

Consistent Model Selection Based on Parameter Estimates

November 16, 2002

Wenxin Jiang¹

Northwestern University

Xiangyang Liu

Johnson & Johnson

Abstract

We consider model selection based on estimators that are asymptotically normal. Such a method can be applied to the context of estimating equations, since a complete specification of the probability model or likelihood function is not required. We construct a cost function for the models in consideration, and show that the minimizer of the cost function is a consistent estimator of the model. Despite the absence of a likelihood function, the cost function is shown to be related to an approximate posterior probability conditional on the parameter estimates, which enables a Bayesian-type evaluation of all candidate models and not just to present one best choice. The proposed method is modular and easily adapted to different problems, since only one set of estimates of the parameters and asymptotic variance is needed as the input, which can be obtained from very different estimation techniques for very different models, both linear and nonlinear. We also show that by ranking Z-statistics, the scope of model searching can be reduced to achieve computing efficiency. We provide data analysis examples from two clinical trials and illustrate these variable selection techniques in the contexts of partial likelihood analysis and generalized estimating equations. A third example of used automobile prices illustrates an application of the methodology in selecting graphical models.

Keywords: Asymptotic normal estimators; Cost functions; Estimating equations; Graphical models; Generalized estimating equations; Model selection; Partial likelihood; Sandwich estimator; Schwarz's Bayesian information criterion (BIC); Wald statistic.

AMS classification codes: 62F99; 62F12; 62-09

¹Address for correspondence: Wenxin Jiang, Department of Statistics, Northwestern University, Evanston, IL 60208, USA.

E-mail: wjiang@northwestern.edu

1 INTRODUCTION

During the last decades of the twentieth century, non-likelihood-based inferential methods have become increasingly popular: quasi-likelihood (see, e.g., McCullagh and Nelder, 1989), estimating equations (see, e.g., Godambe, 1991), generalized estimating equations (GEE, Liang and Zeger, 1986), (generalized) method of moments (see, e.g., Matyas, 1999), least squares, partial likelihood (Cox, 1975), M-estimation (see, e.g., Hampel et al., 1986), to name a few. Many of these methods have developed accompanying computing codes, sometimes incorporated in standard software packages. These methods offer robustness of inferential results of different kinds: robustness against outliers (e.g., in M-estimation), or robustness against departures from distributional assumptions (e.g., in GEE, where a likelihood function is typically not available and the model is only specified by the first two moments). In these situations without specifying a likelihood function, usual model selection criteria such as the Akaike information criterion (AIC) (Akaike, 1974) and Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978) cannot be directly applied. *Is it possible to develop a convenient and modular model selection method, that can be applied to all these different kinds of inferential techniques and directly utilize the outcomes from these standard data analyses, without separate programming and computation on the original data?*

In this paper, we propose a method of model selection based on the estimators of the model parameters, instead of based on the likelihood function and show that the proposed method is consistent (Theorem 1). The proposed model selection criterion is shown to have a Bayesian-type interpretation (Section 3), even if a likelihood function may not be available, which allows evaluation of relative plausibility of candidate models based on an approximate posterior probability conditional on the parameter estimates. Conditions are also provided under which the selection of a compound model (such as a chain graph model) can be done component by component. We also show the validity of an efficient searching algorithm in our context, which only requires searching over a logarithm amount of all model candidates (Theorem 2). The proposed methodology has wide applications — three data examples will be given later in the contexts of generalized estimating equations, partial likelihood analysis, and graphical models where a likelihood

function is not specified. As shown in these examples, the proposed method is modular and convenient to adapt to different problems since all it requires as inputs are one set of estimates for the parameters and the asymptotic variance, which can be generated by very different methods of statistical inference. Therefore we believe that the proposed method will be a useful addition to the wide variety of existing model selection methods, a brief (and noncomplete) survey of which is included at the end of this section.

The development of this methodology is based on the following simple observation. In situations without a specified likelihood function, the estimators are typically still asymptotically normal. Such estimators include those generated by the methods of (generalized) estimating equations, least squares, partial likelihood, and method of moments, to name a few. For example, an estimator solving an estimating equation typically has an asymptotic normal distribution, with a variance estimable from the sandwich formula (see Carroll, Ruppert and Stefanski, 1995, Section A.3). It is therefore natural to construct model selection criteria based on these estimators.

To motivate the problem, we consider the variable selection problem in nonlinear regression. Here it is only assumed that the mean function of the response conditional on the covariates is of the form $E(Y|X_1, \dots, X_p) = \psi(\sum_{j=1}^p \beta_j X_j)$; *a complete probability model is not specified*. Here ψ is the inverse link function.

Example (*Poisson regression with random effects*).

Let $Y|\theta, X_1, \dots, X_p \sim \text{Poisson}(\theta e^{\sum_{j=1}^p \beta_j X_j})$, where θ , called the random effect, has an unknown distribution with mean 1 and is assumed to be independent of the covariates X_j 's. Denote $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ and $\mathbf{X} = (X_1, \dots, X_p)^T$. Then $E(Y|\mathbf{X}) = e^{\mathbf{X}^T \boldsymbol{\beta}}$ [here $\psi = e^{(\cdot)}$]. Let $(Y_i, \mathbf{X}_i, \theta_i)$, $i = 1, \dots, n$, be n i.i.d. (independent and identically distributed) copies of (Y, \mathbf{X}, θ) , where the (Y_i, \mathbf{X}_i) 's are observed, but not the random effects θ_i 's. Without specifying a probability distribution for the random effects, a likelihood function is not available for inference on $\boldsymbol{\beta}$ based on the observed data. However, it is well-known (see White, 1994, Sections 5.1 and 6.2.b) that an estimating equation based on Poisson regression neglecting the random effects will still generate a consistent and asymptotically normal (AN) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

That is, let $\hat{\boldsymbol{\beta}}$ be a solution of the equation $\mathbf{S}(\boldsymbol{\beta}) = \sum_1^n \mathbf{s}_i = \mathbf{0}$, where $\mathbf{s}_i = \mathbf{X}_i(Y_i - e^{\mathbf{X}_i^T \boldsymbol{\beta}})$. Then we have $\hat{\boldsymbol{\beta}} \xrightarrow{D} AN(\boldsymbol{\beta}, \text{var}(\hat{\boldsymbol{\beta}}))$, where the asymptotic variance $\text{var}(\hat{\boldsymbol{\beta}})$ can be estimated by the “sandwich-type” estimate $\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\nabla_{\boldsymbol{\beta}}^T \sum_1^n \mathbf{s}_i)^{-1} (\sum_1^n \mathbf{s}_i \mathbf{s}_i^T) (\nabla_{\boldsymbol{\beta}}^T \sum_1^n \mathbf{s}_i)^{-T} |_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ (see Carroll, Ruppert and Stefanski, 1995, Section A.3).

We consider model selection in such a situation without a likelihood function. The problem is to construct a selection criterion for obtaining the “true model”, which is the set of nonzero β_j 's, without using a likelihood function. We also would like this criterion to penalize the complexity appropriately, similar to BIC.

Note that corresponding to each candidate model M , there is a “null hypothesis” restricting certain components of the vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ equal to zero. Such a hypothesis can be tested by the Wald statistic W based on the estimator $\hat{\boldsymbol{\beta}}$. Our proposed criterion is based on W , with an additional term to penalize the model complexity proportional to the number of parameters times $\log n$. Such a method is shown to be *consistent* in the sense of selecting a candidate model that will recover the underlying true model with probability tending to one as sample size increases. In addition, the method is associated with a Bayesian-type interpretation similar to BIC. We also apply a method of Z -ranking (Zheng and Loh 1995) which leads to a logarithmic reduction of the scope of model searching.

