymbol{\eta}_i$, Equation (4) is deterministic, i.e., a degenerate likelihood in $\{\boldsymbol{\beta}\}$ and $\{\lambda\}$. To overcome this problem, we add independent standard normally distributed errors to the $\mathbf{z}_i$:

$$\begin{pmatrix} z_{1,i} \\ z_{2,i} \\ \vdots \\ z_{K,i} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i' & \mathbf{0} & \cdots & \mathbf{0} & \eta_{1i} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \mathbf{0} & \mathbf{x}_i' & \cdots & \mathbf{0} & 0 & \eta_{1i} & \eta_{2i} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_i' & 0 & 0 & 0 & \cdots & \eta_{1i} & \cdots & \eta_{Ki} \end{pmatrix}$$
$$\cdot (\boldsymbol{\beta}_1' \; \boldsymbol{\beta}_2' \; \cdots \; \boldsymbol{\beta}_K' \; \lambda_{11} \; \lambda_{21} \; \lambda_{22} \; \cdots \; \lambda_{K1} \; \cdots \; \lambda_{KK})^{\mathrm{T}}$$
$$+ \begin{pmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \\ \vdots \\ \varepsilon_{K,i} \end{pmatrix}, \quad \boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_K). \quad (5)$$

The implied covariance matrix $\boldsymbol{\Sigma}$ is then no longer $\boldsymbol{\Delta}\boldsymbol{\Delta}'$ but $\boldsymbol{\Delta}\boldsymbol{\Delta}' + \mathbf{I}$. This is innocuous in our case because the variance of latent variables $\mathbf{z}$ is not likelihood identified from discrete observations $\mathbf{y}$. The likelihood only identifies ratios of stage-specific regression coefficients $\{\boldsymbol{\beta}\}$ and the square root of diagonal elements in $\boldsymbol{\Sigma}$, i.e., the stage-specific standard deviations. Therefore, the likelihood identifies correlations between selection stages, and any correlation pattern may be achieved based on $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}' + \mathbf{I}$.

The MCMC algorithm outlined next navigates the (likelihood) unidentified space. We margin down

to the likelihood-identified space postprocessing the MCMC output (McCulloch and Rossi 1994, Edwards and Allenby 2003).

Our algorithmic approach is similar to Brown et al. (1998), Frühwirth-Schnatter and Tüchler (2008), and Tüchler (2008). Brown et al. (1998) introduce variable selection based on the marginal likelihood for multivariate regression problems. Frühwirth-Schnatter and Tüchler (2008) develop a variable selection prior for both the mean and the hierarchical covariance structure in linear random coefficient models. Tüchler (2008) generalizes the approach to random coefficient logistic models. We develop this later work for application to selection models where parsimonious Cholesky factors are structurally meaningful because of their Wold interpretation.

An appealing feature of the approach we adopt relative to the shrinkage priors of George and McCulloch (1993, 1995) is that the point hypothesis that an effect should be excluded—and not just shrunk to a small value—is well defined without (likelihood) identifying the trade-off between $\{\beta\}$ and $\{\lambda\}$. In this model it is essential that variable selection can proceed in the unidentified space because valid identification constraints, e.g., all diagonal elements in Σ equal to a constant, cannot be expressed through restrictions on Δ prior to knowing its structure.[19]

The MCMC algorithm has the following five basic steps:

Step 1. $\{z\} \mid \{\beta\}, \{\lambda\}, \{\eta\}, \tau, \{y\}$.
Step 2. $\tau \mid \{z\}, \{\eta\}, \alpha$.
Step 3. $\{\beta\}, \{\lambda\} \mid \{z\}, \{\eta\}$.
Step 4. $\{\eta\} \mid \{\beta\}, \{\lambda\}, \{z\}$.
Step 5. $\alpha \mid \tau$.

All sampling steps, except Step 2, are implemented as standard Gibbs samplers. Step 2 uses an independence Metropolis–Hastings step (see Rossi et al. 2005). We sample the elements of τ from the posterior integrated by $\{\beta\}$ and $\{\lambda\}$. This (marginal) posterior is proportional to the marginal likelihood of the latent variables $\{z_i\}$ given a particular model as defined by what variables are included and excluded on what stage, times its (hierarchical) prior probability. Integration by $\{\beta\}$ and $\{\lambda\}$ is crucial for the sampler to work because of the perfect dependence between the regression coefficients $\{\beta\}$, $\{\lambda\}$, and the indicators in τ when the latter take the value 0. We provide a detailed derivation of the (marginal) posterior required for this step and other sampling details in the appendix.

In Step 3 we draw from the unconstrained joint posterior density of $\{\beta\}$ and $\{\lambda\}$. We margin down to the likelihood identified space by computing $\Delta\Delta' + I$, i.e., Σ at every draw from the unconstrained posterior, transforming into a correlation matrix and adjusting the coefficients $\{\beta\}$ accordingly to obtain likelihood identified parameters.

We illustrate the performance of our algorithm using simulated data in Web Appendix 1. We find that simply estimating a saturated model in the hope of seeing which effects matter is likely to result in spurious regression coefficients and extreme correlations between selection stages, even if selection stages are conditionally independent (see also Footnote 11 and Greene 2002).

