Longitudinal variable selection by cross-validation in the case of many covariates

E. Cantoni, C. Field, J. Mills Flemming
and E. Ronchetti

No 2005.01

Cahiers du département d'économétrie
Faculté des sciences économiques et sociales
Université de Genève

Février 2005

Département d'économétrie
Université de Genève, 40 Boulevard du Pont-d'Arve, CH -1211 Genève 4
http://www.unige.ch/ses/metri/

Longitudinal variable selection by cross-validation in the case of many covariates

E. Cantoni[1], C. Field[2], J. Mills Flemming[2] and E. Ronchetti[1]

[1] Department of Econometrics, University of Geneva,
CH-1211 Geneva 4, Switzerland
and
[2] Department of Mathematics and Statistics,
Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1

February 2005

Abstract

Longitudinal models are commonly used for studying data collected on individuals repeatedly through time. While there are now a variety of such models available (Marginal Models, Mixed Effects Models, etc.), far fewer options appear to exist for the closely related issue of variable selection. In addition, longitudinal data typically derive from medical or other large-scale studies where often large numbers of potential explanatory variables and hence even larger numbers of candidate models must be considered. Cross-validation is a popular method for variable selection based on the predictive ability of the model. Here, we propose a cross-validation Markov Chain Monte Carlo procedure as a general variable selection tool which avoids the need to visit all candidate models. Inclusion of a "one-standard error" rule provides users with a collection of good models as is often desired. We demonstrate the effectiveness of our procedure both in a simulation setting and in a real application.

1 INTRODUCTION

Longitudinal modelling techniques are commonly used for studying data collected on individuals repeatedly through time. Such data arises frequently in medical studies where large numbers of potential explanatory variables are often considered. A variety of modelling approaches have been proposed for handling such data and yet variable selection, key to any statistical analysis, is typically neglected. An exception here is the recent work of [3] and [12]. Many commercially available software packages (Splus, SAS, Stata, etc.) now include routines for analyzing longitudinal data but pay little attention to variable selection, even in cases where many potential explanatory variables and hence even larger numbers of candidate models, should be considered.

Although the final choice of model(s) must take into account subject matter and other nonstatistical aspects, data-based statistical methods are very useful tools for variable selection. Probably the most widely used method for estimating the predictive ability of a model is cross-validation. Here we utilize cross-validation in a longitudinal setting with special attention to the computational complexity associated with considering many candidate models which makes approaches as in [3] and [12] inappropriate from a computational point of view. The basic idea is as follows (see [15]). Given a sample of K subjects, we randomly split the data into a construction sample of size K_c and a validation sample of size $K - K_c$. We use the construction sample to fit the model, and use the validation sample to evaluate the prediction error of the particular model. We repeat this procedure for M splits. With few explanatory variables we can proceed directly. However, with large numbers of variables we cannot compute the prediction error for all models. We therefore do cross-validation using a MCMC random search procedure that allows us to sensibly sample the model space. This proposal is built on ideas originally proposed by Qian and Field (see [6], [14]) and moves efficiently through the model space by turning the cross-validation procedure into one of random sample generation from a finite population. Essentially a probability distribution for the various candidate models is defined based on a prediction error criterion. An MCMC method, based on either the Gibbs sampler or the Metropolis-Hastings algorithm, can then be used to generate a sample from this probability distribution. The convergence of the MCMC method ensures that variable selection from the random sample generated is consistent with that from all candidate models, provided that the MCMC sample is sufficiently large.

One benefit of doing the M splits of the data in our procedure is that it provides (at no additional cost) a measure of the standard error of our prediction error estimate. We use this standard error as the basis of a ”one-standard error” rule as is often used with cross-validation ([10]). Depending on the goal(s) of the user, one might choose the most parsimonious model whose error is no more than one standard error above the error of the best model. Alternatively, one might choose to average predictions over a collection good models (e.g. those within one standard error of the best model) as in the spirit of bagging, see [1].

Marginal Longitudinal Generalized Linear Models, available in most software packages, are a popular option for analyzing longitudinal data and are fit using Generalized Estimating Equations (GEE). Given their popularity we have tailored our cross-validation MCMC procedure to this setting while noting that many other environments are possible.