Section 2 describes the formalism of model selection and the consistency results of the proposed methods. A Bayesian-type interpretation similar to BIC is discussed in Section 3. Section 4 considers evaluation of compound models by their components. Numerical studies with simulations and three real data examples are contained in Section 5.

2 MODEL SELECTION

We will consider situations when all candidate models of interest are submodels of a “full model”. For example, in variable selection problems one selects a subset, from a “full model” which contains all candidate explanatory variables of interest, to model the response. Sometimes there is a natural “saturated model” that can be used as the full model. For example, when modeling the covariance structure, an unstructured

covariance matrix is used in the full model and one is interested in selection among submodels that propose covariance matrices with certain patterns. In categorical data analysis the full model can be the one that incorporates all interactions among the discrete variables.

Consider a “full model” which contains p parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. Call this the *full model* Ω , where $\Omega = \{1, 2, \dots, p\}$ comprises the subscripts of the parameters. It is assumed that all candidate models of interest are submodels of this full model. A *candidate model* can be written as $M = \{j_1, \dots, j_{d_M}\}$, which is a subset of Ω of size d_M , containing the subscripts of all nonzero components of $\boldsymbol{\beta}$. This corresponds to a proposal of the following model: $\beta_j \neq 0$ if $j \in M$, and $\beta_j = 0$ if $j \in \Omega - M$. Denote $c_M = \text{card}(\Omega - M) = p - d_M$, and $\Omega - M = \{k_1, \dots, k_{c_M}\}$. Then the resulting proposed model can be represented by a constraint $L_M \boldsymbol{\beta} = \mathbf{0}$, where L_M is a $c_M \times p$ matrix of the form $L_M = (e_{k_1}, \dots, e_{k_{c_M}})^T$, where e_q is a $p \times 1$ vector with the q th component equal to one and all other components equal to zero. An alternative, more general representation of the model is the parameter space Θ_M , which can be a general manifold. We here mainly focus on a linear manifold (or a hyperplane) $\Theta_M = \{\boldsymbol{\beta} : L_M \boldsymbol{\beta} = \mathbf{0}\}$ as a parameter space.

We assume the availability of an asymptotic normal (AN) estimator. That is, there is an AN estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$, such that $\hat{\boldsymbol{\beta}} | \boldsymbol{\beta} \xrightarrow{D} AN(\boldsymbol{\beta}, n^{-1}V(\boldsymbol{\beta}))$ for any parameter $\boldsymbol{\beta}$. Here n is the sample size of the data based on which the estimator $\hat{\boldsymbol{\beta}}$ is computed and $V(\boldsymbol{\beta})$ is assumed to have order one in the matrix sense. We also assume that we have available a consistent estimate $n\hat{\text{var}}(\hat{\boldsymbol{\beta}})$ for $V(\boldsymbol{\beta})$ and denote $\text{se}(\hat{\beta}_j) = \sqrt{\{\hat{\text{var}}(\hat{\boldsymbol{\beta}})\}_{jj}}$ for each $j = 1, \dots, p$.

Suppose that the observed data is “generated” by a special parameter $\boldsymbol{\beta}$ under a *true model* $M_0 = \{j_1^0, \dots, j_{d_0}^0\}$, where M_0 contains the subscripts of all nonzero components of $\boldsymbol{\beta}$. In model selection, we would like to identify M_0 from the data, among all candidate models M in a certain *scope* Φ . Here a scope is a class of subsets of Ω . A trivial example of a scope Φ is $po(\Omega)$, the class of all subsets of Ω , which includes 2^p candidate models. Another example is the *Z-scope* constructed on Ω , which is a random scope denoted as $\Phi_Z = \{M_{(1)}, \dots, M_{(p)}\}$. Here $M_{(j)} = \{i : i \in \Omega, |Z_i| \geq |Z_{(j)}|\}$ for each $j = 1, \dots, p$, where $Z_i = \hat{\beta}_i / \text{se}(\hat{\beta}_i)$ is a usual Z-statistic, and the $Z_{(j)}$ is the j th largest of the Z_i 's. The Z-scope basically contains p increasingly complex candidate models, the j th of which contains the parameters with j largest

Z-values. This scope has size p , which is much smaller than 2^p and is used in Zheng and Loh (1995) for linear model selection with sub-Gaussian noises, in which case they refer to the Z_j 's as the t statistics.

Now we define the *cost* function $C(M)$ of a model M . Let $C(M) = W(M) + d_M h_n$, where $W(M) = \inf_{\beta \in \Theta_M} (\hat{\beta} - \beta)^T \{\hat{\text{var}}(\hat{\beta})\}^{-1} (\hat{\beta} - \beta)$ or $(L_M \hat{\beta})^T \{\hat{\text{var}}(L_M \hat{\beta})\}^{-1} (L_M \hat{\beta})$ is the Wald statistic for the null hypothesis $L_M \hat{\beta} = \mathbf{0}$ corresponding to the model M , $d_M = \text{card}(M) = \dim(\Theta_M)$ is the size of the model. We mainly consider the case when the *growth function* $h_n (> 0)$ is chosen to be the BIC-type $\log n$, while other choices are also possible, as seen in Lemma 2 in the Appendix. Similar to the ordinary BIC, we obtain an estimator \hat{M} of the model by $\hat{M} = \arg \min_{M \in \Phi} C(M)$ over a scope Φ .

The following is a summary of some results on how faithful the estimator \hat{M} is in recovering the true model M_0 . The proofs of the results are given in the Appendix.

First, we claim that \hat{M} is consistent as an estimator of the true model M_0 , if the true model is contained in the scope of model search:

Theorem 1 (Consistency). *If $M_0 \in \Phi$ (a finite scope), then, with probability tending to one as n increases, the true model becomes the unique minimizer of the cost; i.e., $P[\hat{M} \text{ is unique and equal to } M_0] \rightarrow 1$ as $n \rightarrow \infty$.*

The consistency result actually remains valid if we relax the condition to a statement ‘in probability’: We only need $P[M_0 \in \Phi] \rightarrow 1$ as $n \rightarrow \infty$, if Φ is data-driven such as the Z-scope Φ_Z . Indeed, the true model is contained in the Z-scope “in probability” (see Lemma 1 in Appendix). Therefore we obtain the consistency result when we search for the best model $\hat{M}_Z = \arg \min_{M \in \Phi_Z} C(M)$ over the Z-scope Φ_Z :

Theorem 2 (Consistency with the Z-scope). *Suppose $M_0 \subset \Omega$. Let $\hat{M}_Z = \arg \min_{M \in \Phi_Z} C(M)$, where Φ_Z is a Z-scope constructed on Ω . Then $P[\hat{M}_Z \text{ is unique and equal to } M_0] \rightarrow 1$ as $n \rightarrow \infty$.*

Note that $\text{card}(\Phi_Z) = p$, which is much smaller than $\text{card}(po(\Omega)) = 2^p$. Hence Lemma 1 and Theorem 2 suggest that we only need to search among p models to find a consistent estimator \hat{M}_Z . These results ensure that in the very general situations we consider (e.g., nonlinear regression models without a complete probability model specified), there still exist reliable (consistent) and efficient model searching techniques.