4. Illustration: Successive Sample Selection in a Large-Scale Survey

Our empirical illustration uses data from the OUTFIT4 survey. OUTFIT is a large-scale periodic survey conducted by Spiegel publishers among the German consumer population. According to Spiegel publishers, the OUTFIT surveys are one of the major strategic planning instruments for manufacturers and retailers of apparel in Germany. The sampling population consists of 50.9 million German consumers between the ages of 14 and 64. The random sample obtained for OUTFIT4 contains 8,359 consumers. Interviews were conducted by mail and face-to-face between April and June 1997. OUTFIT tracks general attitudes toward various characteristics of apparel and other fashion accessories as well as the awareness, consideration, purchase, and repurchase of major apparel brands in Germany, e.g., Lacoste. *Purchase* is defined as having purchased the brand at least once, and *repurchase* is defined as having purchased the brand coupled with a stated willingness to repurchase the brand. In addition, the survey contains demographic information about age, gender, education, household income, the number of wage-earning members in a household, the size of the household, whether the head of the household is responding, city size, and whether the household lives in the former West Germany (FRG) or in the former East Germany (GDR).

We study the hypothetical but realistic problem of targeting prospects for the brand Lacoste based on demographic information for the purpose of alerting them to a new, vertically integrated e-commerce channel. Operationally, such targeting usually proceeds by identifying postal codes that, based on the distribution of demographic variables across postal codes, are more likely to be responsive. The distribution of demographic information across postal codes can be obtained from commercial data providers.

---

[19] The direct selection of correlations in this context is less useful because even sparse Δ can result in a full covariance matrix (see example (2b) in §3.1).

Marginally, 74.2% of respondents are aware of the brand, 35.6% consider the brand (47.9% of aware respondents), 23.7% have purchased the brand at least once (66.6% of considering respondents), and 17.9% have purchased the brand repeatedly (75.5% of purchasing respondents). The key to targeting is understanding how the demographic descriptors of households relate to the probabilities of passing the successive selection awareness, consideration, purchase, and repurchase. Choosing a particular pattern of covariates as marking a target does not in itself causally change the probabilities of the targets. Targeting allocates information to consumers that might have some use for this information in the context of their existing predispositions. In our example, it would not make much sense to alert someone to Lacoste's new e-commerce channel, if he or she is unlikely to consider the Lacoste brand after having become aware of it. However, alerting someone to Lacoste's new e-commerce channel may create instant awareness of Lacoste as an apparel brand. It may also increase purchase and repurchase probabilities further downstream.

Our decision problem is illustrative, and the data do not contain information about any actual campaigns. Therefore, we cannot measure the extent to which the action of alerting customers to the new channel causally changes the response probabilities of individual recipients of the information. However, the data are sufficient to answer the potentially practically more relevant targeting question. The assumptions we make are that informing any consumer about Lacoste's new channel may (i) at best create instant awareness of the Lacoste brand and at worst not have any impact on awareness, and (ii) at best increase the probability of a purchase and repurchase and at worst not have any impact.

We first investigate posterior evidence of unobserved dependence among selection stages causing the selections to be endogenous. We then compare statistical inference that takes unobserved dependence into account to that from simple independent analyses that ignore endogenous selection. We also show empirical evidence of identification problems in the saturated model that controls for dependence but does not exclude variables. Finally, we discuss implications for targeting.

4.1. Inference About Unobserved Dependence

The posterior distribution of the model indicator vector τ shows that the independence model is straightforwardly rejected by the data. Overall, we have six off-diagonal λ-coefficients that measure dependence in our application. In the posterior, models that include six, five, four, three, two, one, and none of these off-diagonal λ-coefficients are visited with probabilities 12.8%, 24.8%, 28.9%, 31.6%, 1.9%, 0%,

Table 1 Posterior Means of Error Correlations (Identified Σ) Across Selection Stages: Dependence Model with Variable Selection (DEP)

	Awareness	Consideration	Purchase	Repurchase
Awareness	1			
Consideration	**−0.933**	1		
	(0.038)			
Purchase	−0.074	**0.369**	1	
	(0.124)	**(0.146)**		
Repurchase	0.052	−0.059	−0.052	1
	(0.207)	(0.219)	(0.157)	

Note. Posterior standard deviations are in parentheses.

Table 2 Posterior Means of λ-Coefficients Connecting η and z

	Awareness (η_1)	Consideration (η_2)	Purchase (η_3)	Repurchase (η_4)
Awareness (z_1)	**0.99**	0	0	0
	(0.003)			
Consideration (z_2)	**−0.94**	**0.32**	0	0
	(0.04)	**(0.10)**		
Purchase (z_3)	−0.07	**0.93**	0.26	0
	(0.12)	**(0.07)**	(0.20)	
Repurchase (z_4)	0.05	−0.04	−0.03	**0.95**
	(0.21)	(0.18)	(0.13)	**(0.08)**

Note. Posterior standard deviations are in parentheses.

and 0%, respectively.[20] Therefore, the independence model has zero support in our MCMC sample from the posterior.[21]

Table 1 presents the correlations across stages. Correlations that are credibly different from 0 appear in boldface. We find a strong negative correlation between unobserved causes of awareness and consideration of Lacoste caused by a significant link from η_1 to z_2 (see Table 2). We also find a weaker positive correlation between unobserved causes to consideration and purchase of Lacoste caused by a significant link from η_2 to z_3 (see Table 2). All other unobserved dependence relationships are equal to 0. That is, respondents that are unexpectedly aware of Lacoste are relatively less likely to consider it than predicted from their observed covariates. Respondents that are unexpectedly considering Lacoste are somewhat more likely to purchase Lacoste.