The paper is organized as follows. In Section 2 we develop our cross-validation MCMC procedure. In Section 3 we present the results of a simulation study that examines the performance of our procedure in a number of different settings. An application on real data from the Coronary Artery Risk Development in Young Adults (CARDIA) Study is presented in Section 4. Conclusions and directions for future research are provided in Section 5.

2 METHODOLOGY

2.1 Estimators for Prediction

We consider a longitudinal data analysis setting, where Y_{it} is the discrete or continuous outcome for subject i at time t, for $i = 1, \cdots, K$ and $t = 1, \cdots, n_i$. For each outcome Y_{it}, we also measure a set of covariates x_{it}. We write $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ for the $n_i \times 1$ vector of responses, and $X_i = (x_{i1} \cdots x_{in_i})^T$ for the $n_i \times p$ matrix of covariates of subject i. Purely dependent data are obtained with $K = 1$ (only one cluster) and purely independent data are obtained with $n_i = 1$ for all i. We fit a Marginal Model to the data by modelling the marginal mean $E(Y_{it}) = \mu_{it}$, and assuming that $g(\mu_{it}) = x_{it}^T\beta$ for a known link function g, and $Var(Y_{it}) = \sigma^2\nu(\mu_{it})$. We suppose that $Corr(Y_i) = A_i^{-1}Var(Y_i)A_i^{-1}$, with $A_i = \text{diag}(\nu^{1/2}(\mu_{i1}), \cdots, \nu^{1/2}(\mu_{in_i}))$, and that the subjects are independent. An estimator $\hat{\beta}$ is the solution of the general estimating equations proposed by [11]:

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0, \tag{1}$$

where $S_i = Y_i - \mu_i$, $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$, $D_i = D_i(X_i, \beta) = \partial\mu_i/\partial\beta$ is a $n_i \times p$ matrix, and $V_i = V_i(\mu_i, \alpha) = A_i R_i(\alpha) A_i$ is a $n_i \times n_i$ matrix. The matrix

$R_i(\alpha)$, for an s-parameter α, is said to be the working correlation matrix, as opposed to the "true" correlation matrix $Corr(Y_i)$.

One may wish to consider a more compact model

$$g(\mu_{it}) = x_{itv}^T \beta_v \tag{2}$$

in situations when some of the components of β are zero. We take $v = (v_1, \cdots, v_p)$ with some components equal to 1 and others equal to 0 and let d_v denote the total number of 1s occuring in v ($d_v \leq p$). Then β_v is a vector of length d_v containing the non-zero components of β and similarly x_{itv} is a vector of length d_v. There are $2^p - 1$ possible different models each of which corresponds to a subset (or candidate model) v.

2.2 The Cross-Validation MCMC Procedure

Consider the original sample (X_i, y_i), $i = 1, \cdots, K$ and split the data into a construction sample of size K_c and a validation sample of size $K_{val} = K - K_c$. Note that we are splitting on subjects, the natural sampling unit in the context of longitudinal data (rather than individual observations). The estimation procedure described in Section 2.1 provides estimators for prediction in the validation sample. That is, we use the construction sample to obtain estimates of β_v for the particular candidate model v of concern. Then for each observation (X_i,y_i) in the validation sample, we can compare the observed value y_i with the prediction $\hat{y}_i$. A suitable choice upon which to base our prediction error criterion is hence a loss function of the form

$$\sum_{i=1}^{K_{val}} (y_i - \hat{y}_i)^T V_i^{-1} (y_i - \hat{y}_i), \tag{3}$$

where V_i is the covariance matrix (which depends on the working correlation $R_i(\alpha)$) that we estimate (only once) based on the full set of observations. This procedure enables one to arrive at a measure of prediction error for each candidate model in our chain. In summary, the average prediction error criterion is calculated as follows:

- Generate M random splits of the dataset into a construction sample of size K_c and a validation sample of size $K - K_c$, where K_c is approximately $K^{3/4}$ (see [15] for further details).

- For each split, use the construction sample to fit the v model. The parameters of the model are estimated by the procedure given in the section above.

- Compute the prediction error for each of the M splits and then the average prediction criterion over the M splits for model v. We denote this quantity $PE(v)$.

Note that the same M splits are used for each model v that we evaluate.

