Inferential Models for Linear Regression

Zuoyi Zhang
Department of Statistics
Purdue University
zyzhang@stat.purdue.edu

Huiping Xu
Division of Biostatistics
Indiana University School of Medicine
huipxu@iupui.edu

Ryan Martin
Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago
rgmartin@math.uic.edu

Chuanhai Liu
Department of Statistics
Purdue University
chuanhai@stat.purdue.edu

Abstract

Linear regression is arguably one of the most widely used statistical methods. However, important problems, especially variable selection, remain a challenge for classical modes of inference. This paper develops a recently proposed framework of *inferential models* (IMs) in the linear regression context. In general, the IM framework is able to produce meaningful probabilistic summaries of the statistical evidence for and against assertions about the unknown parameter of interest, and these summaries are shown to be properly calibrated in a frequentist sense. Here we demonstrate by example that the IM framework is promising for linear regression analysis---including model checking, variable selection, and prediction---and for uncertain inference in general.

Keywords and phrases: Auxiliary variable; Credibility; Prediction; Predictive random set; Variable selection.

1. Introduction

Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{U}, \tag{1}$$

where $\mathbf{Y} \in \mathsf{R}^n$ is the vector of response variables, $\mathbf{X}$ is a fixed $n \times p$ matrix of predictor variables, $\boldsymbol{\beta} \in \mathsf{R}^p$ is an unknown vector of regression coefficients, $\sigma > 0$ is an unknown scale parameter, and $\mathbf{U} \in \mathsf{R}^n$ is an unobservable vector of noise, assumed to follow a standard n-dimensional Gaussian distribution, i.e., $\mathbf{U} \sim N_n(0, \mathbf{I}_n)$. Throughout we shall use upper-case bold font ($\mathbf{X}$) for deterministic matrices, upper-case bold italics font ($\mathbf{Y}$) for random vectors, and lower-case bold font ($\mathbf{y}$) for deterministic vectors. In addition to Gaussianity, we shall

assume that $\mathbf{X}$ is of full rank, with $p < n$. In this paper we focus on probabilistic inference in several problems that frequently arise in linear regression analysis, such as model checking, variable selection, and prediction. More precisely, we develop a framework that produces probabilistic summaries of the statistical evidence for and against assertions, or hypotheses, of interest which frequently arise in these aspects of linear regression analysis.

By now, inference in the basic regression problem is well-understood from both frequentist and Bayesian perspectives. However, for the variable selection problem, a fully satisfactory theory/method has yet to emerge. It is not our goal to review the extensive literature on variable selection, but it can be insightful to see where the fundamental difficulty arises. The most popular strategies are stepwise selection procedures and the *lasso* (Tibshirani 1996) and its many variants; see Hastie et al. (2009) for a thorough review of these strategies. These methods have a common drawback, which is that they cannot assign any meaningful measures of uncertainty---probabilistic or otherwise---to the set of variables selected. From a Bayesian perspective, probabilistic summaries of various models can be obtained by introducing a prior probability over the model space and a conditional prior on the model parameters, and performing a Markov chain Monte Carlo scan of the model space. For relatively small p this scheme is feasible (e.g., Clyde and George 2004), but it typically requires a convenient choice of prior for parameters given the model, which may overly influence the posterior calculations. Furthermore, as p increases, estimates of posterior model probabilities become less reliable heaton.scott.2009, making it questionable whether the ``mostly likely'' model has been identified. Since there seems to be no fully satisfactory approach among the existing methods, it makes sense to consider something new and different.

This paper provides an alternative approach to linear regression analysis based on the relatively new *inferential model* (IM) framework; see Martin and Liu (2011) and also Martin et al. (2010) and Zhang and Liu (2011). The crux of the IM approach is its direct attack on the underlying source of uncertainty. In the linear model (1), the source of uncertainty—the unobserved error term $\mathbf{U}$— is clearly specified. This unobservable quantity $\mathbf{U}$, called the *auxiliary variable*, or *a-variable* for short, plays a fundamental role in the IM framework. In particular, prior-free posterior-probabilistic inference about assertions related to variable selection and model checking can be realized by predicting the unobserved value $\mathbf{u}^*$ of $\mathbf{U}$.

The remainder of the paper is organized as follows. In Section 2 we review the general IM approach with a simple illustrative example. Section 3 presents a simplification of the basic IM for linear regression, based on a concept of conditioning. An IM-based approach for model checking is developed in Section 4 and applied to a real-data example taken from Efron et al. (2004). Section 5 presents several new IM-based strategies for variable selection with some numerical results for real- and simulated-data examples. These results indicate that the IM approach described in this paper can select the correct model much

more frequently than popular existing methods based on the LARS algorithm of Efron et al. (2004). Inference on future observations from an IM perspective is the subject of Section 6, and some concluding remarks are given in Section 7.

2. Review of inferential models

As mentioned in Section 1, the unobserved auxiliary variable plays a fundamental role in the IM framework. This section makes this statement more precise and gives a simple and general three-step construction of IMs and a review of some important properties. More details can be found in Martin and Liu (2011).

2.1 Association models

The starting point of an IM analysis is a relationship between data, unknown parameters, and the unobserved a-variables. In general, we write this as

$$\mathbf{Y} = a(\boldsymbol{\theta}, \mathbf{U}), \tag{2.1}$$

where $\mathbf{Y} \in \mathsf{Y}$ is the observable data, $\boldsymbol{\theta} \in \Theta$ is the unknown parameter, and $\mathbf{U} \in \mathsf{U}$ is the unobservable a-variable. In addition to (2.1), a distribution μ for $\mathbf{U}$ is required, which we call the a-measure. The idea is that if $\mathbf{U} \sim \mu$, then the induced distribution on $\mathbf{Y}$ via (2) matches the specified sampling distribution. The quintuple $(\mathsf{Y}, \Theta, \mathsf{U}, a, \mu)$ defines what we call an *association model*. In the regression case, the association model is characterized by the relation (1), with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, and the a-measure $\mu = N_n(0, \mathbf{I}_n)$.