---

[20] The λ-coefficients on the diagonal are always included; i.e., the respective elements of τ are fixed to 1.

[21] The exact posterior probability of the independence model is, of course, strictly positive but sufficiently small to be of no practical relevance. Similarly, the saturated model has zero support in our MCMC sample. The posterior is represented by 250,000 draws obtained by saving every second draw after burning off the first 500,000 iterations. We ran the MCMC multiple times initialized at different τ patterns, initially including (τ-start = 1) or excluding (τ-start = 0) all coefficients {β} and {λ}, and we obtained essentially equivalent results.

Table 3   Coefficients from the DEP and INDEP Analyses, and the SAT Model

| | Before selection | | | After selection(s) | | | | | | | | |
| | Awareness | | | Consideration | | | Purchase | | | Repurchase | | |
Variable	D	I	S	D	I	S	D	I	S	D	I	S
Constant	**2.08** (**0.11**)	**2.05** (**0.12**)	**2.13** (**0.12**)	**0.17** (**0.09**)	**0.56** (**0.14**)	**0.61** (**0.21**)	−0.15 (0.16)	−0.08 (0.11)	0.38 (0.26)	0.53 (0.30)	**0.75** (**0.17**)	0.67 (0.31)
Female	0	0	−0.04 (0.04)	**−0.28** (**0.03**)	**−0.34** (**0.03**)	**−0.33** (**0.04**)	**−0.21** (**0.07**)	0	−0.07 (0.07)	0	0	−0.08 (0.08)
Age	**−0.01** (**0.001**)	**−0.01** (**0.001**)	**−0.01** (**0.001**)	0	**−0.005** (**0.001**)	**−0.01** (**0.002**)	0	0	−0.001 (0.002)	0.007 (0.004)	0	**0.01** (**0.002**)
Education	**0.13** (**0.02**)	**0.14** (**0.02**)	**0.14** (**0.02**)	−0.02 (0.02)	0	−0.02 (0.02)	**0.12** (**0.03**)	**0.12** (**0.02**)	**0.14** (**0.02**)	0	0	**−0.09** (**0.03**)
HH income	**0.06** (**0.01**)	**0.06** (**0.01**)	**0.06** (**0.01**)	0	**0.03** (**0.007**)	**0.03** (**0.01**)	**0.07** (**0.01**)	**0.06** (**0.01**)	**0.07** (**0.01**)	0	0	−0.01 (0.01)
No. of wage earners in HH	−0.06 (0.05)	−0.08 (0.06)	**−0.10** (**0.03**)	0	0	−0.01 (0.04)	0	0	−0.06 (0.04)	0	0	−0.04 (0.05)
Size of HH	−0.05 (0.03)	−0.03 (0.03)	**−0.05** (**0.02**)	0	0	**−0.04** (**0.02**)	−0.04 (0.03)	0	**−0.05** (**0.02**)	−0.05 (0.04)	0	−0.05 (0.03)
Head of HH	0	0	0.02 (0.04)	0	0	0.01 (0.05)	0	−0.14 (0.08)	−0.01 (0.05)	0.18 (0.11)	0	**0.22** (**0.08**)
Size of city	**0.08** (**0.01**)	**0.08** (**0.01**)	**0.08** (**0.01**)	**−0.04** (**0.01**)	0	−0.01 (0.02)	0	0	**0.06** (**0.01**)	−0.05 (0.04)	0	**−0.09** (**0.02**)
Lives in the former GDR	**−1.42** (**0.04**)	**−1.44** (**0.04**)	**−1.43** (**0.04**)	**0.61** (**0.05**)	−0.1 (0.07)	0.05 (0.29)	−0.11 (0.13)	0	**−0.90** (**0.12**)	0	0	**0.58** (**0.12**)

Notes. Posterior standard deviations are in parentheses. D, DEP; I, INDEP; S, SAT; HH, household.

4.2.   Inference About the Mean Structure

Table 3 compares the regression coefficients obtained from our proposed model to those from independent analyses ($\lambda_{ij} = 0$ for all $i \neq j$) of the various stages. Independent analyses are the default in practice when exclusion restrictions are not obvious a priori and controlling for endogenous selection results in an empirically underidentified model. Exclusion restrictions, i.e., marginal inclusion probabilities below 0.5, are indicated by a coefficient equal to 0. Included coefficients credibly different from 0 appear in boldface.[22] The third parameter column reported for each stage shows parameter estimates from the *saturated* (SAT) model. We comment on these results after comparing our proposed model to *independent* analyses.

The proposed *dependent* (DEP) analysis almost perfectly agrees with the *independent* (INDEP) analysis predicting awareness. This is because awareness is observed prior to any selection. The results imply that younger, better educated, wealthier consumers

---

[22] An alternative posterior summary reports the posterior at the most likely model defined by the τ-pattern visited most often by the MCMC. Although our data essentially rule out the independence model as well as the saturated model as data-generating mechanisms, they do not point to one dominating modal model. The modal model is visited in 1,047 of 250,000 MCMC iterations with an implied posterior probability of only 0.5%. We therefore prefer the reported marginal summaries to a summary of the posterior at the modal model.

who live in larger cities are more likely aware of the Lacoste brand. The large negative coefficient associated with living in the former East Germany, i.e., the former GDR, suggests that these consumers are generally less likely to be aware of the Lacoste brand than their West German counterparts. As to be expected from the error correlations in Table 1, DEP and INDEP analyses produce different covariate patterns predicting consideration and purchase. Both models agree that repurchase cannot be predicted with the data at hand. We discuss the drastically different results from the SAT model below.