An exhaustive variable selection procedure requires the evaluation of $PE(v)$ for each candidate model and is therefore not feasible when there are a large number of candidate models, that is, when p is moderate to large. To overcome this difficulty we propose an MCMC random search procedure built on defining an appropriate transition kernel,

$$P(v) = B \exp\{-PE(v|Y, X)\}, \tag{4}$$

where $B = (\sum_v \exp\{-PE(v|Y, X)\})$ and $PE(v)$ represents the average prediction error criterion. We then proceed with variable selection from a sample of candidate models generated using the probability distribution $P(v)$. This approach is based on methods in [14].

For the probability distribution $P(v)$, the evaluation of the constant B is not computationally feasible when there are a large number of candidate models. Fortunately, one can generate a sample from $P(v)$, by applying an MCMC method even though B does not have a computable form. To use an MCMC method, one needs a properly determined transition kernel which generates a reversible Markov Chain. If the transition kernel satisfies a so-called detailed balance condition and has a support covering that of $P(v)$, it can be shown that $P(v)$ is the stationary distribution of the generated Markov chain. Therefore after an initial burn-in period the generated Markov chain becomes ergodic and can be used for most purposes as an i.i.d. sample from $P(v)$, even though the models in the chain are not independent. Since we are looking for models with minimum PE, our objective is simply to move through the model space in order to find them. As a result, an i.i.d. sample and removal of the burn-in period are actually not necessary. A similar approach is taken in [9] where a Gibbs sampler is used for Bayesian variable selection in the context of multiple regression.

We will apply one of the most frequently used MCMC methods, the Metropolis-Hasting algorithm, for generating a sample from the distribution $P(v)$ defined on the set of all candidate models. Alternatively one could use Gibbs sampling (see [14] for further details). Note that generating a sample for $P(v)$ amounts to generating a sequence of $1 \times p$ binary vectors. Hence the following algorithm for generating a sample $\{v^{(1)}, v^{(2)}, \cdots, v^{(J)}\}$ can be used:

- Arbitrarily choose a starting model $v^{(0)} = (v_1^{(0)}, \cdots, v_p^{(0)})$, compute

$PE(v^{(0)})$ and its corresponding standard error (over the M splits) hereafter denoted $\sigma(PE)$.

- Repeat for $j = 1 \cdots, J$: To get the model $v^{(j)}$, first generate a candidate model $\tilde{v}$ from an operating transition kernel $q(v|v^{(j-1)})$ for v, and generate a u from Uniform(0,1). Then set $v^{(j)} = \tilde{v}$ if

$$\begin{aligned} u \leq r(v^{(j-1)}, \tilde{v}) = &\ \min\{([P(\tilde{v})/P(v^{(j-1)})]^{c/\sigma(PE)} * q(v^{(j-1)}|\tilde{v})/q(\tilde{v}|v^{(j-1)}), 1\} \\ = &\ \min\{\exp(c * \tfrac{(PE(v^{(j-1)})-PE(\tilde{v}))}{\sigma(PE)} * q(v^{(j-1)}|\tilde{v})/q(\tilde{v}|v^{(j-1)}), 1\}, \end{aligned}$$

  otherwise set $v^{(j)} = v^{(j-1)}$.

- Return the model sequence $\{v^{(1)}, v^{(2)}, \cdots, v^{(J)}\}$.

The constant $c = -\log(prob)$ is used to calibrate the MCMC chain so that it visits a reasonable number of candidate models. Note that the larger the value of c, the higher the probability of moving to a candidate model with larger PE. We refer to [7] for a discussion on calibration of the Metropolis-Hasting algorithm. We define the operating transition kernel for all candidate models v which differ from the present model $v^{(j-1)}$ only in that they either include an additional covariate or exclude a present one. The probability $q(v|v^{(j-1)})$ is calculated based on the p-value (obtained from the t-test for the full model) for the covariate under consideration for inclusion or exclusion. Let p^{t-test} be a vector of size p containing the p-value of the t-test in the full model (for each explanatory variable). For a model $v^{(j-1)}$ define the set of its neighboring models as $M_{v^{(j-1)}} = \{m^1, \ldots, m^p\}$, where each m^i is such that $\sum_{k=1}^{p} |v_k^{(j-1)} - m_k^i| = 1$. For each $m^i \in M_{v^{(j-1)}}$, the transition kernel is then defined by

$$q(m^i|v^{(j-1)}) = \frac{(1 - p_i^{t-test}) * E_i + p_i^{t-test} * (1 - E_i)}{\sum_{l=1}^{p}[(1 - p_l^{t-test}) * E_l + p_l^{t-test} * (1 - E_l)]},$$