Observe that if we only knew the actual value $\mathbf{u}^*$ for a given data set, then we would know all that one could ever know about $\boldsymbol{\theta}$ from observing $\mathbf{Y} = \mathbf{y}$. For this reason, we shift focus from the unknown parameter to the unobserved value of the a-variable. There are also some philosophical reasons for focusing on $\mathbf{u}^*$ rather than $\boldsymbol{\theta}$. In fact, Martin and Liu (2011) argue that predictive probabilistic inference about $\boldsymbol{\theta}$ is not possible unless there is some unobserved but predictable quantity associated with $\mathbf{y}$ and $\boldsymbol{\theta}$. So, in a certain sense, shifting focus to the a-variable is the only way to accomplish our goal of prior-free probabilistic inference.

2.2 A three-step construction of IMs

Once an association model (2.1) is specified, the construction of a corresponding IM is fairly straightforward. This section outlines a simple three-step procedure.

A-step *The association step begins with the association model (2.1) and, for given* $\mathbf{Y} = \mathbf{y}$*, defines a mapping from* U *to subsets of* Θ *as follows:*

$$\Theta_{\mathbf{y}}(\mathbf{u}) = \{\boldsymbol{\theta} \in \Theta : \mathbf{y} = a(\boldsymbol{\theta}, \mathbf{u})\}, \quad \mathbf{u} \in \mathsf{U}. \tag{2.2}$$

Intuitively, $\Theta_{\mathbf{y}}(\mathbf{u})$ corresponds to the set of candidate $\boldsymbol{\theta}$ values which corresponds to the observed $\mathbf{y}$, the particular $\mathbf{u}$, and the association model (2.1). Note that the true value of $\boldsymbol{\theta}$ must be contained in $\Theta_{\mathbf{y}}(\mathbf{u}^{\text{å}})$.

P-step *The prediction step starts with an assessment of what is known about* $\mathbf{u}^{\text{å}}$. *In particular, it is known that* μ^* *is a sample from* μ. *But trying to (accurately) predict* μ^* *with another draw* $\mathbf{U} \sim \mu$ *is a hopeless endeavor. We acknowledge this difficulty and choose, instead, to try to predict* μ^* *with a random set. Let* S *be a mapping from* U *to subsets of* U *that satisfies* $\mathsf{S}_{\mathbf{u}} \ni \mathbf{u}$ *for all* $\mathbf{u}$. *That is, for the P-step we produce a sample* $\mathbf{U} \sim \mu$ *and construct the predictive random set (PRS)* $\mathsf{S}_{\mathbf{U}}$ *to predict the unobserved* μ^*. *To ensure that the resulting IM has desirable properties, some conditions on the mapping* S *must be imposed; see Section 2.3.*

C-step *The combination step puts together the results of the A- and P-steps above. That is, incorporating the additional uncertainty about* μ^* *in* $\mathsf{S}_{\mathbf{u}}$ *into the set* $\Theta_{\mathbf{y}}(\cdot)$ *gives the expanded set of candidate* θ *values:*

$$\Theta_{\mathbf{y}}(\mathsf{S}_{\mathbf{u}}) = \bigcup_{\mathbf{u}' \in \mathsf{S}_{\mathbf{u}}} \Theta_{\mathbf{y}}(\mathbf{u}'). \tag{2.3}$$

When $\mathbf{U} \sim \mu$, the set $\Theta_{\mathbf{y}}(\mathsf{S}_{\mathbf{U}})$ is random and the IM output corresponds to certain probabilities for this random set. Let $A \subseteq \Theta$ be an assertion about the parameter $\boldsymbol{\theta}$. This A plays a role similar to a (null) hypothesis in the classical framework. We compute the probability that $\Theta_{\mathbf{y}}(\mathsf{S}_{\mathbf{U}})$ is a subset of A (resp. A^c) as a measure of the statistical evidence for (resp. against) the assertion A. In particular, for given A we compute

$$\text{Bel}_{\mathbf{y},\mathsf{S}}(A) = \mu\{\mathbf{u} : \Theta_{\mathbf{y}}(\mathsf{S}_{\mathbf{u}}) \subseteq A\}, \tag{2.4}$$

the belief function, and

$$\text{Pl}_{\mathbf{y},\mathsf{S}}(A) = 1 - \text{Bel}_{\mathbf{y},\mathsf{S}}(A^c), \tag{2.5}$$

the plausibility function. The functions are similar to those that appear in the Dempster–Shafer theory (Dempster 2008, Shafer 1976), i.e., $\text{Bel}_{\mathbf{y},\mathsf{S}}(A) \le \text{Pl}_{\mathbf{y},\mathsf{S}}(A)$ for all A, but our unique use of the PRS S has some important consequences. Together, $(\text{Bel}, \text{Pl})_{\mathbf{y},\mathsf{S}}$ characterize the IM output.

For illustration, consider a simple special case of (1), namely, $Y = \theta + U$, where $U \sim N(0,1)$. For the A-step, we have a simple mapping $\Theta_y(u) = \{y - u\}$, a singleton. For the P-step, consider the PRS defined by

$$\mathsf{S}_u = \{u' : |u'| \le |u|\}, \quad u \in \mathsf{R}. \tag{2.6}$$

Then the result of the C-step is the set $\Theta_y(\mathsf{S}_u) = [y - |u|, y + |u|]$. Take, for example, a single point assertion $A = \{\theta_0\}$. Then it is easy to see that the belief function for A is zero. But the plausibility function is