DEP analysis implies that females and consumers living in larger cities are less likely to consider Lacoste and that consumers residing in the former GDR are substantially more likely to consider Lacoste. INDEP analysis agrees on the negative effect of being female but spuriously includes a negative influence of age and a positive effect of household income on consideration. INDEP analysis also misses the negative effect size of city and the substantial positive influence of residing in the former GDR on consideration.

Recall the strong negative dependence between the awareness stage and the consideration stage from Table 1. It implies that consumers that are unexpectedly aware of the Lacoste brand, i.e., require a positive influence of unobservables to pass the awareness stage successfully, are less likely to pass

the consideration stage, everything else equal. The large negative effect of living in the former GDR on awareness implies that some consumers unexpectedly aware of the Lacoste brand, as by the model, may actually live there. Failing to take into account the unobserved heterogeneity in η_1 influencing both awareness and consideration, i.e., z_1 and z_2, INDEP analysis biases the estimate of the GDR fixed effect and other fixed effects on the consideration stage.

The direction of the observed biases in the coefficients follows from our discussion in §2. Negative (positive) unobserved dependence, if ignored, results in spuriously carrying forward effects from the previous stage maintaining (changing) their sign. Consequently, older people who are less likely to be aware of Lacoste, everything else equal, spuriously appear to also consider Lacoste less likely when analyzed independently. Similarly, wealthier consumers who are more likely to be aware of Lacoste, everything else equal, spuriously appear also more likely to consider Lacoste. In the case of city size and living in the former GDR, the structural coefficients as estimated by DEP analysis are in the opposite direction of the influence, positive and negative, respectively, spuriously carried forward by INDEP analysis. Therefore, the negative effects of city size and the positive effect of living in the former GDR on consideration are biased upward and downward, respectively, in the INDEP analysis of consideration.

In the case of the GDR coefficient, INDEP analysis results in a borderline significant sign reversal of the effect. That is, INDEP analyses suggest that the Lacoste brand may generally be a worse fit for consumers in the former GDR. In contrast, DEP analysis implies that consumers living in the former GDR—and predicted to be aware of the Lacoste brand based on other covariates—are actually much more likely to consider the Lacoste brand than their West German counterparts and are therefore potentially good targets.

On the purchasing stage, INDEP analysis ignores the positive, unobserved dependence on consideration and consequently misses that women are less likely to purchase Lacoste. "Female" has a negative effect on consideration that is spuriously carried forward to the next stage with a sign change by the INDEP analysis. The sign change occurs because of the positive, unobserved dependence and biases the negative coefficient to 0.

We now turn to a comparison to inference from the SAT model (columns labeled "S" in Table 3). It is apparent that the number of coefficients credibly different from 0 is larger overall in the saturated model which might be an unexpected result from including more variables in the model. The discrepancy between the SAT model and the other models on the repurchase stage seems especially striking.

Table 4   Posterior Means of Error Correlations (Identified Σ) Across Selection Stages: Saturated Model (SAT)

	Awareness	Consideration	Purchase	Repurchase
Awareness	1			
Consideration	−0.22 (0.37)	1		
Purchase	**0.99** (**0.12**)	−0.20 (0.35)	1	
Repurchase	**−0.98** (**0.04**)	0.22 (0.34)	**−0.99** (**0.05**)	1

Note. Posterior standard deviations are in parentheses.

The explanation for these differences are the multiple, extreme error correlations estimated by this model (see Table 4).

Such extreme contractions of the error space with correlations that cannot be statistically distinguished from one are characteristic of a lack of empirical identification.[23] With a sufficiently flexible mean structure, the multivariate likelihood is unbounded in the direction of error correlation matrices that approach rank deficiency (see also Web Appendix 1). This property of the likelihood creates a situation similar to conditioning on strong, nonzero error correlations when, in fact, successive stages are only weakly correlated or even independent. The consequence is a severely biased mean structure populated with coefficients to correct for imbalance from maximizing error correlations.[24] Because of the obvious identification problems in this model, we exclude it in the following discussion of targeting decisions.

4.3. Targeting Decisions

Next we discuss the implications for the management problem of targeting prospects for the purpose of alerting them to a new e-commerce channel. A priori, it makes sense to target consumers that are likely to consider the Lacoste brand; there is the possibility that a suitably designed promotion of the new e-commerce channel creates brand awareness. We first illustrate unobserved heterogeneity in the consideration probabilities. We keep all observed covariates constant,

[23] The determinant of the error correlation matrix from the saturated model is by two orders of magnitude smaller than that from the dependence model with variable selection, and the fact that we can estimate the saturated at all is entirely owed to the Bayesian approach and the use of proper subjective priors.

[24] In personal communication, Dan Ackerberg (2011) pointed out that our prior over different τ patterns, besides formalizing a lack of prior knowledge, translates into an MCMC that better navigates multimodal likelihood surfaces than an MCMC built from standard priors. We thank an anonymous reviewer for pointing out that the summaries reported for the saturated model in Tables 3 and 4 are likely to only capture a particular, well-pronounced local mode of an overall diffuse posterior that standard MCMC techniques fail to fully explore.

which are identified as active on the consideration stage in the DEP analysis. We then compare the probability of consideration given awareness, between a group of consumers more likely to be aware of the Lacoste brand and a group less likely to be aware of the brand as explained by other covariates.