where $E_i = 1_{\sum_{k=1}^{p}(m_k^i - v_k^{(j-1)})=1} = 1$ if $\sum_{k=1}^{p}(m_k^i - v_k^{(j-1)}) = 1$, that is if m^i includes an extra variable with respect to v^{j-1}, and 0 otherwise. For example, suppose we have a full model with $p = 3$ covariates where $p^{t-test} = (.05, .65, .15)$. Note that if the p-value is small we would typically like to keep the covariate in the model so we choose to add it with probabililty 1 - p-value. Now suppose that $v^{(j-1)} = (1, 0, 0)$ (i.e. the model that contains only the first covariate) and the model under consideration in the set of neighboring models is $m^1 = (1, 1, 0)$. Then, $q(m^1|v^{(j-1)}) = (1 - .65)/(.05 + (1 - .65) + (1 - .15))$ and similarly $q(v^{(j-1)}|m^1) = .65/(.05 + .65 + (1 - .15))$.

Alternatively one could use uniform probabilities, $q(v|v^{(j-1)}) = \frac{1}{p}$ (where p is the number of covariates) but our experience has shown this approach to be less efficient. Note that various choices for the operating kernel are possible. It is our feeling that we have chosen the most sensible one for this application.

2.3 The "one-standard error" Rule

For each model v in the chain our MCMC cross-validation procedure provides both a measure of predictive ability (PE(v)) and the associated standard error, σ computed over the M splits for the model. This latter statistic has traditionally not been available and hence represents one of the innovations of our procedure. Supposing that v_0 is the model in the chain with the smallest PE, we can define a set of indistinguishable models as being comprised of all those models whose PE is within σ of PE(v_0). We refer to this approach as the "one-standard error" rule in keeping with that suggested on page 214 of [10]. We view the resulting set as representing a collection of good models for the data.

The researcher now can summarize the collection of good models in a number of ways. In many prediction settings a search for a single best model for a particular set of data is neither sensible nor reasonable. Instead, as in the spirit of bagging ([1]), models provided by the "one-standard error" rule can be used together to obtain good prediction estimates. By looking at the ensemble of variables selected, the researcher can see which variables occur across most or all of the models giving a set of core variables to be included for any analysis. There may also be situations where one or the other of a pair of variables is selected indicating that each has similar explanatory power hence suggesting new composite variables. If one final model is required for further analysis, a sensible strategy would be to include all variables occurring in the majority of models creating a consensus model. In Section 3 and 4 we illustrate these ideas in the context of simulated and actual data.

2.4 Extensions

We note that our procedure could be used with robust GEE ([2]) rather than GEE. In addition there are other choices available for PE, Efron's .632 estimator (p. 321 of [4]), for example. For an interesting and insightful discussion of some of the relationships between these estimators we suggest [5]. If one wished to move more quickly to other models as comprised of those models that differ from the present one in that they include or exclude 2 covariates rather than just 1.

Table 1: Marginal frequencies of appearance of the x variables in a chain of length 5000. Note that x_1 through x_7 are significant.

	no interaction			interaction		
	mean	median	sd	mean	median	sd
x_1	0.83	1	0.28	0.59	0.5	0.33
x_2	0.76	1	0.3	0.8	1	0.29
x_3	0.78	1	0.29	0.67	0.66	0.25
x_4	0.87	1	0.19	0.76	1	0.29
x_5	0.83	1	0.23	0.79	0.98	0.27
x_6	0.81	1	0.24	0.89	1	0.18
x_7	0.84	1	0.23	0.89	1	0.17
x_8	0.17	0.16	0.07	0.25	0.25	0.11
x_9	0.23	0.2	0.15	0.26	0.24	0.1
x_{10}	0.24	0.19	0.22	0.21	0.2	0.12
x_{11}	0.19	0.15	0.16	0.28	0.25	0.14
x_{12}	0.4	0.36	0.16	0.38	0.36	0.14
x_{13}	0.37	0.35	0.13	0.35	0.33	0.12
x_{14}	0.37	0.35	0.11	0.35	0.34	0.09
x_{15}	0.38	0.37	0.14	0.36	0.35	0.13
x_{16}	0.35	0.36	0.09	0.38	0.36	0.1
x_{17}	0.37	0.38	0.11	0.4	0.38	0.15
x_{18}	0.39	0.35	0.14	0.37	0.34	0.13
x_{19}	0.4	0.39	0.08	0.37	0.34	0.17
x_{20}	0.36	0.36	0.1	0.38	0.37	0.14

3 SIMULATION STUDY

To assess the performance of our cross-validation MCMC procedure we have carried out a simulation study designed to measure performance in a number of different settings as well as to serve as a general indicator of utility.