$$\begin{aligned}
&\mathrm{Pl}_{y,\mathsf{S}}(\{\theta_0\}) = 1 - \mu\{u : \Theta_y(\mathsf{S}_u) \subseteq \{\theta_0\}^c\} \\
&= 1 - \mu\{u : y - |u| > \theta_0 \text{ or } y + |u| < \theta_0\} \\
&= 1 - \mu\{u : |u| < |y - \theta_0|\} \\
&= 2\{1 - \Phi(|y - \theta_0|)\},
\end{aligned}$$

where Φ denotes the distribution function of $N(0,1)$. The plausibility function for singleton assertions is an important quantity in an IM analysis. As Martin and Liu (2011) show, this function can be used to construct a so-called *plausibility interval* for θ, an IM-based counterpart to the frequentist confidence or the Bayesian credible interval. Figure 1 shows a graph of $\mathrm{Pl}_{y,\mathsf{S}}(\{\theta_0\})$ as a function of θ_0 for $y = 3$. The message is that only values of θ relatively close to y are plausible. Note that the $100(1-\alpha)\%$ plausibility interval defined in Martin and Liu (2011) is just the α -level set of $\mathrm{Pl}_{y,\mathsf{S}}(\{\theta_0\})$.

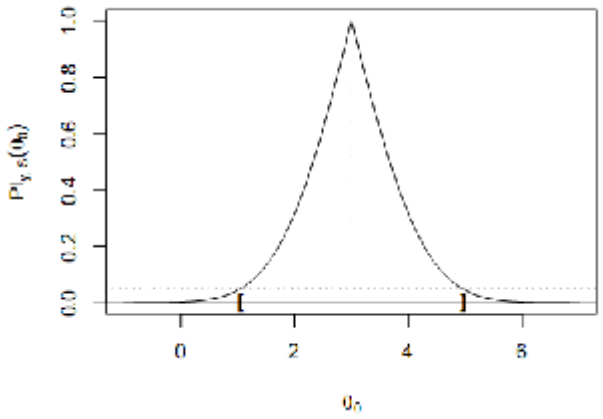

Figure 1: A plot of $\mathrm{Pl}_{y,\mathsf{S}}(\{\theta_0\})$ as a function of θ_0 when $y = 3$. The dotted horizontal line at $\alpha = 0.05$ defines the cutoff points for the 95% plausibility interval for θ marked by brackets on the horizontal axis.

2.3 Properties

As mentioned above in the P-step, the PRS influences the properties of the resulting IM. In fact, properties of the IM are almost completely determined by properties of the PRS. Towards this, define $Q_{\mathsf{S}}(\mathbf{u}) = \mu\{\mathbf{u}' : \mathsf{S}_{\mathbf{u}'} \not\ni \mathbf{u}\}$, for $\mathbf{u} \in \mathsf{U}$. Then the PRS $\mathsf{S}_{\mathbf{u}}$ is said to be *credible* at level $\alpha \in (0,1)$ if $\mu\{\mathbf{u} : Q_{\mathsf{S}}(\mathbf{u}) \geq 1 - \alpha\} \leq \alpha$. In words, credibility implies that the probability that S_{U} misses its target $\mathbf{u}$ is large only for a relatively small proportion of possible $\mathbf{u}$ values. Theorem 1 of Martin and Liu (2011) can be used to show that the PRS in (2.6) and the others used in Section 5.3 are credible.

The property of interest for IMs is what is called validity and is essentially a long-run frequency calibration property of the belief and plausibility function values. This property is what makes the numerical values of these functions meaningful across different users or experiments. Let $\mathrm{P}_{\boldsymbol{\theta}}$ denote the sampling model for the data $\mathbf{Y}$. Then the IM is *valid for* A if, for each $\alpha \in (0,1)$, the belief function satisfies

$$\sup_{\boldsymbol{\theta} \notin A} \mathrm{P}_{\boldsymbol{\theta}}\{\mathrm{Bel}_{\mathbf{Y},\mathrm{S}}(A) \geq 1-\alpha\} \leq \alpha \tag{2.7}$$

and the plausibility function satisfies

$$\sup_{\boldsymbol{\theta} \in A} \mathrm{P}_{\boldsymbol{\theta}}\{\mathrm{Pl}_{\mathbf{Y},\mathrm{S}}(A) \leq \alpha\} \leq \alpha. \tag{2.8}$$

The IM is *valid* if it is valid for all A. Theorem 2 of Martin and Liu (2011) shows that if the PRS is credible, then the resulting IM is valid.

An important application of this validity result is the case where the statistical analysis must result in an ``accept/reject'' decision. In this case, a natural strategy is to pick a small $\alpha \in (0,1)$ and conclude that an assertion A is *true* if $\mathrm{Bel}_{\mathrm{y},\mathrm{S}}(A) \geq 1-\alpha$ or *false* if $\mathrm{Pl}_{\mathrm{y},\mathrm{S}}(A) \leq \alpha$. The validity property then guarantees that such a procedure controls the ``Type I error'' probabilities at the α level. This approach, with $\alpha = 0.05$, will be used in the numerical examples that follow.

As in the classical hypothesis testing problem where Type I error probability cannot be the only consideration, in the IM context we cannot focus solely on the validity property. We say that an IM is *efficient for* A if the inequalities ``$\leq \alpha$'' in (2.7) and (2.8) are both equalities. This additional property can be achieved, at least for some assertions, such as singletons, via stronger conditions on the PRS. A powerful technique for obtaining efficient IMs is via an initial dimension reduction step of the a-variable $\mathbf{U}$. Two such procedures are available–*conditioning* and *marginalization*–and these techniques will be used occasionally in what follows. The interested reader may refer to Martin et al. (2011a,b) for the details.