Specifically, we look at male targets that live in a medium-sized city in the former GDR as gender, city size, and East versus West Germany are the only statistically credible drivers of consideration. Because of negative dependence between awareness and consideration, we expect consumers that are predicted to be less likely aware of the brand to less likely consider the brand, given awareness. Holding covariates effective on the consideration stage constant, we obtain consumers that are less likely aware of the brand by looking at older (age > 40), less educated (some high school or less), and less wealthy individuals (monthly income below or equal to DM4,500 (approximately €2,250)). Conversely, we obtain consumers that are more likely aware of the brand, and fit the description of covariates held constant, by looking at younger (age ≤ 40), more educated (high school diploma or more), and wealthier individuals (monthly income above DM4,500).

The DEP analysis indicates that there is much more heterogeneity in the consideration probabilities between the two groups than the INDEP analysis (see Table 5). Independent analysis underestimates the differential in consideration probabilities between the group that is predicted to be likely aware and the group predicted to be unlikely aware by a factor of 2.4 (= 0.2693/0.1071; see the last column of Table 5). The spurious coefficients estimated by the INDEP analysis only partially compensate for ignoring unobserved heterogeneity in η_1 that connects awareness and consideration.

Next we extend this example considering all four selection stages. We compare the expected success from targeting males, 40 years or younger, with a high school diploma or university degree, with monthly household incomes above DM4,500 (approximately €2,250), and in cities with more than 100,000 inhabitants in the former West and East Germany. The description of this target set T is motivated by a decision maker looking to inform consumers who are likely to be aware of the Lacoste brand.

We assume an average price of €100 for a Lacoste item, an average purchase interval of two years, and a discount rate of 7.5%. We compute the expected average revenue from targeting consumers as described above in the former West and East Germany and compare the predictions from INDEP and DEP analyses. We evaluate two scenarios. In the optimistic first scenario, we assume that alerting targeted consumers to the new channel of the Lacoste brand creates instant brand awareness; i.e., we condition on having passed the awareness selection (see Table 6).[25] In the pessimistic second scenario, we assume that alerting consumers to the new channel does not affect the awareness of the Lacoste brand. That is, if the message is sent to someone not aware of the brand, the message is lost (see Table 7). In both cases, we assume that alerting consumers to the new channel does not affect consideration, purchase, or repurchase probabilities.

The economic implications of targeting these prospects in East and West Germany are different across DEP and INDEP analyses. In the optimistic scenario of Table 6, DEP analysis predicts an average revenue difference of €12.94 from targeting members of the set T living in the former East Germany relative to West Germany. This expected total advantage is the sum of the average revenue difference from purchasing (€7.40) and repurchasing (€5.54). In contrast, INDEP analyses estimate the average revenue difference to be negative, i.e., −€5.02. That is, INDEP analyses underestimate the average revenue difference between targeting T in the former East and West Germany by about €18, or 140% per customer.

Moving to the pessimistic scenario regarding the effect of a potential channel campaign on brand awareness in Table 7, we see that both DEP and INDEP analyses estimate the average revenue difference between targeting T in the former East and West Germany to be negative. This is because of the relatively smaller probability of being aware of the Lacoste brand in the former East Germany (see Table 3). However, INDEP analyses overestimate the negative effect by approximately 100% (see the last two rows in Table 7).

Overall, our case study illustrates how ignoring unobserved heterogeneity in unobservables connecting successive selection stages biases the comparison

Table 5 Mean Consideration Probabilities Conditional on Awareness

	Predicted to be likely aware	Predicted to be less likely aware	Δ
DEP	0.6720	0.4027	0.2693
INDEP	0.5799	0.4728	0.1071

Note. Δ, difference between consumers predicted to be more and less likely aware of the brand.

[25] We note that conditioning on awareness is different from introducing direct mail as a hypothetical cause to awareness on the first stage of our model. Specifically, conditioning on awareness preserves the negative implications of "unlikely" awareness for the consideration stage. In contrast, introducing direct mail as a hypothetical cause, overriding everything else on the first stage, implies that selection based on unobservables affecting both awareness and consideration is no longer relevant when awareness is created by direct mail.

Table 6   Expected Average Revenues (€) for a One-Year (Purchase) and Two-Year Horizon (Repurchase) Conditional on Being Aware of the Brand (*Optimistic*)

	Formula	GDR	FRG	Δ
Purchase				
DEP	$E_{i\in T}P_i(y_2=1, y_3=1 \mid y_1=1) \times 100$	55.64	48.24	7.40
INDEP	$E_{i\in T}P_i(y_2=1)P(y_3=1) \times 100$	48.03	51.10	−3.06
Repurchase				
DEP	$E_{i\in T}P_i(y_2=1, y_3=1, y_4=1 \mid y_1=1) \times 100/1.075^2$	36.16	30.62	5.54
INDEP	$E_{i\in T}P_i(y_2=1)P(y_3=1)P(y_4=1) \times 100/1.075^2$	31.14	33.10	−1.96
Sum				
DEP		91.80	78.86	12.94
INDEP		79.17	84.20	−5.02

Note. Δ is the difference between East and West Germany.