We consider a marginal longitudinal model (as described in Section 2.1) with log link, for $i = 1, \cdots, K = 30$ and $t = 1, \cdots, n_i = n = 10$. The response Y_{it} is Poisson. The dimension of x_{it} is $p = 20$ with this set of explanatory variables including a combination of those which are time-dependent or time-independent as well as those which are continuous or discrete. The correlation between observations on the same subject is exchangeable. The subjects are assumed independent. We take $c = -\log(.5)$ in our cross-validation MCMC procedure. Our starting model is the one retaining all the variables for which the individual t-tests are significant.

We begin by considering a setting in which X_1, X_2, X_8 and X_9 are binomial(.5), X_3, X_{10} and X_{11} are three-level variables with probabilities (0.5,0.35,0.15) and $X_4 \rightarrow X_7$ and, $X_{12} \rightarrow X_{20}$ are N(0,1) variables. The true model generating the data includes 7 significant variables, $X_1 \rightarrow X_7$. We run 50 simulations and for each utilize an MCMC chain of length 5000. The true model generating the data is usually visited quite early on giving us confidence in the fact that the chain is of sufficient length. Each chain provides 5000 models and we compute the frequency of appearence of all 20 explanatory variables in these models. We run 50 simulations and in the left half of Table 1 we report summaries (mean, median, standard deviation) of the distribution of the marginal frequencies of appearance of all 20 explanatory variables over the 50 simulations. This approach to summarizing results is similar in spirit to that of [8] where promising covariates were identified as those with more frequent appearance in the Gibbs sample. Clearly, our procedure does an excellent job of identifying those variables which are significant. In addition (results not shown), all variables appear in models at least 15% of the time suggesting that we are moving around the design space quite well. Similar results are obtained when the number of significant variables generating the true model is reduced. We next consider a setting in which there are interactions. That is, X_1, X_8 and X_9 are binomial(.5), X_2, X_{10} and X_{11} are three-level variables with probabilities (0.5,0.35,0.15), $X_3 \rightarrow X_5$ and, $X_{12} \rightarrow X_{18}$ are N(0,1) variables, $X_6 = X_1 * X_3$, $X_7 = X_4 * X_5$, $X_{19} = X_8 * X_{12}$ and $X_{20} = X_{13} * X_{14}$. The true model generating the data includes 7 significant variables, $X_1 \rightarrow X_7$. Again we run 50 simulations and for each utilize an MCMC chain of length 5000. In the right half of Table 1 we report the marginal frequencies of appearance of all 20 explanatory variables over the 50 simulations. We assume that, in order to include an interaction effect one must also include the corresponding individual effects in the model. Again, our procedure does an excellent job of identifying those variables which are significant.

Our cross-validation MCMC procedure can achieve a variety of objectives. For instance, suppose that one requires a single best model for the dataset of interest. We hereafter refer to such a model as a *consensus* model and suggest that it need not simply be the model in our chain with the smallest PE, though this is one possibility. Instead, one could choose to define the consensus model dependent upon the objective(s) of the end user. For example, suppose that the most parsimonious model for the dataset is desired. One could then define the consensus model to include only those variables occuring in *all* of the indistinguishable models, as obtained by application of the “one-standard error” rule. At the other end of the spectrum, one could define the consensus model as being comprised of all variables occuring in *any*

Table 2: Consensus Model and number of Indistinguishable Models, 7 significant out of 20.