Finally, despite the frequency-calibration properties described above, the main thrust of the IM approach is that the inferential output can be interpreted probabilistically. That is, for a given data set $\mathbf{Y} = \mathbf{y}$, the quantities $\mathrm{Bel}_{\mathrm{y},\mathrm{S}}(A)$ and $\mathrm{Pl}_{\mathrm{y},\mathrm{S}}(A)$ are not simply tools to construct frequentist decision procedures. In fact, $\mathrm{Bel}_{\mathrm{y},\mathrm{S}}(A)$, for example, has a meaningful interpretation as the amount of statistical evidence available in $\mathbf{Y} = \mathbf{y}$ supporting the assertion A about $\boldsymbol{\theta}$–no notion of ``repeated experiments'' is needed for interpretation. Neyman–Pearson type of decision rules lack this property, and Fisher's p-value gives only indirect evidence for/against the assertion A, viewed as a null hypothesis.

3. IMs for linear regression

The regression model (1.1) clearly demonstrates the association between the observable $\mathbf{Y}$, the unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, and the unobservable but predictable $\mathbf{U}$. This results in a system of n equations in $p+1$ unknowns, the association model will have some functions of the a-variable which are actually observed. To see this, let $\mathbf{P}$ and $\mathbf{P}^{\perp}$ denote the familiar projection matrices onto the linear space $C_{\mathbf{X}}$ spanned by the columns of $\mathbf{X}$ and the linear space $C_{\mathbf{X}}^{\perp}$ orthogonal to $C_{\mathbf{X}}$, respectively. If $\mathsf{P}\cdot\mathsf{P}$ denotes the usual ℓ_2-norm, then the association model (1) can be re-written as

$$\mathbf{PY} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{PU}, \tag{3.1}$$

$$\left\|\mathbf{P}^{\perp}\mathbf{Y}\right\| = \sigma\left\|\mathbf{P}^{\perp}\mathbf{U}\right\|, \tag{3.2}$$

and

$$\left\|\mathbf{P}^{\perp}\mathbf{Y}\right\|^{-1}\mathbf{P}^{\perp}\mathbf{Y} = \left\|\mathbf{P}^{\perp}\mathbf{U}\right\|^{-1}\mathbf{P}^{\perp}\mathbf{U}. \tag{3.3}$$

As the direction of the residual vector obtained from the least-square fitting of the linear regression model to the observed data, the function $\left\|\mathbf{P}^{\perp}\mathbf{U}\right\|^{-1}\mathbf{P}^{\perp}\mathbf{U}$ of $\mathbf{U}$ in (3.3) is fully observed. Thus, predicting the n-vector $\mathbf{U}$ amounts to predicting $(\mathbf{PU}, \left\|\mathbf{P}^{\perp}\mathbf{U}\right\|)$, as functions of $\mathbf{U}$ involved in (3.1) and (3.2), conditioned on (3.3). This approach effectively reduces the dimension of a-variables to be predicted and, upon following the three steps in Section 2.2, produces what Martin et al. (2011) call a conditional IM.

Well-known sampling distribution properties in the linear model context suggest the following transformations. Define

$$S = \left\|\mathbf{P}^{\perp}\mathbf{Y}\right\| / \sqrt{n-p} \quad \text{and} \quad \mathbf{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

to be the usual least-squares estimates of σ and $\boldsymbol{\beta}$, respectively. Set

$$M = \left\|\mathbf{P}^{\perp}\mathbf{U}\right\| / \sqrt{n-p} \quad \text{and} \quad \mathbf{T} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{U} / M.$$

It is known that M and $M\mathbf{T}$ are independent, with $(n-p)M^2 \sim ChiSq(\nu)$ and $\mathbf{T} \sim t_p(0, \mathbf{W}, \nu)$, where $ChiSq$ and t_p denote the chi-square and a p-variate Student-t distributions, respectively, the scale matrix satisfies $\mathbf{W} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, and in both cases the degrees of freedom ν equals $n-p$. Moreover, M and $\mathbf{T}$ are independent of $\left\|\mathbf{P}^{\perp}\mathbf{U}\right\|^{-1}\mathbf{P}^{\perp}\mathbf{U}$. According to Martin et al. (2011a), (3.1) and (3.2) can be equivalently written as

$$\mathbf{B} = \boldsymbol{\beta} + S\mathbf{T} \tag{3.4}$$

and

$$S = \sigma M, \tag{3.5}$$

which gives an association model for inference about $(\boldsymbol{\beta}, \sigma^2)$ with observables S and $\mathbf{B}$ and a-variables M and $\mathbf{T}$. We note that the reduction of the a-variable dimension here, via conditioning, is equivalent to an initial reduction via sufficiency. But, in this IM context, it is not the reduction of the data to sufficient statistics that is important: the advantage of this conditional IM is that it is easier to efficiently predict the new a-variable $(\mathbf{T}, M) \in \mathsf{R}^{p+1}$ compared to $\mathbf{U} \in \mathsf{R}^n$, when $n > p+1$.

It is typically the case that σ is a nuisance parameter–an unknown quantity but not of primary interest. In such cases, it is desirable to further reduce the dimension of the a-variable $(M, \mathbf{T})$ via marginalization. The details behind the IM-based marginalization strategy, found in Martin et al. (2011b), are quite deep, but the result is that we effectively ignore (3.5) and focus exclusively on the *marginal* association model (3.4). The corresponding marginal IM is defined through the three-step process based on trying to predict the unobserved value $\mathbf{t}^*$ of $\mathbf{T}$. One can similarly marginalize out certain components of $\boldsymbol{\beta}$ if only a subset are of interest. For example, the intercept is often of no real interest so it can be marginalized away; we do this in Section 5.1.

To summarize, direct reasoning with the underlying source of uncertainty leads to association models which are effectively the same as those that would be obtained via the familiar classical arguments, e.g., sufficiency. The advantage of the IM approach will present itself when we discuss challenging inference problems such as variable selection. The prior-free probabilistic inference one can achieve with IMs is something new that cannot be obtained by classical arguments.