Table 7   Expected Average Revenues (€) for a One-Year (Purchase) and Two-Year Horizon (Repurchase) Conditional on Awareness Distribution (*Pessimistic*)

	Formula	GDR	FRG	Δ
Purchase				
DEP	$E_{i\in T}P_i(y_1=1, y_2=1, y_3=1) \times 100$	38.31	46.87	−8.56
INDEP	$E_{i\in T}P_i(y_1=1)P(y_2=1)P(y_3=1) \times 100$	32.63	49.61	−16.98
Repurchase				
DEP	$E_{i\in T}P_i(y_1=1, y_2=1, y_3=1, y_4=1) \times 100/1.075^2$	24.89	29.75	−4.86
INDEP	$E_{i\in T}P_i(y_1=1)P(y_2=1)P(y_3=1)P(y_4=1) \times 100/1.075^2$	21.15	32.14	−10.99
Sum				
DEP		63.20	76.62	−13.42
INDEP		53.78	81.75	−27.97

Note. Δ is the difference between East and West Germany.

of different potential targets to an extent that could well lead a decision maker to ignore attractive targets that live in the former GDR.

5. Discussion

Selected and successively selected samples are common in marketing. We propose a model for *multiple* successive selection stages that generalizes the well-discussed two-stage case. Our generalized model builds on exclusion restrictions for observed and unobserved covariates, where the latter are the source of unobserved dependence between selection stages. We handle the common lack of prior knowledge about what covariates to exclude from what stage by formulating a prior over all possible sets of exclusion restrictions and we develop an MCMC to update this prior with data. The MCMC efficiently navigates different sets of exclusion restrictions but does not require every model visited to be well identified empirically for the implied model comparison to be meaningful.[26] This way, all information about likely model structures contained in the data is reflected in the posterior.

We highlight the relationship between the managerial goals of scoring, the targeting and influencing

consumers, and requirements for inference from successively selected data based on a DAG analysis of the selection problem. The results explain why scoring decisions are relatively more robust to ignoring unobserved dependence across selection stages and why influencing decisions may be completely misguided. We find targeting decisions based on a smaller set of covariate patterns to be similarly sensitive to ignoring unobserved dependence across stages.

In our illustrative application, we show how modeling unobserved dependence between successive selections effectively captures consumer heterogeneity, and we investigate a targeting problem. We find that accounting for versus ignoring successive dependent selections can reverse the relative attractiveness of targeting prospects in the former East and West Germany. Thus, the more active the approach to marketing, the more important proper accounting for unobserved dependence across successive selections becomes. Although modeling unobserved dependence between successive selections may appear to provide only a small advantage to firms that passively score customers in a stable environment, it becomes crucial for active targeting and marketing mix decisions.

By developing a prior over all possible sets of exclusion restrictions, we shift the burden of formulating dogmatic prior beliefs about specific exclusions to the weaker statement that some exclusion restrictions

---

[26] The marginal likelihood derived in the appendix is well defined independent of the likelihood identification of (all) the regression coefficients in a particular model as defined by the elements of τ.

may be supported by the data. Relative to estimating all possible models before comparing or averaging the results, our prior more practically handles the common lack of definite prior knowledge about exclusion restrictions in models of endogenous selection. Moreover, it can be easily adapted to express partial knowledge about the model structure. Finally, we find that our prior alleviates problems encountered because of weak empirical identification when fitting a full, unrestricted model in the hope of learning which effects can be excluded.

We suggest that priors over sets of models are a generally useful tool in the context of weak empirical identification of (some) members of this set and limited knowledge about the detailed mechanisms underlying the influence of covariates—a situation quite common in empirical marketing models with multiple, potentially dependent outcomes. Finally, we have assumed exogenous x variables, i.e., independence between x and η in Equation (1), throughout this paper. We leave the problem of jointly accounting for endogenous selection and endogenous observed covariates for future research.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/mksc.1120.0754.

Appendix. MCMC Details

The MCMC alorithm comprises the following steps.

Step 1. $\{\mathbf{z}\} \mid \{\boldsymbol{\beta}\}, \{\lambda\}, \{\boldsymbol{\eta}\}, \boldsymbol{\tau}, \{\mathbf{y}\}$.

Step 1 augments unobserved latent continuous variables $\{\mathbf{z}_i\}$ drawn from conditionally independent truncated normal distributions (Albert and Chib 1993). When a particular respondent failed to pass the *previous* selection stage,

the posterior distribution on the current stage is without truncation. However, because of conditional independence across stage-specific *equations* after conditioning on $\{\boldsymbol{\eta}_i\}$, we can simply ignore latent variables z that are not informed by the likelihood.

Step 2. $\boldsymbol{\tau} \mid \{\mathbf{z}\}, \{\boldsymbol{\eta}\}, \alpha$.

Conditional on random effects $\boldsymbol{\eta}$, stage-specific equations are independent. When coupled with independent priors for regression coefficients $\boldsymbol{\beta}$ and $\{\lambda\}$, the marginal likelihood $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \boldsymbol{\eta})$ required for sampling $\boldsymbol{\tau}$ factors into the stage-specific marginal likelihoods $p(\mathbf{z}_1 \mid \boldsymbol{\eta}) \times \cdots \times p(\mathbf{z}_K \mid \boldsymbol{\eta})$. In the following we derive this marginal likelihood for one stage omitting the stage-specific subscript. We start with the unnormalized, conditional posterior of $(\boldsymbol{\beta}, \{\lambda\})$, keeping track of the prior's and the likelihood's normalizing constants. We use a standard, diffuse multivariate normal prior. Its normalizing constant is $(2\pi)^{-k/2} |\mathbf{A}|^{1/2}$, where k corresponds to the number of active covariates, including unobserved covariates $\boldsymbol{\eta}$, and $\mathbf{A}$ is the corresponding $k \times k$ diagonal prior information matrix. The normalizing constant of the likelihood is 1 because errors ε are distributed standard normal:

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \boldsymbol{\eta}, \mathbf{z})p(\mathbf{z} \mid \boldsymbol{\eta}) = p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta})p(\boldsymbol{\beta}, \boldsymbol{\lambda}),$$

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \boldsymbol{\eta}, \mathbf{z})$$

$$\propto \exp\left[\frac{-1}{2}\left(\mathbf{z} - [\mathbf{X}, \boldsymbol{\eta}]\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix}\right)'\left(\mathbf{z} - [\mathbf{X}, \boldsymbol{\eta}]\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix}\right)\right] \quad (6)$$

$$\times \exp\left[\frac{-1}{2}\left(\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix} - \begin{bmatrix}\bar{\boldsymbol{\beta}}\\\bar{\boldsymbol{\lambda}}\end{bmatrix}\right)'\mathbf{A}\left(\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix} - \begin{bmatrix}\bar{\boldsymbol{\beta}}\\\bar{\boldsymbol{\lambda}}\end{bmatrix}\right)\right]$$

$$\times (2\pi)^{-k/2}|\mathbf{A}|^{1/2}.$$

We rewrite the unnormalized posterior from Equation (6) (e.g., Zellner 1971):

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \boldsymbol{\eta}, \mathbf{z}) \propto \exp\left[\frac{-1}{2}\left(\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix} - \begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix}\right)'\right.$$

$$\cdot ([\mathbf{X}, \boldsymbol{\eta}]'[\mathbf{X}, \boldsymbol{\eta}] + \mathbf{A})$$

$$\cdot \left.\left(\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix} - \begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix}\right)\right] \times \exp\left[\frac{-n\tilde{s}^2}{2}\right]$$

$$\times (2\pi)^{-k/2}|\mathbf{A}|^{1/2},$$

$$n\tilde{s}^2 = \left(\mathbf{z} - [\mathbf{X}, \boldsymbol{\eta}]\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix}\right)' \quad (7)$$

$$\cdot \left(\mathbf{z} - [\mathbf{X}, \boldsymbol{\eta}]\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix}\right) + \left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix} - \begin{bmatrix}\bar{\boldsymbol{\beta}}\\\bar{\boldsymbol{\lambda}}\end{bmatrix}\right)'$$

$$\cdot \mathbf{A}\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix} - \begin{bmatrix}\bar{\boldsymbol{\beta}}\\\bar{\boldsymbol{\lambda}}\end{bmatrix}\right),$$

$$\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix} = ([\mathbf{X}, \boldsymbol{\eta}]'[\mathbf{X}, \boldsymbol{\eta}] + \mathbf{A})^{-1}\left([\mathbf{X}, \boldsymbol{\eta}]'\mathbf{z} + \mathbf{A}\begin{bmatrix}\bar{\boldsymbol{\beta}}\\\bar{\boldsymbol{\lambda}}\end{bmatrix}\right).$$

We then integrate the resulting multivariate normal distribution by $(\boldsymbol{\beta}, \{\lambda\})$ to obtain the marginal likelihood:

$$\int p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \boldsymbol{\eta}, \mathbf{z})p(\mathbf{z} \mid \boldsymbol{\eta})d(\boldsymbol{\beta}, \boldsymbol{\lambda})$$

$$= \int \exp\left[\frac{-1}{2}\left(\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{\lambda}\end{bmatrix} - \begin{bmatrix}\tilde{\boldsymbol{\beta}}\\\tilde{\boldsymbol{\lambda}}\end{bmatrix}\right)'\right.$$

$$\cdot \left([\mathbf{X}, \boldsymbol{\eta}]'[\mathbf{X}, \boldsymbol{\eta}] + \mathbf{A}\right) \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} \right) \Bigg] d(\boldsymbol{\beta}, \boldsymbol{\lambda}) \quad (8)$$

$$\times \exp\left[\frac{-n\tilde{s}^2}{2}\right] (2\pi)^{-k/2} |\mathbf{A}|^{1/2},$$

$$p(\mathbf{z} \mid \boldsymbol{\eta}) = |\mathbf{A}|^{1/2} |([\mathbf{X}, \boldsymbol{\eta}]'[\mathbf{X}, \boldsymbol{\eta}] + \mathbf{A})|^{-1/2} \exp\left[\frac{-n\tilde{s}^2}{2}\right].$$

We note that $(\boldsymbol{\beta}, \{\boldsymbol{\lambda}\})$ could be integrated out from Equation (6) directly. However, the resulting marginal likelihood is a multivariate normal distribution of dimensionality $n \times n$ where n is the number of observations on the current stage. Reexpressing the unnormalized posterior first as in Equation (7) leads to a much more manageable representation of the marginal likelihood.

We use a Metropolis–Hastings step to update individual elements of the vector of exclusion restrictions $\boldsymbol{\tau}$ setting the proposal density equal to the hierarchical prior $p(\tau \mid \alpha)$. Note that we only need to compute $p(\mathbf{z} \mid \boldsymbol{\eta})$ when we propose to include a previously excluded covariate, and vice versa. When an element of $\boldsymbol{\tau}$ changes from 1 to 0 (0 to 1), the corresponding column is deleted from (added to) $[\mathbf{X}, \boldsymbol{\eta}]$ and $\mathbf{A}$. The penalty for increasing the dimensionality of the model comes through the determinant of $\mathbf{A}$ in Equation (8). This determinant converges to 0 as covariates are added to the model when the diagonal elements of $\mathbf{A}$ are smaller than 1. We use 0.01 for the diagonal elements of $\mathbf{A}$ and center the prior on 0.