The method is also modular in the sense that it can be easily adapted to various problems. What is needed is only one set of input: $\hat{\beta}$ and $\text{var}(\hat{\beta})$ from a full model, which can be obtained from very different estimation techniques. This will be testified later by the variety of examples in Section 5.

2.1 Related Works

There have been extensive work on model selection. We here provide a brief (and noncomplete) survey with focus on works that are related to the current paper.

Previously, frequentist properties of model selection have been studied in linear regression when a probability model (usually normal) for the response is specified, see, e.g., Shibata (1981), Zhang (1992) and Zheng and Loh (1995). There are also several known selection approaches (e.g., bootstrap, cross validation, and methods that are robust against outliers) that have been proposed for non-likelihood based models. Some useful references are Shao (1993, 1996), Machado (1993), Ronchetti and Staude (1994), Burman and Nolan (1995), Ronchetti (1997), Shi and Tsai (1998), Pan (1999), and Tibshirani and Knight (1999). These methods typically involve modification of standard estimation procedures or special treatment of the observed data. Non-likelihood-based model selection procedures are often constructed based on predictive performances such as cross entropy, expected predictive bias, predictive mean square error (or their AIC-type approximations), rather than based on consistency considerations. See, e.g., Pan (2001a,b) for GEE and estimating equations, Hurvich and Tsai (1995) for quasi-likelihood, and Sommer and Huggins (1996) for a C_P -type method based on Wald statistic (which is applied to linear and logistic regressions), without the consistency properties or Bayesian interpretation.

In the general situations with nonlinear models without a likelihood (or even without a scalar objective function), our proposed method offers both consistent model selection and a Bayesian-type evaluation of model candidates. In contrast to other consistent model selection techniques, our approach can also naturally accommodate a wide variety of estimation techniques (e.g., GEE, partial likelihood, method of moments, or methods that are robust against outliers), and directly utilize their inferential results for model selection, without reprogramming on the original data.

3 RELATION TO THE BAYES FACTOR

The choice $h_n = \log n$ is related to the BIC and the Bayes factor (see, e.g., Kass and Raftery, 1995). Such a relation will be explored in Proposition 1 of this section, which will enable us to assess the relative plausibility for all the candidate models based on an approximate posterior probability, despite the absence of a likelihood function. This way, not only a ‘point estimate’ \hat{M} can be provided in our approach, but also other competing candidate models together with their approximate posterior probabilities relative to \hat{M} .

Here we do not consider the full posterior probability $P(M|\text{full data})$ of model M , since there is no complete probability model assumed for the data (often only some moments are specified). However, there is a “*partial posterior*” $P(M|\hat{\beta}) = P(\hat{\beta}|M)P(M)/P(\hat{\beta})$, based on the AN estimator $\hat{\beta}$ (instead of based on the full data). (Here P is the probability function or density function, for, respectively, a discrete or continuously-supported random quantity.) As we will show below, despite the absence of a complete probability model, the “*partial posterior*” $P(M|\hat{\beta})$, or $P(\hat{\beta}|M)$, can still be approximately evaluated based on a Laplace approximation and the AN property of $\hat{\beta}$.

Under model M , the parameter β lies in a d_M -manifold (e.g., the hyperplane $\{\beta : L_M\beta = \mathbf{0}\}$) and can be (smoothly) parameterized as $\beta = \theta(\beta_M)$, say, by a d_M -dimensional parameter β_M in B_M . Here $B_M = \Re^{d_M}$, say, and β_M consists of all $(\beta)_j$, $j \in M$. Note that in this notation, Θ_M can be represented by $\theta(B_M)$, and the Wald statistic becomes $W(M) = \inf_{\beta_M \in B_M} \{\hat{\beta} - \theta(\beta_M)\}^T \{\text{var}(\hat{\beta})\}^{-1} \{\hat{\beta} - \theta(\beta_M)\}$. We can also write

$$P(\hat{\beta}|M) = \int_{B_M} P(\hat{\beta}|\theta(\beta_M))P(\beta_M|M)d\beta_M. \quad (1)$$

Under regularity conditions and applying the Laplace approximation (see, e.g., Draper, 1995, eqn. (11); Kass, Tierney and Kadane, 1990), we obtain $\log P(\hat{\beta}|M) = (1/2)d_M \log(2\pi/n) + \log P(\hat{\beta}|\theta(\hat{\beta}_M)) + O(1)$ where $\hat{\beta}_M = \arg \max_{\beta_M \in B_M} \log P(\hat{\beta}|\theta(\beta_M))$. Hence

$$-2 \log P(\hat{\beta}|M) = d_M \log(n) - 2 \sup_{\beta_M} \log P(\hat{\beta}|\theta(\beta_M)) + O(1) \equiv -2\text{PBIC} + O(1). \quad (2)$$

Now by the asymptotic normality of $\hat{\beta}|\beta$,

$$-2 \log P(\hat{\beta}|\beta) \text{ is asymptotically equivalent to } (\hat{\beta} - \beta)^T \{\text{vâr}(\hat{\beta})\}^{-1} (\hat{\beta} - \beta) + \log |2\pi \text{vâr}(\hat{\beta})|. \quad (3)$$

Then $-2 \sup_{\beta_M} \log P(\hat{\beta}|\theta(\beta_M))$ is asymptotically equivalent to $\inf_{\beta_M} (\hat{\beta} - \theta(\beta_M))^T \{\text{vâr}(\hat{\beta})\}^{-1} (\hat{\beta} - \theta(\beta_M)) + \log |2\pi \text{vâr}(\hat{\beta})|$ or $W(M) + \log |2\pi \text{vâr}(\hat{\beta})|$.

These discussions immediately lead to the following proposition.

Proposition 1 (*Partial Bayes Factor*). *Based on the AN approximation (3) and the Laplace approximation (2), we have*

(i) $-2 \log P(\hat{\beta}|M) = -2\text{PBIC} + O(1) = W(M) + d_M \log(n) + \log |2\pi \text{vâr}(\hat{\beta})| + O(1)$.

(ii) $-2 \log P(M|\hat{\beta}) = -2 \log P(\hat{\beta}|M) - 2 \log P(M) + 2 \log P(\hat{\beta})$.

(iii) *Let the ‘partial Bayes factor’ for models M_1 and M_2 be $P(\hat{\beta}|M_1)/P(\hat{\beta}|M_2) = \text{PBF}_{12}$. Then*

$$-2 \log \text{PBF}_{12} = \{W(M_1) + d_{M_1} \log n\} - \{W(M_2) + d_{M_2} \log n\} + O(1) = C(M_1) - C(M_2) + O(1).$$

(iv) *If $-2 \log\{P(M_1)/P(M_2)\} = O(1)$, then $-2 \log\{P(M_1|\hat{\beta})/P(M_2|\hat{\beta})\} = -2 \log \text{PBF}_{12} + O(1)$.*

(v) *Let $\hat{M} = \arg \min_{M \in \Phi} C(M)$ where $C(M) = W(M) + d_M \log n$. Suppose $-2 \log\{P(M_1)/P(M_2)\} = O(1)$ for all M_1, M_2 in Φ . Then $\hat{M} = \arg \max_{M \in \Phi} \log Q(M|\hat{\beta})$ where $\log Q(M|\hat{\beta}) = \log P(M|\hat{\beta}) + O(1)$.*

Therefore the minimizer of $C(M)$ is equivalent to the maximizer of the leading order terms of the partial posterior $P(M|\hat{\beta})$ due to (v), and equivalent to the maximizer of the *partial BIC* PBIC in (2).