4. Model checking

4.1 An IM-based approach

The conditional IM, with association model given by (3.4) and (3.5) has a p-dimensional a-variable to predict. For (computational) simplicity we shall assume $n - p$ is large, so that the standardized residual process

$$\mathbf{r} = s^{-1}\mathbf{P}^{\perp}\mathbf{y} \tag{4.1}$$

can be considered as a sample from $N(0,1)$. In this case, we view $\mathbf{r}$ as a stochastic process along a selected direction in $C_{\mathbf{X}}$. In particular, we are concerned with potential deviations from the linearity assumption, so we consider directions of individual explanatory variables and directions determined by any pair of these variables. The basic idea is to see if the regression of $\mathbf{Y}$ on $\mathbf{X}$, namely $\mathrm{E}(\mathbf{Y} \mid \mathbf{X})$, can be approximated significantly better, compared to the linear model, by including higher-order (e.g., quadratic) terms. To be more specific, for

any two predictor variables $\mathbf{x}_i$ and $\mathbf{x}_j$ (columns of $\mathbf{X}$) we consider four directions: $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_i + \mathbf{x}_j$, and $\mathbf{x}_i - \mathbf{x}_j$. There is a total of p^2 directions to be considered and, in what follows, we let $\mathbf{x}$ be a generic notation for one such direction.

Let $\mathbf{r}(\mathbf{x})$ be the residual process in (4.1) but ordered according to the magnitudes of the coordinates of $\mathbf{x}$. Write $h_i = -\log \Phi(\mathbf{r}(\mathbf{x})_i)$ for $i = 1, \ldots, n$. If $\mathbf{r}$ does not strongly depend on $\mathbf{x}$ in the sense that a plot of $\mathbf{r}(\mathbf{x})_i$ versus the index i shows no discernible pattern, then $h_1, \ldots, h_n$ should look like an independent sample of unit exponential random variables, a simple exponential process along the $\mathbf{x}$ direction. Then it follows that

$$u_{(i)} = \frac{\sum_{j=1}^{i} h_j}{\sum_{j=1}^{n} h_j}, \quad i = 1, \ldots, n-1$$

should resemble an ordered sample of $n-1$ uniform random variables. Therefore, evidence for/against the computed $u_{(1)}, \ldots, u_{(n-1)}$ looking like an ordered sample of uniforms can be used as evidence for/against proper fit of the postulated model. An efficient approach for predicting a set of ordered uniform variates has been recently developed in Zhang (2010); see also Zhang and Liu (2011), Martin et al. (2010) and Martin and Liu (2011, Sec. 6.2). With this efficient PRS in place, the three-step construction of the IM as well as the belief and plausibility function computation are fairly straightforward.

4.2 The diabetes data example

Consider the diabetes data example in Efron et al. (2004). This data set consists of observations for $n = 442$ diabetes patients with response variable (a quantitative measurement of disease progression one year after baseline) and ten covariates: age, sex, body mass index (bmi), average blood pressure (map) and six blood serum measurements (tc, ldl, hdl, tch, ltg, and glu). Our analysis begins with checking if the fit of the full model with ten covariates is satisfactory. Figure 2 shows the analysis of the residual process along the $age + sex$ direction. That is, the residual process $\mathbf{r}$ in (4.1) is sorted according to the values of $\mathbf{x} = age + sex$. The plausibility function for the assertion that the corresponding $u_{(1)}, \ldots, u_{(441)}$ looks like a sample of ordered uniforms is 0.0578, a small number, suggesting the potential existence of an $age \times sex$ interaction. The analysis in Section 5.3 shows that the model that includes this interaction term is satisfactory as a full model from which the variable selection problem can begin.

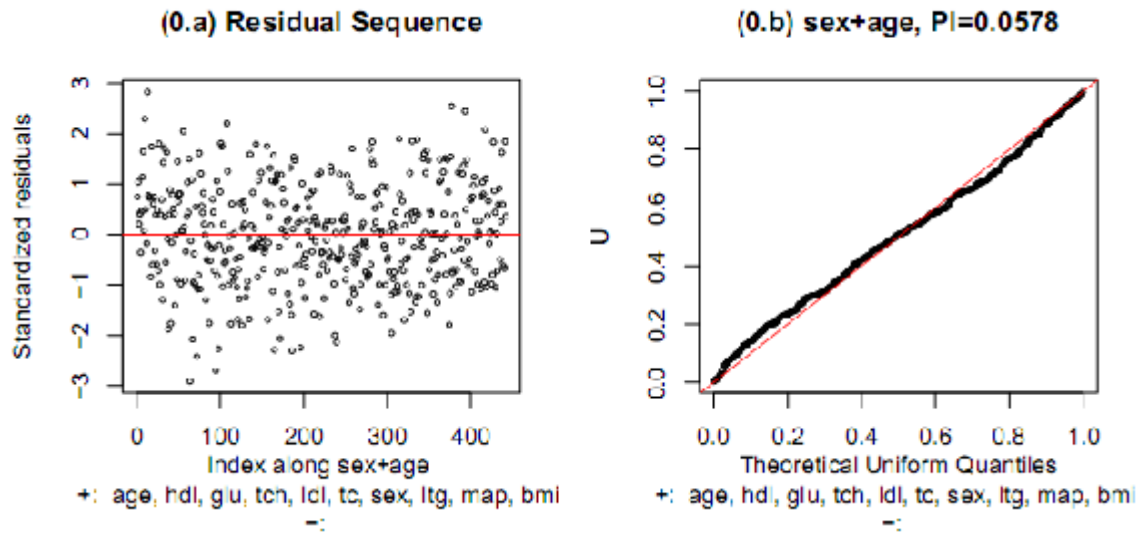

Figure 2: Plots for model checking: (0.a) the sequence of normalized residuals from fitting the full model with the given ten covariates age, hdl, …, and bmi, as a stochastic process, along the direction of $sex+age$, and (0.b) the accumulated exponential process for checking normality and trend along direction of $sex+age$, for which the plausibility function evaluates to 0.0578.