Sim #	Consensus model	Indist.	Sim #	Consensus model	Indist.
1	1 2 3 5 7 18	15	26	1 3 4 5 6 7	22
2	2 3 4 5 6 7 14	25	27	1 2 3 4 5 6 7	20
3	3 4 5 6 7	23	28	1 2 3 4 6 7 13	9
4	1 2 4 6 7	11	29	1 3 4 5 6 7 10 13 19	26
5	1 2 3 4 5 6 7	18	30	1 2 3 4 5 6 7	58
6	1 2 3 4 5 6 7 20	15	31	3 4 5 6 7	5
7	1 2 3 4 5 6 7 16 17	17	32	1 2 4 5 6 7 14	23
8	3 4 5 6 7	6	33	1 3 4 5 6 7 15 20	21
9	1 3 4 5 7 12	14	34	1 2 4 5 6 7 9 10 17 18	9
10	1 4 5 6 7 12 15 18	14	35	1 3 4 5 6 8	5
11	2 3 5 6 7	44	36	3 4 5 6 7 20	7
12	1 2 4 5 6 7 9	8	37	1 2 3 4 5 6 7 12	27
13	1 2 3 5 6 7 18	22	38	1 2 3 4 5 7	13
14	1 2 3 4 5 6 7	11	39	1 3 4 5 6 7	12
15	1 2 4 7	14	40	1 2 4 5 7	20
16	1 2 3 4 5 6 14	6	41	1 2 3 4 5 6 7	16
17	1 3 4 5 6 7	11	42	3 4 5 6 8 11 13 15	10
18	1 2 3 4 5 6 7 16 19	46	43	1 2 3 4 5 6 7	14
19	1 2 3 4 5 6 7 12	6	44	1 2 4 5 6 7	10
20	1 2 3 4 5 6 7 13	14	45	1 2 3 4 5 6 7 11	9
21	1 2 3 4 5 7 12	8	46	1 2 3 4 5 6 7	10
22	1 2 4 5 6 7 10	26	47	1 2 3 4 5 6 7 15	9
23	1 2 3 4 5 6 7 15	45	48	1 2 3 4 5 6 7	14
24	1 2 3 4 5 6 7 9 20	31	49	1 2 4 5 6 7 13 18	11
25	3 4 5 6 7	8	50	2 3 4 5 6 7 17 20	15

of the indistinguishable models. Such a model would clearly be more conservative in nature. In Table 2 we report the consensus model along with the correponding number of indistinguishable models when the consensus model is taken to include all variables occuring in more than *50%* of the indistinguishable models. There are only 15 instances in which the true model is not visited and yet it is within σ of the consensus model. This reflects our choice of c and is reasonable given that there are over one million possible candidate models. Furthermore, we visit the true model about 60% of the time. In Table 3 we report similar results in the presence of interaction. To summarize, our procedure provides a rich summary of predictive ability enabling one to address a vast array of questions pertaining to variable selection.

Table 3: Consensus Model and number of Indistinguishable Models, 7 significant (with interactions) out of 20.

Sim #	Consensus model	Indist.	Sim #	Consensus model	Indist.
1	2 3 4 5 6 7 11 13 14	37	26	1 2 4 5 6 7 11 14 17 20	83
2	2 3 4 6 7 12	8	27	1 2 5 6 7 16	10
3	1 2 4 6 7	39	28	2 3 4 5 7	13
4	1 2 3 4 5 7	25	29	4 6 7 11 12	14
5	2 3 4 6 7	12	30	1 5 6 7	21
6	1 2 3 4 5 7	11	31	1 2 3 4 5 6 7 13	8
7	2 3 4 5 6 7	31	32	1 2 4 5 6 7 9 13 16	9
8	2 4 5 6 7	20	33	1 2 4 5 6 7 10 12 19	22
9	1 2 3 4 5 6 7	8	34	2 4 5 6 7 9 20	20
10	2 3 4 5 6 7 18	32	35	1 4 6 7	12
11	1 2 3 4 5 6 7 16 18 20	4	36	1 2 3 4 5 6 7 12 14 19	9
12	1 2 4 5 6 7 11 19	30	37	1 2 5 6 7 17	14
13	2 3 5 7 17	15	38	1 2 4 5 6 7	17
14	1 2 3 5 7 18	19	39	1 2 5 6 7 18	12
15	2 3 4 5 6 7 9 17	10	40	1 2 3 4 5 6 7	13
16	1 2 3 4 5 6 7	7	41	2 3 4 5 6 7 8 15	12
17	1 5 6 7	15	42	2 3 4 5 6 7	19
18	1 2 4 5 6 7 15 20	12	43	1 2 3 4 5 6 7	10
19	1 2 3 4 5 6 7	5	44	1 2 3 5 6 12	12
20	2 4 5 6 7 12 13 16 17 19 20	24	45	2 3 4 5 6 7 14	17
21	1 2 3 4 5 6 8 12 19	38	46	1 2 4 5 6 7	6
22	2 3 4 5 6 7	9	47	2 3 4 5 6 7 10	5
23	5 6 7 8 14 16	9	48	1 2 3 4 5 6 7 11	12
24	1 2 3 4 6 7 10	7	49	2 4 5 6 7	15
25	1 3 5 6 7 12	11	50	2 3 4 5 6 7	9