These observations allow us to provide an assessment of how likely a candidate model M is relative to the optimal candidate \hat{M} , by reporting the leading order posterior ratio or *partial posterior ratio* PP-Ratio = $P(M|\hat{\beta})/P(\hat{M}|\hat{\beta}) \approx \exp[-0.5\{C(M) - C(\hat{M})\}]$ based on the parameter estimate $\hat{\beta}$. For example, in some numerical examples later, we report all candidate models M together with the ratios $P(M|\hat{\beta})/P(\hat{M}|\hat{\beta})$, for all M such that the ratio is larger than 0.05 (corresponding to a difference of less than about 6 in the scale of the cost function). They are considered as reasonable model candidates that are competitors of \hat{M} and investigated further.

4 SELECTING COMPOUND MODELS IN PARTS

Suppose we are considering complex composite models that have several parts, each being a simpler component model. In model selection, do we have to treat the composite models by themselves, with higher dimensional parameters, or can we break the task down to selecting the separate components (with lower dimensional parameters) and combine them later? Obviously, the latter method, if legitimate, is more attractive due to computing ease and modularity.

In these situations the model parameter $\beta \in \Theta$ can be decomposed into several parts, $\beta^T = (\beta_a^T)_{a=1}^A$, say, where each sub-parameter β_a lies in a sub-parameter space Θ^a and is estimated by $\hat{\beta}_a$ from solving a separate estimating equation $\mathbf{G}_a(\beta_a) = \mathbf{0}$. A model M , which represents a manifold $\Theta_M \subset \Theta$, can also be decomposed into a Cartesian product $\otimes_{a=1}^A \Theta_M^a$ of sub-manifolds $\Theta_M^a \subset \Theta^a$, each corresponding to a sub-model M^a for the part- a parameter β_a . Symbolically we can write $M = \otimes_{a=1}^A M^a$.

One example is our third data analysis example where models represented by the chain graphs can be parameterized in a collection of parameter vectors from all the blocks. A brute force application of the proposed procedure would involve stacking all A parts of the estimating equations, using the sandwich formula to estimate the asymptotic variance, and computing the Wald statistic and the cost function for each compound model candidate. This could be clumsy to implement, especially when A is large, or when the parts of the compound model are to be treated in dissimilar manners. A natural attempt is to select the compound model “part by part”, which would be much more convenient. This can be done if the cost function and the Wald statistic for a compound candidate model decompose into A additive parts for the corresponding component models.

The next proposition summarizes situations when the cost functions are decomposable into natural parts, which happen when the estimating functions for the components of the model satisfy a martingale-type orthogonality condition. Selection of a composite model can then be performed by combining the selections from all the components, which makes the task computationally easier and more modular.

Proposition 2 (*Decomposition of the Cost Function*). *Suppose a sequence of estimating functions $\{\mathbf{G}_a\}_1^A$*

(evaluated at the true parameter) form a martingale, i.e., suppose $\mathbf{G}_a \in \mathcal{B}_a$ and $E(\mathbf{G}_a | \mathcal{B}_b) = \mathbf{0}$ for all $b < a$, where $\{\mathcal{B}_a\}_1^A$ is an increasing sequence of sigma fields. Then $C(\otimes_1^A M^a) = \sum_1^A C(M^a)$, $W(\otimes_1^A M^a) = \sum_1^A W(M^a)$, $d_{\otimes_1^A M^a} = \sum_1^A d_{M^a}$. Here $W(M_a) = \inf_{\beta_a \in \Theta_M^a} (\hat{\beta}_a - \beta_a)^T \hat{\text{var}}(\hat{\beta}_a)^{-1} (\hat{\beta}_a - \beta_a)$.

A sketch of proof is given in the Appendix.

5 NUMERICAL STUDIES

There are wide applications of the proposed methodology. We will illustrate the methodology with three real data sets. In each of the three examples, a complete probability model is not assumed and we use estimation techniques without using a likelihood function. It is noted that the estimation methods involved are very different, which are, respectively, generalized estimating equations (GEE), partial likelihood, and a combination of least squares and method of moments. However, all these estimators are typically AN and easily incorporated in our method for model selection.

It is noted that in these situations when a likelihood is not available (e.g., in GEE or partial likelihood analysis), the most commonly used method for model selection is based on examining the individual Z-values of the parameter estimates. No previous method penalizes model complexity and guarantees consistency for model selection in these situations as the proposed method does. Nor do previous methods provide a measure of the relative plausibility of model candidates as the proposed method does based on leading order partial posterior probabilities.

Before we present the real data examples, we first describe some simulation results about the finite sample performance of the proposed method.

5.1 Simulations

The performance of the proposed technique is tested following a model based on the example in the Introduction. In one situation, for example, the “unknown” distribution of the “unobserved” random effect θ is taken to be gamma with mean 1 and variance 2 in order to generate the simulated data sets. Five covariates $(X_1, X_2, X_3, X_4, X_5)$ are of interest ($p = 5$), including a constant 1 and four independent

uniform random variables in $[0, 1]$. An “observed” data set includes n i.i.d. copies of (Y, X_1, X_2, X_3, X_5) , where Y is generated from $\text{Poisson}(\theta e^{\sum_{j=1}^5 \beta_j X_j})$ where $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (1, 1, 1, 0, 0)$. On the other hand, in data analysis, suppose due to the lack of knowledge one chooses not to specify a probability model for the random effects, then a likelihood based treatment cannot be implemented. Instead, we perform variable selection with the use of Z-ranking based on the parameter estimates from the QL/M (quasi-likelihood/moment) estimating equations (see, e.g., Breslow 1984 or Lawless 1987), which does not require a probability model (such as gamma) for the random effects. A successful selection corresponds to $\hat{M} = \{1, 2, 3\}$. The proportion of successes are computed out of 500 repeated experiments, under several choices of the sample size n . When the sample size n increases from 50, 100 to 200, the proportion of successful model selections increases from 0.752, 0.914 to 0.972. We also compared the performance of the usual BIC method based on the true likelihood function, pretending that we know the true probability model (that is, a gamma mixture of Poisson regression, or a negative binomial regression model). The corresponding proportions of successes were 0.844, 0.900, and 0.964. In simulations not reported here, alternative sets of true parameters were also used, as well as covariates that are correlated with a varying degree of multicollinearity. In all the cases considered, the small sample performance of the proposed model selection method was comparable to (and sometimes better than) the model-dependent BIC method based on the true likelihood function (which in practice may not be available).

5.2 GEE—Epileptic Seizures Data

This example illustrates the application of the proposed methodology for model selection in generalized estimating equations, where a likelihood function is not assumed.

We use the epileptic seizures data analyzed by Diggle, Liang and Zeger (1994). It comprises data from a clinical trial with 59 epileptic patients. For each patient, the number of epileptic seizures was recorded during a baseline period of eight weeks. Patients were then randomized to treatment with the anti-epileptic drug progabide, or to placebo in addition to standard chemotherapy. The number of seizures was then recorded in four consecutive two-week intervals. Diggle et al. (1994, p.167) used log-linear regression with

the GEE method to study the overall treatment effect on the rate of epileptic seizures. (The effect was found to be nonsignificant at the 0.05 level.) The model they use assumes a constant trend for the rate of seizures after the randomization. In this section we will use the proposed method to investigate whether such a constant model is the most appropriate one, as compared to other models in low order polynomials.