5. Variable selection

5.1 Ordering the covariates

According to the discussion in Section 2, probabilistic inference about $\boldsymbol{\beta}$ in general, or variable selection in particular, should be carried out by predicting at least p a-variables, namely, the quantity $\mathbf{T}$ in the association model (13). A change-of-variables will make the notation more convenient. Recall the matrix $\mathbf{W} = ((w_{ij})) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. We define the following modified notation:

$$Z_i \leftarrow \frac{B_i}{\sqrt{S^2 w_{ii}}}, \quad \theta_i \leftarrow \frac{\beta_i}{\sqrt{S^2 w_{ii}}}, \quad \text{and} \quad w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{w_{ii} w_{jj}}}, \quad i, j = 1, \ldots, p.$$

Then, in terms of the new notation, the association model is

$$\mathbf{Z} = \quad + \quad, \quad \text{with} \quad \sim t_p(0, \quad, \nu), \tag{5.1}$$

almost exactly as before, except for now all quantities involved have been properly scaled. Observe that $\beta_i = 0$ if and only if $\theta_i = 0$. Having a common scale for the U_i's will come in handy when we specify PRS in (5.3) below.

In what follows we shall assume that the p equations in (5.1) have been sorted according to the magnitude of coordinates in $\mathbf{Z}$, i.e.,

$$|Z_1| \leq |Z_2| \leq \cdots \leq |Z_p|. \tag{5.2}$$

We shall also assume that the model always includes an intercept, and that this `` β_0 '' term has been marginalized out as described at the end of Section 3. So, henceforth, p shall stand for the number of variables in the model, not counting the intercept. In the diabetes example to follow, we have the original ten variables plus an interaction term, thus making $p = 11$. The degrees of freedom ν in (5.1) will also change accordingly.

A key observation is that, after the rescaling, the components of $\mathbf{U}$ in (5.1) have identical marginal distributions. Therefore, the set-valued mapping S used to build a PRS for $\mathbf{u}^*$ should satisfy the following property:

$$\text{proj}_i(\mathsf{S}_{\mathbf{U}}) \equiv \text{proj}_j(\mathsf{S}_{\mathbf{U}}), \quad i, j = 1, \ldots, p. \tag{5.3}$$

where $\text{proj}_i(\mathsf{S}_{\mathbf{u}})$ denotes the projection of the set $\mathsf{S}_{\mathbf{u}} \subset \mathsf{R}^p$ down to the u_i-space. In this case, the evidence against $\{\theta_i = 0\}$ can be no more than the evidence against $\{\theta_j = 0\}$ whenever $i < j$ and, hence, $|Z_i| \leq |Z_j|$. Therefore, the ordering of the rescaled least-squares estimates $\mathbf{Z}$ and the symmetry property (5.3) of the PRS suggest a certain nesting of the candidate models, and so we shall consider the sequence of assertions

$$A_k = \{\boldsymbol{\theta} : \theta_1 = \cdots = \theta_k = 0\}, \quad k = 1, \ldots, p. \tag{5.4}$$

for the variable selection problem in Section 5.3. But first, in Section 5.2, we give a variable selection application of the model checking strategy in Section 4.

5.2 Variable selection via iterative model building

We consider an iterative model building process based on the ordering of the covariates in Section 5.1. First, set $k = 1$; now proceed as follows:

1. Drop the covariates corresponding to $\theta_1, \cdots, \theta_k$;
2. Compute the normalized residuals $\mathbf{r}$ in (4.1) from the least-squares fit;
3. Apply the method of Section 4 to check the residual processes along all p^2 covariate directions, even those first k that were dropped in Step 1.
4. If one of the p^2 goodness-of-fit evidentiary measures falls below $\alpha = 0.05$, then stop deleting variables; otherwise, set $k \to k+1$ and go back to Step 1.

We apply this procedure to the diabetes example described in Section 4.2. For each $k = 1, \ldots, 11$, plausibility is computed for all the 11×11 residual processes. Figure 3 shows the strongest ones for each $k = 0, 1, \ldots, 5$. The $k = 0$ case stands for the full model, i.e., all the variables are selected. The results in panels (0.a) and (0.b) shows that there is not very strong evidence against the assumption that the enlarged full model, with the $age \times sex$ interaction, is adequate as the baseline regression model. For the $k = 4$ case, i.e., after deleting the four variables hdl, glu, tch, and ldl, panels (4.a) and (4.b) show strong evidence against the assertion that the model excluding the these variables fits the data.

This result suggests that except for the first three variables–hdl, glu, and tch–all other variables are significant for explaining the outcome variable of interest.