4 APPLICATION TO REAL DATA

We consider here data from the Coronary Artery Risk Development in young Adults (CARDIA) Study which was originally designed to document levels of risk factors for coronary artery disease and potential determinants of these risk factors in young adults in the United States. [13] make available a portion of the data corresponding to 5,078 black and white young adults and propose a weighted generalized estimating equations approach that accounts for dropouts. We choose to examine a subset of this data in order to demonstrate the utility of our cross-validation MCMC procedure. We take a random sample of size 500 from 3693 individuals for which smoking status (yes/no) has been recorded at 4 different time points (0, 2, 5 and 7 years). Previous analyses ([13]) have indicated that a logistic GEE approach as discussed in Section 2.1 is appropriate. Available covariates are age in years (AGE), a 3 level factor for birth cohort (BIRTH equals 1 if born in 1963-1967, 2 if born in 1955-1962 and 3 if born in 1955-1958), a 3 level factor for attained level of education (EDUCATION is 1 for High School or less, 2 for Some College and 3 for College degree obtained) and a 4 level factor for race/sex group (RACESEX is 1 for black males, 2 for black females, 3 for white males and 4 for white females). Our cross-validation MCMC procedure allows us to efficiently consider candidate models and draw a number of insightful conclusions. We include AGE, BIRTH, EDUCATION, RACESEX as well as interaction between AGE and RACESEX for a total of p=11 covariates (not including the intercept) . Both the consensus model (defined to include all variables occuring in at least 50% of the indistinguishable models) as well as the collection of indistinguishable models are shown in Table 4. Results suggest that attained College degree, black females and age * black female interaction are all significant predictors of smoking status. Due to the interaction we would conclude that variables 1,5,6 and 9 would be most important. Without our cross-validation MCMC procedure available one would likely use a t-test for variable selection. The resulting model with (p=.05) would include only variables 4 and 5. Clearly our procedure provides a great deal of additional information and is likely much more reliable than that of the t-test.

5 CONCLUSIONS

In this paper we propose a cross-validation Markov Chain Monte Carlo procedure as a general variable selection tool which avoids the need to visit all candidate models. This proves particularly useful in the presence of many covariates. We adapt our approach to the context of longitudinal data analysis, while emphasizing that it represents a general technique for variable selection that can be applied whenever a loss function is available (e.g. a measure of predictive ability). Several parameters (M, K_{val}, c, etc., see Section 2.2) allow the user to tune different aspects of the procedure, thereby providing a very flexible tool. Moreover, in contrast to other available techniques, it generates a rich output: not simply a "best" model, but a collection

Table 4: Consensus Model followed by Indistinguishable Models for CARDIA Example. Covariates are indicated by number where 1 corresponds to AGE, 2 and 3 are factors for BIRTH, 4 and 5 are factors for EDUCATION, 6 through 8 are factors for RACESEX, and 9 through 11 represent the interactions between AGE and RACESEX.

	AGE	BIRTH		EDUCATION		RACESEX			AGE*RACESEX		
	1	2	3	4	5	6	7	8	9	10	11
Consensus					x	x			x		
Indist. 1				x	x						
Indist. 2				x	x	x			x		
Indist. 3			x	x	x	x			x		
Indist. 4	x			x	x	x			x		
Indist. 5		x	x		x	x			x		
Indist. 6			x		x	x			x		
Indist. 7	x		x		x	x			x		
Indist. 8		x			x				x		
Indist. 9					x	x			x	x	
Indist. 10		x		x	x	x			x		
Indist. 11		x			x	x					
Indist. 12					x	x	x		x		
Indist. 13			x		x	x	x		x		
Indist. 14	x	x			x	x			x		
Indist. 15		x			x	x			x		
Indist. 16					x	x			x		
Indist. 17					x				x		
Indist. 18	x				x	x			x		
Indist. 19					x	x					
Indist. 20		x			x						
Indist. 21					x						

of interesting models defined with the help of the estimated variability of the optimality measure (e.g. according to the “one-standard error” rule). The information conveyed by this collection of models can be used in many ways; for example, to extract a “consensus” model as defined in Section 3.