Step 3. $\{\boldsymbol{\beta}\}, \{\boldsymbol{\lambda}\} \mid \{\mathbf{z}\}, \{\boldsymbol{\eta}\}$.

Step 3 is a standard conjugate update of regression coefficients that is facilitated by independence between selection equations conditional on $\{\boldsymbol{\eta}_i\}$. In this update, all columns in the design matrix collecting observed $(\mathbf{x})$ and unobserved covariates $(\boldsymbol{\eta})$ in Equation (5) that correspond to elements of $\boldsymbol{\tau}$ equal to 0 are deleted. The dimensionality of the subjective multivariate normal prior for $(\{\boldsymbol{\beta}\}, \{\boldsymbol{\lambda}\})$ is adjusted accordingly.

Step 4. $\{\boldsymbol{\eta}\} \mid \{\boldsymbol{\beta}\}, \{\boldsymbol{\lambda}\}, \{\mathbf{z}\}$.

Step 4 updates the random effects $\{\boldsymbol{\eta}_i\}$ using standard conjugate results. Similar to augmenting the latent variables $\{\mathbf{z}_i\}$ in Step 1, we only need to augment these elements of a random effect vector $\boldsymbol{\eta}_i$ that correspond to latent variables $\mathbf{z}_i$, which in turn are informed by corresponding observations $\mathbf{y}_i$.

Step 5. $\alpha \mid \boldsymbol{\tau}$.

Finally, Step 5 is a standard conjugate update of the hierarchical prior probability α that a particular element of the indicator vector $\boldsymbol{\tau}$ is equal to 1. We use a noninformative Beta$(1, 1)$ prior for α.

References

Ackerberg D (2011) Personal communication with authors. July, Montréal.

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* 88:669–679.

Brown PJ, Vannucci M, Fearn T (1998) Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B* 60: 627–641.

Cao Y, Gruca TS (2005) Reducing adverse selection through customer relationship management. *J. Marketing* 69:219–229.

Chib S, Seetharaman PB, Strijnev A (2004) Model of brand choice with a no-purchase option calibrated to scanner-panel data. *J. Marketing Res.* 41:184–196.

Dellaportas P, Forster JJ, Ntzoufras I (2002) On Bayesian model and variable selection using MCMC. *Statist. Comput.* 12:27–36.

Edwards YD, Allenby GM (2003) Multivariate analysis of multiple response data. *J. Marketing Res.* 40:321–334.

Erdem T (1996) A dynamic analysis of market structure based on panel data. *Marketing Sci.* 15:359–378.

Frühwirth-Schnatter S, Tüchler R (2008) Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statist. Comput.* 18:1–13.

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* 88:881–889.

George EI, McCulloch RE (1995) Stochastic search variable selection. Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in Practice* (CRC Press, Boca Raton, FL), 203–214.

George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. *Statistica Sinica* 7:339–374.

Ghose A, Yang S (2009) An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Sci.* 55:1605–1622.

Greene WH (2002) *LIMDEP Econometric Modeling Guide: Version 8.0*, Vol. 1 (Econometric Software, New York).

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47:153–162.

Keane MP (1992) A note on identification in the multinomial probit model. *J. Bus. Econom. Statist.* 10:193–200.

Lee EJ, Eastwood DB, Lee J (2004) A sample selection model of consumer adoption of computer banking. *J. Financial Services Res.* 26:263–275.

Li K, Prabhala NR (2007) Self-selection models in corporate finance. Eckbo BE, ed. *Handbook of Corporate Finance: Empirical Corporate Finance*, Vol. 1, Chapter 2 (Elsevier, Amsterdam).

Maddala GS (1983) *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge University Press, Cambridge. UK).

McCulloch R, Rossi PE (1994) An exact likelihood analysis of the multinomial probit model. *J. Econometrics* 64:207–240.

Riesenbeck H, Perrey J (2009) *Power Brands: Measuring, Making, Managing Brand Success*, 2nd ed. (Wiley-VCH, Weinheim, Germany).

Rossi PE, Allenby GM, McCulloch R (2005) *Bayesian Statistics and Marketing* (John Wiley & Sons, Chichester, UK).

Tüchler R (2008) Bayesian variable selection for logistic models using auxiliary mixture sampling. *J. Comput. Graphical Statist.* 17:76–94.

Vella F (1998) Estimating models with sample selection bias: A survey. *J. Human Resources* 33:127–169.

Williams D (2009) Tracking purchase funnel metrics in a global market. Report, Gfk Cstom Research North America, New York. http://www.gfkcustomresearchbrasil.com/imperia/md/content/gfkcustomresearchbrasil/2010_afi_brazill_dick_williams.pdf.

Wold H (1954) *A Study in the Analysis of Stationary Time Series*, 2nd ed. (Almqvist and Wiksell Book Co., Uppsala, Sweden).

Zeithammer R, Lenk P (2009) Statistical benefits of choices from subsets. *J. Marketing Res.* 46:816–831.

Zellner A (1971) *An Introduction to Bayesian Inference in Econometrics* (John Wiley & Sons, New York).

Zhao Y, Zhao Y, Song I (2009) Predicting new customers' risk type in the credit card market. *J. Marketing Res.* 46:506–517.