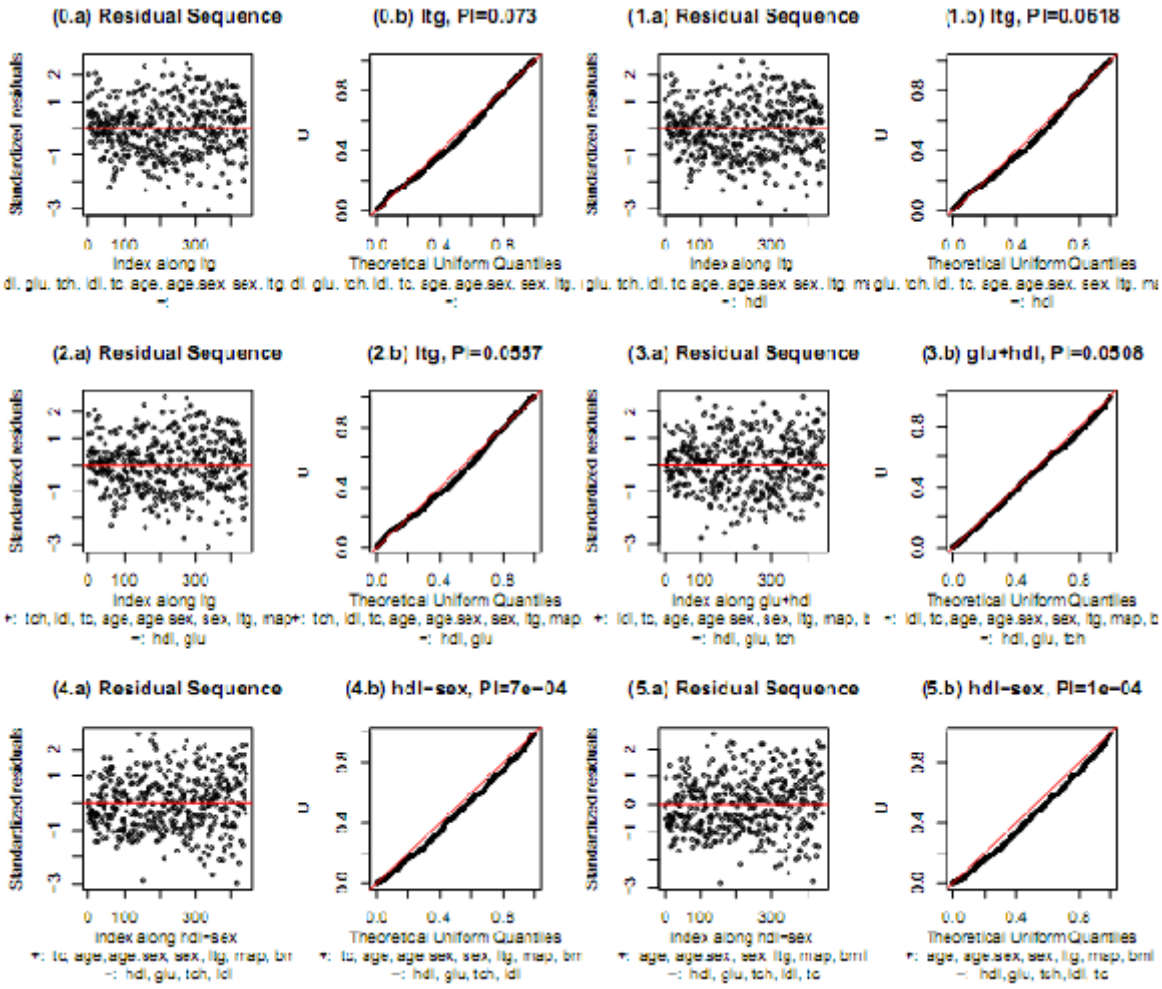

Figure 3: The first 6 sets of plots for model checking. The goodness of fit is measured by evidence against the chosen model and is reported in the (*.b) panels as ``PI.'' Figure subtitles show which variables are included (+) and excluded (−).

Although apparently new, this approach is informal and not unlike existing step-wise variable selection procedures from an operational point of view, although there is a probabilistic interpretation to the evidentiary measure of the importance of each variable. A formal approach is considered next, where the IM is used to produce meaningful summaries of evidence for/against assertions relevant to the variable selection problem.

5.3 IMs for variable selection: a preview

In this section we present a relatively simple IM-based approach for variable selection. Some more sophisticated approaches will be investigated in future work. In the subsections that follow, we present IMs for producing evidence

for/against the null model and certain non-null models, as determined by the class of assertions A_k, $k = 1, \ldots, p$, in (5.4). Since A_k determines a lower-dimensional subspace of R^p, the belief function of A_k must be zero. We shall, therefore, only discuss evaluation of the plausibility function. Numerical results are given in Sections 5.3.4 and 5.3.5.

Note that the sequence of assertions in (5.4) depends on the observed data through the ordering (5.2). The general IM theory, however, does not immediately apply to data-dependent assertions. For simplicity, we shall ignore this dependence in what follows. See duncan.thesis for an IM analysis involving data-dependent assertions.

5.3.1 Plausibility for the null model

To start, we consider determination between the null model–the model with only an intercept–and some non-null model. Towards this, we shall construct an IM and evaluate the plausibility function at $A_p = \{\boldsymbol{\theta} : \theta_1 = \cdots = \theta_p = 0\}$. If this plausibility e is small, we proceed by investigating the covariate corresponding to Z_p in Section 5.3.2; otherwise, we stop and select no variables.

Here we give the three simple steps to construct an IM for inference about A_p. The PRS we use is called a ``box PRS'' and is based on the ℓ_∞-norm, defined by $\|\mathbf{u}\|_\infty = \max_{1 \le i \le p} |u_i|$. Note that for the observed $\mathbf{Z} = \mathbf{z}$, we have $\|\mathbf{z}\|_\infty = |z_p|$.

A-step *Associate the observed* $\mathbf{z}$ *with the unknown* $\boldsymbol{\theta}$ *and unobservable* $\mathbf{u}$ *in (5.1) to obtain the singleton set* $\Theta_{\mathbf{z}}(\mathbf{u}) = \{\mathbf{z} - \mathbf{u}\}$.

P-step *Predict the unobserved* $\mathbf{u}^*$ *with a PRS* $\mathsf{S}_{\mathbf{U}} = \{\mathbf{u} : \|\mathbf{u}\|_\infty \le \|\mathbf{U}\|_\infty\}$, *where* $\mathbf{U}$ *has the re-scaled Student-t distribution in (5.1).*

C-step *Combine* $\Theta_{\mathbf{z}}$ *and* S *to get* $\Theta_{\mathbf{z}}(\mathsf{S}_{\mathbf{u}}) = \{\boldsymbol{\theta} : \|\mathbf{z} - \boldsymbol{\theta}\|_\infty \le \|\mathbf{u}\|_\infty\}$. *Then the plausibility function at* A_p *is given by*

$$\begin{aligned} \mathrm{Pl}_{\mathbf{z},\mathsf{S}}(A_p) &= 1 - \mu\{\mathbf{u} : \Theta_{\mathbf{z}}(\mathsf{S}_{\mathbf{u}}) \subset A_p^c\} \\ &= 1 - \mu\{\mathbf{u} : \|\mathbf{u}\|_\infty < \|\mathbf{z}\|_\infty\} = 1 - F(|z_p|), \end{aligned}$$

where F denotes the distribution function of $\|\mathbf{U}\|_\infty$ when $\mathbf{U} \sim \mu := t_p(0, \mathbf{W}, \nu)$. This probability can be easily approximated using Monte Carlo methods.