Let Y_{ij} be the number of epileptic seizures the i -th patient has in the j -th period [$j=0$ (baseline), 1, 2, 3, 4]; $z_i=1$ if the i -th patient is in the treatment group, 0 otherwise. We use a log-linear structure (similar to Diggle et al. 1994) for our analysis, which allows different and arbitrary trends for the rates of seizures in the two groups of patients. For $i=1, 2, \dots, n$, $j=0, 1, 2, 3, 4$, this log-linear structure takes the form $\log E(Y_{ij}) = \mu_j + \gamma_j z_i$.

We consider the following four candidate models for the parameters μ_j and γ_j : for $k=0, 1, 2, 3$, model M_k is parameterized by $\beta = \theta(\beta_{M_k})$ where $\beta_{M_k} = (\mu_0, \gamma_0, \beta_0, \alpha_0, \dots, \beta_k, \alpha_k)$, $\beta = (\mu_0, \gamma_0, \dots, \mu_4, \gamma_4)$, and $\theta(\cdot)$ is defined by $\mu_j = \mu_0 + \mathbf{1}(j > 0) \sum_{l=0}^k \beta_l t_j^l$ and $\gamma_j = \gamma_0 + \mathbf{1}(j > 0) \sum_{l=0}^k \alpha_l t_j^l$ for $j=0, 1, 2, 3, 4$. The model M_k models the post-randomization log(event rate) as a k th-order polynomial in time. Note that the offset term accounting for a different observation period at $j = 0$ is absorbed in μ_0 . Here, $t_0 = -4$, $t_j = 2j - 1$ for $j=1, 2, 3, 4$. Note that the full model (or the saturated model) is M_3 , which contains all candidate models of interest here $[(M_k)_0^3]$ as submodels, and can be parameterized either by β_{M_3} or by β .

In this context no likelihood function is available. We perform model selection based on the GEE estimates $\hat{\beta} = (\hat{\mu}_0, \hat{\gamma}_0, \dots, \hat{\mu}_4, \hat{\gamma}_4)$ from the full model M_3 , with the variance estimates $\text{var}(\hat{\beta})$ based on the sandwich formula. The estimated cost functions $C(M_k)$ for these four models (M_k for $k = 0, 1, 2, 3$) can be easily obtained from minimizing a quadratic function as follows, with results listed in Table 1: $C(M_k) = (2k + 4) \log n + \inf_{\beta_{M_k} \in \mathfrak{R}^{2k+4}} (\hat{\beta} - \beta)^T \{\text{var}(\hat{\beta})\}^{-1} (\hat{\beta} - \beta)$, where $\beta_{M_k} = (\mu_0, \gamma_0, \beta_0, \alpha_0, \dots, \beta_k, \alpha_k)$ and $\hat{\beta} - \beta = \left[\hat{\mu}_0 - \mu_0, \hat{\gamma}_0 - \gamma_0, \dots, \hat{\mu}_4 - (\mu_0 + \sum_{l=0}^k \beta_l t_4^l), \hat{\gamma}_4 - (\gamma_0 + \sum_{l=0}^k \alpha_l t_4^l) \right]^T$.

The reported values in Table 1 were obtained using an AR(1) working correlation structure when computing the GEE estimate $\hat{\beta}$. Also, as in Diggle et al. (1994), we have excluded the patient with ID number 207 which was regarded as an outlier. Analyses with other choices of correlation structure or inclusion of the outlier follow the same procedure and gave similar results. Here the PP-Ratio in the

last column is the partial posterior ratio of a model M when compared to the optimal candidate $\hat{M} = \arg \min_{M \in \Phi_Z} C(M)$, based on the leading order approximation of the partial Bayes factor in Proposition 1.

[Table 1 ABOUT HERE.]

Table 1 highly favors the constant trend model M_0 . The second best model (the linear trend model) is only 1.7% as likely. Therefore the model used by Diggle et al. (1994, page p.167) seems to be compatible with the data, which assumes that the post-randomization event rate of epileptic seizures does not change over time.

5.3 Cox’s Model—Lung Cancer Data

In Cox regression with a proportional hazard model for the failure time, the baseline hazard function is not specified and a complete probability model is not available. The proposed method can be used to perform variable selection in this context based on the partial likelihood estimates.

The example involves the Veteran’s Administration lung cancer data (Kalbfleisch and Prentice, 1980, Chapter 3). In this clinical trial, 137 males with lung cancer were randomized to either a standard or test chemotherapy, among whom 97 had no prior therapy. The models for the patients with and without prior therapy can be different and we report the results on the group without prior therapy. The procedure of treating the other part of the data is similar.

The primary endpoint for therapy comparison was Time to death Y_i (in days). The following covariates are considered: Treatment indicator X_1 (=1 if ‘test’, or 0 if ‘standard’); Tumor type indicators X_2, X_3, X_4 (for ‘squamous’, ‘small cell’ and ‘adeno’ types, respectively); Karnofsky rating X_5 ; Time from diagnosis to randomization X_6 (in months); Age X_7 (in years).

The objective is then to select an appropriate proportional hazard model for these data. Consider a full model under which the partial likelihood function is $L(\beta) = \prod_{i=1}^{97} [\exp(\mathbf{X}_i\beta) / \sum_{j \in R_i} \exp(\mathbf{X}_j\beta)]^{\delta_i}$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{i7})$ is the row vector of covariates associated with patient i , β is the 7×1 column vector of parameters, δ_i is the survival status of patient i at the end of study (0=censored, 1=death), and R_i is the set of labels attached to the individuals at risk at time Y_i .

Let $\hat{\beta}$ be the value of β which maximizes $L(\beta)$ and $\text{var}(\hat{\beta})$ be the sandwich estimate of the asymptotic variance (Lin and Wei, 1989). Then $C(M) = \inf_{\beta \in \Theta_M} (\hat{\beta} - \beta)^T \{\text{var}(\hat{\beta})\}^{-1} (\hat{\beta} - \beta) + \dim(\Theta_M) \log n$. Here Θ_M is the hyperplane explained in Section 2, where the β -components corresponding to the covariates not selected by model M are restricted to be zero and all other components are free to vary. Based on ranking the Z -values, we found that the Z -scope contains the following candidate models: M_k for $k = 1, \dots, 7$, where M_k is the subset of the first k indexes from $\{1, 3, 4, 5, 2, 6, 7\}$. The values of the cost functions $C(M_k)$ for M_1 to M_7 can then be easily computed by minimization of quadratic functions, with results shown in Table 2.

[Table 2 ABOUT HERE.]

Therefore M_4 is found to be the optimal model (with treatment selection, small cell cancer type, Karnofsky rating and adeno lung cancer as predictors). The next best model (14.6% as likely) includes squamous tumor type as an additional covariate. The third best model (without the Karnofsky rating as a predictor) is only 7.6% as likely. A similar analysis was run for the group of patients with prior therapy. It was found that while Karnofsky rating and adeno lung cancer remain as important predictors for both data, treatment selection and small cell cancer type are no longer included in the optimal model for patients with prior therapy.

5.4 Graphical Model—Used Automobile Prices

Now we consider an example for graphical model selection, where no standard distribution can be used to model the multivariate data and a likelihood function is not available.