Acknowledgments

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Swiss National Science Foundation (1214-66989).

References

[1] Breiman L. Bagging predictors. *Machine Learning* 1996; **24**:123–140.

[2] Cantoni E. A robust approach to longitudinal data analysis. *Canadian Journal of Statistics* 2004; **32**:169–180.

[3] Cantoni E, Mills Flemming J, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. *Biometrics* 2005; to appear.

[4] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.

[5] Efron B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 2004; **99**:619–632.

[6] Field C, Qian G. Law of iterated logarithm and consistent model selection criterion in logistic regression. *Statistics and Probability Letters* 2002; **56**:101–112.

[7] Gelman A, Roberts GO, Gilks WR. Efficient Metropolis jumping rule. In *Bayesian Statistics 5,* Bernardo J, Berger J, Dawid A, Smith A (eds). Oxford University Press: 1996; 599–607.

[8] George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 1993; **88**:881–889.

[9] George EI, McCulloch RE. Approaches for Bayesian variable selection. *Statistica Sinica* 1997; **7**:339–374.

[10] Friedman JH, Hastie T, Tibshiran R. *The Elements of Statistical Learning.* Springer: New York, 2001.

[11] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.

[12] Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**:120–125.

[13] Preisser J, Galecki A, Lohman K, Wagenknecht L. Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of the American Statistical Association* 2000; **95**(452):1021–1031.

[14] Qian G, Field C. Using MCMC for logistic regression model selection involving large numbers of candidate models. In *Monte Carlo and Quasi-Monte Carlo Methods 2000,* Fang K, Kickernell F, Niedrreiter H (eds). Springer: 2001; 461–474.

[15] Shao J. Linear model selection by cross-validation. *Journal of the American Statistical Association* 1993; **88**:486–494.

Publications récentes du Département d'économétrie

pouvant être obtenues à l'adresse suivante :

Université de Genève
UNI MAIL
A l'att. de Mme Caroline Schneeberger
Département d'économétrie
40, Bd du Pont-d'Arve
CH - 1211 Genève 4
ou sur
INTERNET : http//www.unige.ch/ses/metri/cahiers

2004.15 KRISHNAKUMAR Jaya, Marc-Jean MARTIN, Nils SOGUEL, Application of Granger Causality Tests to Revenue and Expenditure of Swiss Cantons, Décembre 2004, 27 pages.

2004.14 KRISHNAKUMAR Jaya, Gabriela FLORES, Sudip Ranjan BASU, Spatial Distribution of Welfare Across States and Different Socio-Economic Groups in Rural and Urban India, Mai 2004, 66 pages.

2004.13 KRISHNAKUMAR Jaya, Gabriela FLORES, Sudip Ranjan BASU, Demand 2004System Estimations and Welfare Comparisons : Application to Indian Household Data, Mai 2004, 70 pages.

2004.12 KRISHNAKUMAR Jaya, Going beyond functionings to capabilities: an econometric model to explain and estimate capabilities, Août 2004, 29 pages.

2004.11 MÜLLER Tobias, RAMSES Abul Naga, KOLODZIEJCZYK Christophe, The Redistributive Impact of Altrnative Income Maintenance Schemes : A Microsimulation Study using Swiss Data, Août 2004, 40 pages.

2004.10 GHIGLINO Christian, DUC François, Expectations in an OG Economy, Août 2004, 21 pages.

2004.09 GHIGLINO Christian, OLSZAK-DUQUENNE Marielle, On the Impact of Heterogeneity on Indeterminacy, Août 2004, 29 pages.

2004.08 FIELD Chris, ROBINSON John, RONCHETTI Elvezio, Saddlepoint Approximations for Multivariate M-Estimates with Applications to Bootstrap Accuracy, Août 2004, 26 pages.

2004.07 VICTORIA-FESER Maria-Pia, CONNE David, A Latent Variable Approach for the Construction of Continuous Health Indicators, Août 2004, 8 pages.

2004.06 DELL'AQUILA Rosario, RONCHETTI Elvezio, Resistant Nonparametric Analysis of the Short Term Rate, Juillet 2004, 26 pages.