In this example, unlike in the last two, we not only model the relationship between the response and explanatory variables, but also model the relationship among the explanatory variables themselves. The model can be summarized in a chain graph (see Whittaker 1990, Sec. 3.6; or Cox and Wermuth 1993) with two blocks, the first part being an undirected graph summarizing the relationship among the explanatory variables, and the second part being the response variable pointed to by arrows coming from significant explanatory variables. Typically (as is the case in our example below), the martingale condition of the

estimating functions employed in the two parts is satisfied, the cost function is decomposable, and the model selection can be naturally done separately in two parts due to Proposition 2. One part is for the regression variable selection which is similar to what we did in the first two data analyses, and the other part is for the covariance selection or the detection of zero partial correlations (see Whittaker 1990, Ch. 1). The example we consider is a situation when a usual likelihood-based treatment is not a reasonable choice, because of the difficulty of finding a suitable probability model for the multivariate data set that we consider. The proposed method, on the other hand, can be applied, based on the least-squares regression coefficients and the sample concentration matrix (or inverse sample variance, see Whittaker 1990, Ch. 5).

The data set is constructed from ‘Automobiles for Sale’ that were advertised on June 1, 2000, from a website of a local newspaper serving suburban Chicago at <http://www.pioneerlocal.com/> [see Pioneer Press Online (2000)]. The sample includes 86 used automobiles which are ten years old or less and which have all the following four variables available:

Y or ‘AD’: the advertised price of the used automobile.

X_1 or ‘YR’: the year when the automobile was made, between 1990 and 1999 (2000).

X_2 or ‘MI’: the mileage in thousand miles.

X_3 or ‘LS’: the original listed price (or the ‘new car price’ for the same model) from the Blue Book (2000).

Among the four variables, AD is the response variable while the other three are explanatory. The inter-relationship of the four variables can be summarized as a chain graph, such as the ones in Figure 1. There, a lack of an undirected edge between two different explanatory variables X_i and X_j means that the partial correlation $\text{pcor}(X_i, X_j | \text{all other } X\text{'s})$ is zero; and a lack of an arrow from X_k to Y means that the partial correlation $\text{pcor}(Y, X_k | \text{all other } X\text{'s})$ is zero, or equivalently, that the regression coefficient β_k is zero in the multiple regression of Y on all the explanatory variables. In the multivariate normal case the zero partial correlation (*pcor*) is equivalent to conditional independence. In more general situations, a zero $\text{pcor}(A, B | C)$ describes a kind of linear irrelevance between A and B given C — the precise sense in terms of the best linear prediction is seen, for example, in Whittaker (1990, Ch. 5).

Usually, model selection for detecting zero pcors are performed based on the likelihood ratio tests, see,

e.g., Whittaker (1990, Ch. 6, 8), based on the multinormal assumption. Here, however, the multinormal assumption is hard to justify — the AD and LS prices are skewed to the right, while the YR is bounded and has an irregular histogram. In this case we will use the proposed approach based on the parameter estimates. Since a complete probability model (such as multinormal) is not assumed, the resulting graphs are linear irrelevance graphs, rather than the ordinary graphs for probabilistic conditional independence.

Before the data analysis, we based on our judgment to establish a prior guess of the resulting chain graph, let us refer to it as the ‘guessed graph’, which is Figure 1 (C). The intention was to later compare this guess with what the data actually would tell us. It seemed to us that all three predictors should be strongly influential to the response, so we expected all three arrows to be present. Among the three predictors, the YR and MI should obviously increase with each other fixing LS (the original listed price, which is related to the type/ model of the automobile). Less clear to us was the existence of a relationship between LS and another predictor, given the third. However, we conjectured that there might exist a moderate link between LS and YR given MI. Note that newer automobiles may tend to have higher original prices due to the factor of inflation—this relation should exist marginally (unconditional on MI), but we guessed that even conditionally we may still see a weak partial correlation.

The data analysis involves model selection in two parts due to Proposition 2. (Here the sigma-fields \mathcal{B}_1 and \mathcal{B}_2 are generated by X_1, X_2, X_3 and by X_1, X_2, X_3, Y , respectively.) The two cost functions can be added later for selecting the combined graph. Note that here \mathbf{G}_2 is the gradient of the least-squares criterion function. In the choice of the arrows, we performed least-squares regression using the sandwich formula for estimating the asymptotic variance (this is protective against the heteroschedasticity). (Note that testing a zero least-squares regression coefficient is equivalent to testing a zero partial correlation.) The details are omitted due to the similarity with the two earlier data analysis examples. As a conclusion, we found that the model with all three arrows are dominantly favored. The next best model is less than 1/10,000 times as likely, with a cost increase as high as 22.

In the other part, we perform the ‘covariance selection’ for establishing the undirected graph, where the absence of an edge represents a zero partial correlation given all other variables in the block. Here

it is somewhat more convenient to use the elements of the concentration (or inverse variance) matrix $D_{ij} = [\text{var}\{(X_k)_1^3\}^{-1}]_{ij}$ to parameterize the model. Note that a zero off-diagonal elements D_{ij} is equivalent to the zero pcor between X_i and X_j given all other X 's. For this part we define the parameter (omitting the subscript 1 for this part of the combined model) to be β =(all means and concentration matrix elements), $\hat{\beta}$ =(all the corresponding sample quantities). The asymptotic variance of $\hat{\beta}$ can be obtained from the sandwich formula, by recognizing that these sample estimates can be solved from the method of moment estimating equations $\hat{s} - E(\hat{s}|\beta) = 0$, where \hat{s} includes all the first and second order sample moments. Then, the Wald statistic and the cost function can be computed (from a trivial minimization of a quadratic function) to test a candidate model where a number of concentration elements (or pcors) are constrained to be zero. The cost function is $C(M) = \inf_{\beta \in \Theta_M} (\hat{\beta} - \beta)^T \{\text{var}(\hat{\beta})\}^{-1} (\hat{\beta} - \beta) + \dim(\Theta_M) \log n$ where Θ_M imposes 0-concentration constraints in model M . E.g., for model D in Figure 1, D_{12} and D_{23} are constrained to be zero (as components of β) in Θ_M since $0 = \text{pcor}(X_1, X_3|X_2) = \text{pcor}(X_2, X_3|X_1)$ [note that $(X_1, X_2, X_3) = (\text{YR}, \text{MI}, \text{LS})$.]

In such a problem with a small number (eight) of candidate models, an exhaustive search is feasible and presents a best model A in Figure 1. (Incidentally the Z-scope method gives the same optimal choice as the all-subset method. The same happened to the selection of arrows.) We list the four most likely models with smallest costs in graphs A to D, together with how likely they are relative to the optimal choice, by presenting the leading order posterior ratios based on the parameter estimates. Recall that the guessed model was C, which is about 17% as likely as A. Both A and B (39% as likely as A) are more likely than C, while D is less than 1% as likely as A, which states that LS is linearly irrelevant for both YR and ML. Based on these approximate posterior ratios, it seems that all three top choices (A, B, C) are worth serious consideration, noting that they are all more than 5% as likely as A.

[Figure 1 ABOUT HERE.]

In all these models, a nonzero pcor(YR, MI|LS) is present (sample value being -0.63). In the optimal model, there is also a nonzero pcor(LS, MI|YR) depicted. (The sample value is -0.19 .) This last relation was not expected before analyzing the data. A plausible explanation is that among the automobiles that

are made in a same year, perhaps more expensive automobiles (with higher LS) are used less often for some reason (and thus tend to have a lower MI). On the other hand, the conjectured edge between LS and YR due to the inflation factor is missing in this graph [sample $\text{pcor}(\text{YR}, \text{LS}|\text{MI})=0.14$]. The reason may be that the inflation effect should mostly exist on a marginal scale. [Unconditional on MI, the sample correlation $\text{cor}(\text{YR}, \text{LS})=0.33$ and is larger.] After adjusting for MI, there is no longer an obvious relationship from the subject matter consideration. Nevertheless, a saturated model B, where such a conditional relationship is present, is still a likely model (39% as likely as A).

6 DISCUSSION

In this paper we present a widely applicable, consistent, modular, and computationally efficient method of model selection using the Wald-statistic-based cost functions, which also extends the BIC to inference conditional on the parameter estimates instead of conditional on the full data. This method is applicable to many situations without the specification of a complete probability model, where the likelihood function is not available and the usual BIC cannot be directly implemented. We provided some examples in this paper to illustrate model selection based on the partial likelihood estimates, the GEE estimates, and the parameter estimates used in graphical models. We expect that the proposed method, both being consistent in the frequentist point of view and allowing a Bayesian-type evaluation of the models, will become a very useful tool for model selection for such situations without a likelihood.

ACKNOWLEDGMENTS

This work was based in part on the second author's Ph.D. thesis. The authors are grateful to Martin Tanner for reading the manuscript and providing useful comments, and to the referee and editors for helpful suggestions and for providing additional references.

REFERENCES

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Auto. Control* **19**,

716-723.

Blue Book (2000). *Used Car Guide, Consumer Edition, July-December 2000*. 2000 Carfax, Inc.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38-44.

Burman, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika*, **82**, 877-886.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, New York.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. (With discussions.) *Statistical Science*, **8**, 204-283.

Dawid, A. P. (1998). Conditional independence. *Encyclopedia of Statistical Sciences, Update Volume 2*, ed. Kotz, S., Read, C. B. and Banks, D. L., 146-155. Wiley, New York.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. R. Statist. Soc. B*, **57**, 45-97.

Godambe, V. P. (1991). *Estimating Functions*, Clarendon Press, Oxford.

Hurvich, C. M. and Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, **51**, 1077-1084.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kass, R. E. and Raftery, A. (1995). Bayes factors. *J. Am. Statist. Assoc.*, **90**, 773-795.

- Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior asymptotic expansion on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geisser, J. S. Hodges, S. J. Press and A. Zellner eds., North-Holland, New York.
- Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209-225.
- Liang, K. -Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazard Model, *The Journal of the American Statistical Association*, **84**, 1074-1078.
- Machado, J. A. F. (1993). Robust model selection and M -estimation. *Econometric Theory*, **9**, 478-493.
- Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *J. Comput. Graph. Statist.*, **8**, 109-120.
- Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.
- Pan, W. (2001b). Model selection in estimating equations. *Biometrics*, **57**, 529-534.
- Pioneer Press Online (2000). Automobiles for Sale, June 1, 2000. *2000 Pioneer Press Newspapers and the Chicago Sun-Times Co.* <http://www.pioneerlocal.com/>
- Ronchetti, E, and Staudte, Robert G. (1994). A robust version of Mallows' C_P . *J. Amer. Statist. Assoc.* **89**, 550-559.
- Schwarz, G. (1978). Estimating a dimension of a model. *Ann. Statist.*, **6**, 461-464.
- Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.
- Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91**, 655-665.

- Shi, P. and Tsai, C.-L. (1998). A note on the unification of the Akaike information criterion. *J. R. Stat. Soc. Ser. B* **60**, 551-558.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45-54.
- Sommer, S. and Huggins, R. M. (1996). Variable selection using the Wald test and a robust C_p . *Applied Statistics*, **45**, 15-29.
- Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. Ser. B* **61**, 529-546.
- White, H. (1994). *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge, England.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.
- Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, **87**, 732-737.
- Zhang, P. (1993). On the convergence rate of model selection criteria. *Communications in Statistics—Theory and Method*, **22**, 2765-2775.
- Zheng, X. and Loh, W. -Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association*, **90**, 151-156.

APPENDIX: PROOFS OF THE THEOREMS

Theorems 1 and 2 are straightforward corollaries of the following lemmas, respectively, which allows the growth function (h_n) being more general than the BIC choice $\log n$.

Lemma 1 Suppose $M_0 \subset \Omega$, and Φ_Z is a Z -scope constructed on Ω . We have: $P[M_0 \in \Phi_Z] \rightarrow 1$ as $n \rightarrow \infty$.

Lemma 2 (*Consistency*). If (A) $M_0 \in \Phi$ (a finite scope) and if (B) $1 \prec h_n \prec n$ [this means $1/h_n = o(1)$ and $h_n/n = o(1)$ as n increases], then, with probability tending to one as n increases, the true model becomes the unique minimizer of the cost; i.e., $P[\hat{M} \text{ is unique and equal to } M_0] \rightarrow 1$ as $n \rightarrow \infty$.

Lemma 3 (*Consistency with the Z-scope*). Suppose $M_0 \subset \Omega$. Let $\hat{M}_Z = \arg \min_{M \in \Phi_Z} C(M)$, where Φ_Z is a Z-scope constructed on Ω . If the growth function in $C(M)$ satisfies (B) $1 \prec h_n \prec n$, then $P[\hat{M}_Z \text{ is unique and equal to } M_0] \rightarrow 1$ as $n \rightarrow \infty$.

Remark 1 As can be seen from the proof of Lemma 2, Condition (A) can be replaced by a weaker one “ $\lim_{n \rightarrow \infty} P[M_0 \in \Phi] = 1$ ”, where Φ can be a sequence of random scopes.

Remark 2 (*Choice of the Growth Function*). Note that the choice $h_n = \log n$ satisfies Condition (B) on the growth function. While there are other choices (e.g., $h_n = \log \log n$, or \sqrt{n}) that can satisfy the condition and guarantee consistency, the choice $h_n = \log n$ corresponds to an equivalent of the BIC when a likelihood function exists and $\hat{\beta}$ is the MLE (maximum likelihood estimator). Even when $\hat{\beta}$ is not the MLE, such a choice of the growth function leads to a “partial” BIC in the sense of Section 3 and enables a Bayesian-type model evaluation.

Remark 3 (*Strong Consistency*). The theorems above are focusing on consistency “in probability” only. To obtain results on almost sure consistency, the behavior of the lim sup (roughly speaking) of the Wald statistic as n increases is required. If, for example, a law of iterated logarithm holds, then the almost sure consistency is implied by a growth function slower than n but faster than $\log \log n$.

Remark 4 (*Convergence Rate*). It can be shown that the probability of choosing a wrong model, when either \hat{M} or \hat{M}_Z in the previous lemmas is used, is of order $O(n^{-1/2}) + O((h_n e^{h_n})^{-1/2})$, which agrees with the result of Zhang (1993) obtained in the special case of linear regression. The term $O((h_n e^{h_n})^{-1/2})$ comes from the probability of large deviation of the chi-square random variables. The $O(n^{-1/2})$ term is the error of the chi-square approximation obtained from the technique of Edgeworth expansion. Therefore for our BIC-type procedure with $h_n = \log n$, the probability of model misselection is $O(n^{-1/2})$.

Proof of Lemma 2:

Suppose the true model is M_0 of size d_0 . Consider a candidate model M (of size d_M) from Φ .

Case 1. Suppose M misses some label(s) from M_0 . That is, M fails to include some j in M_0 , for which $\beta_j \neq 0$. Then $L_M \beta \neq \mathbf{0}$ (it has a nonzero component β_j). Then, as $n \rightarrow \infty$, $C(M) = n(c + o_p(1))$ where $c = (L_M \beta)^T \{L_M V(\beta) L_M^T\}^{-1} (L_M \beta) > 0$, noting that $h_n \prec n$. This leads to a positive cost of order n [coming from the Wald statistic $W(M)$].

Case 2. Suppose M includes all labels from M_0 (and therefore $d_M > d_0$ unless $M = M_0$). Then $L_M \beta = \mathbf{0}$, and the Wald statistic $W(M)$ is of order $O_p(1)$. The leading contribution to the cost is from the complexity penalty (noting $1 \prec h_n$). Therefore $C(M) \sim d_M h_n$ (where $d_M > d_0$ unless $M = M_0$).

The true model M_0 is one of these candidate models in Φ [Condition (A)], and obviously belongs to the second case. Therefore, $C(M_0) \sim d_0 h_n$. Comparing this with the orders of $C(M)$ in cases 1 and 2, and noting that $h_n \prec n$, we get $C(M_0) < C(M)$ for all M in Φ not equal to M_0 , with probability tending to 1 as $n \rightarrow \infty$. Therefore, with probability tending to one as n increases, M_0 becomes the unique minimizer of $C(M)$. \square

Proof of Lemma 1:

Suppose the true model is $M_0 = \{j_1^0, \dots, j_{d_0}^0\}$ of size d_0 , which is a subset of Ω . Then the Z statistic $Z_k = \beta_k / \sqrt{n^{-1} \{V(\beta)\}_{kk} + \sqrt{n} o_p(1)}$, for each k in Ω . Therefore, $|Z_k|$ is (positive and) of order \sqrt{n} if $k \in M_0$, and of order $\sqrt{n} o_p(1)$ if $k \in \Omega - M_0$. Therefore, with probability tending to one as $n \rightarrow \infty$, the subscripts of the d_0 largest $|Z_k|$'s give all labels in M_0 , i.e., $P[M_{(d_0)} = M_0] \rightarrow 1$. This implies $P[M_0 \in \Phi_Z] \rightarrow 1$, since $M_{(d_0)}$ is a member of Φ_Z . \square

Proof of Lemma 3:

Denote A as the event that there is a minimizer \hat{M}_Z (of $C(M)$ over Φ_Z) which is not equal to M_0 . Then, $P[A] \leq P[M_0 \notin \Phi_Z] + P[M_0 \in \Phi_Z, A] \leq P[M_0 \notin \Phi_Z] + P[A | M_0 \in \Phi_Z]$. As $n \rightarrow \infty$, the first term converges to zero due to Lemma 1. The second term converges to zero due to Lemma 2, since, by the conditioning on $[M_0 \in \Phi_Z]$, Condition (A) is also satisfied (for $\Phi = \Phi_Z$).

Hence $P[A] \rightarrow 0$ as $n \rightarrow \infty$, leading to the proof. \square

Proof of Proposition 2:

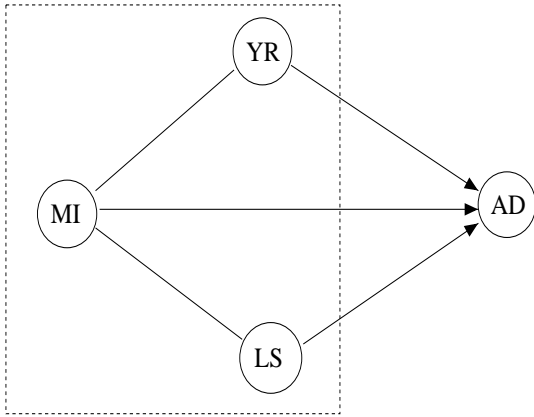
The martingale condition implies that the asymptotic covariance matrix of $\{\mathbf{G}_a\}_1^A$ is block diagonal and so is the asymptotic covariance matrix of $\{\hat{\beta}_a\}_1^A$. Then the quadratic form in the Wald statistic and the cost function decomposes into the corresponding A parts, leading to the proof of the proposition. \square

Table 1: Epileptic seizure data: Cost function $C(M)$ and approximate posterior ratio (PP-Ratio) for four candidate models relative to the best choice

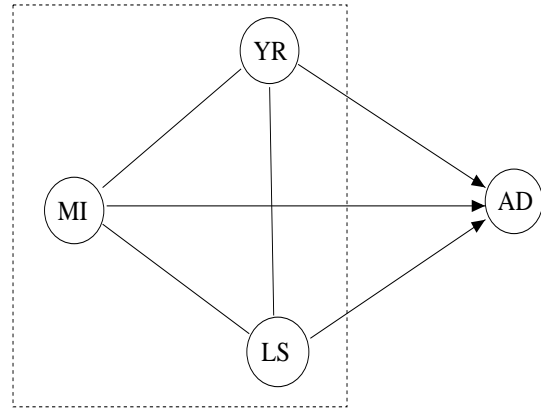
M	d_M	$C(M)$	PP-Ratio
M_0	4	12.449	-
M_1	6	20.544	0.017
M_2	8	28.685	0.000
M_3	10	36.698	0.000

Table 2: Lung cancer data (without prior therapy): Cost function $C(M)$ and approximate posterior ratio (PP-Ratio) for candidate models in the Z-scope relative to the best choice

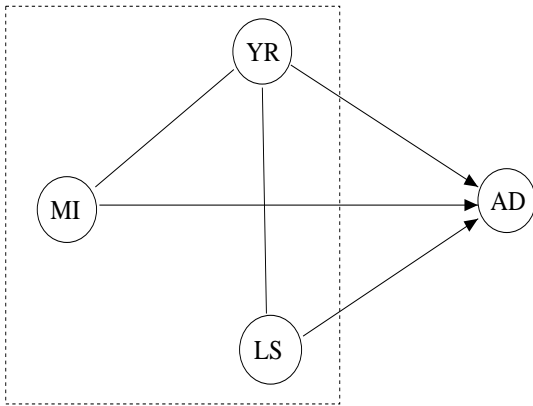
Covariates	d_M	$C(M)$	PP-Ratio
1	1	32.230	0.000
13	2	30.276	0.008
134	3	25.668	0.076
1345	4	20.508	-
12345	5	24.358	0.146
123456	6	27.476	0.031
1234567	7	32.023	0.003



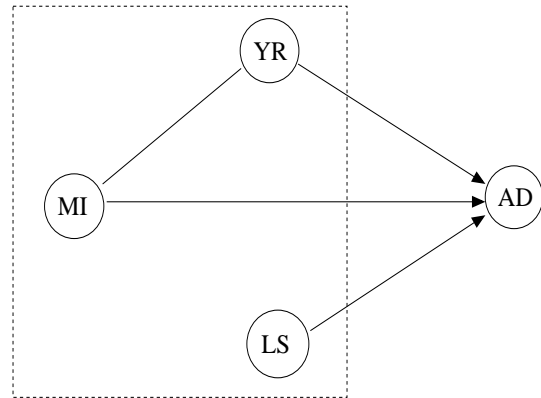
Model A: has the lowest cost 56.0



Model B (cost=57.9): 39% as likely as Model A



Model C (cost=59.5): 17% as likely as Model A



Model D (cost=65.4): 0.9% as likely as Model A

Figure 1: Used automobile prices data: Some chain graph models with small costs