5.3.2 Plausibility for non-null models

Suppose, starting with the null model case $k = p$, the plausibility for each $A_p, \ldots, A_{k+1}$ is small. That is, we have already selected the covariates

corresponding to $z_p,\ldots,z_{k+1}$. The goal here is to use the plausibility of A_k for deciding whether to select the variable corresponding to z_k.

It is tempting to consider the marginal IM that marginalizes out $\theta_{k+1},\ldots,\theta_p$. But care must be taken to account for the selective process imposed by ordering the rescaled least-squares estimates $\mathbf{z}$. Without accounting for the imposed constraint, the association model for inference about $\boldsymbol{\theta}_{1:k}=(\theta_1,\ldots,\theta_k)$ would be given by

$$\mathbf{Z}_{1:k} = \quad_{1:k} + \quad_{1:k}, \qquad \quad_{1:k} \sim t_p(0, \quad_{1:k,1:k}, \nu), \tag{5.5}$$

where $\mathbf{W}_{1:k,1:k}$ is the $k\times k$ block in the upper left-hand corner of $\mathbf{W}$. While the actual values of $\theta_{k+1},\ldots,\theta_p$ are unknown, they do have some impact on $\boldsymbol{\theta}_{1:k}$ and, hence, on the distribution of $\mathbf{U}_{1:k}$. Indeed, it is known that

$$\left\|\boldsymbol{\theta}_{1:k}+\mathbf{U}_{1:k}\right\|_\infty \le \ |Z_{k+1}|. \tag{5.6}$$

Therefore, the appropriate association model would be (5.5) but with the a-measure μ taking the constraint (5.6) into account. This discussion leads to the following three-step IM construction.

A-step *Per (5.5), the set of candidate* $\boldsymbol{\theta}_{1:k}$*'s is* $\Theta_{\mathbf{z}}(\mathbf{u}_{1:k})=\{\mathbf{z}_{1:k}-\mathbf{u}_{1:k}\}$.

P-step Predict the unobserved $\mathbf{u}^*_{1:k}$ with the PRS $\mathsf{S}_{\mathbf{U}}$ determined by the Student-t distribution for $\mathbf{U}=\mathbf{U}_{1:k}$ in (5.5) with the constraint (5.6) and the set-valued mapping $\mathsf{S}_{\mathbf{u}}=\{\mathbf{u}_{1:k}': \left\| \quad_{1:k}+\quad'_{1:k}\right\|_\infty \le \left\| \quad_{1:k}+\quad_{1:k}\right\|_\infty\}$, $\mathbf{u}=\mathbf{u}_{1:k}\in\mathsf{R}^k$. It may seem a bit unnatural that the PRS itself depends on the unknown $\boldsymbol{\theta}_{1:k}$. But since this PRS will be combined with the set $\Theta_{\mathbf{z}}$ anyway, this dependence causes no technical problems. Indeed, the credibility and validity results described above go through in this more general case.

C-step *Combining the results of the A- and P-steps gives*

$$\Theta_{\mathbf{z}}(\mathsf{S}_{\mathbf{u}})=\{\boldsymbol{\theta}_{1:k}: \left\|\mathbf{z}_{1:k}\right\|_\infty \le \left\|\boldsymbol{\theta}_{1:k}+\mathbf{u}_{1:k}\right\|_\infty\}, \quad \mathbf{u}=\mathbf{u}_{1:k}\in\mathsf{R}^k.$$

Then the plausibility function at A_k is given by

$$\begin{aligned}\mathrm{Pl}_{\mathbf{z},\mathsf{S}}(A_k) &= 1-\mu\{\mathbf{u}_{1:k}:\Theta_{\mathbf{z}}(\mathsf{S}_{\mathbf{u}})\subset A_k^c\}\\ &= 1-\mu\{\mathbf{u}_{1:k}:\left\|\mathbf{u}_{1:k}\right\|_\infty < |z_k|\} = 1-F(|z_k|),\end{aligned}$$

where F denotes the distribution function of $\left\|\mathbf{U}_{1:k}\right\|_\infty$ when $\mathbf{U}_{1:k}\sim\mu$ and μ is defined as the constrained Student-t distribution from above. Again, this probability can be easily approximated using Monte Carlo methods.

5.3.3 Summary of the method

The previous two subsections have described how one can evaluate the plausibility function at assertions characterizing the null and certain non-null models. Here we describe how these quantities are to be used for variable selection. Most importantly, we want to emphasize that, unlike the informal approach in Section 5.2, this approach is not iterative or step-wise. For the given covariate order, one evaluates the plausibility for each model in the nested sequence. Then to select a set of variables, one will simply pick the smallest model whose plausibility is below a specified α threshold. This is certainly different from the classical forward and backward step-wise variable selection procedures that ignore the accumulation of error probabilities due to multiple decisions. In fact, it is more like a Bayes approach where models in the nested sequence (5.4) are each assigned a score and the smallest model with sufficiently large score is chosen.

In addition to the problem-specific interpretation of these quantities, we expect that a validity result will ensure the frequentist Type I error of this selection rule is bounded by α. However, the fact that the assertions A_k in (5.4) are data-dependent, the fixed-assertion validity theorem of Martin and Liu (2011) does not apply. Based on the simulation results in Section 5.3.5, we claim that validity does hold, but more theoretical work is needed to justify this claim.

5.3.4 The diabetes data example, cont.

When applied to the diabetes example with our initial full model, the above method produces the results tabulated in Table 1. The table shows the ordering of the covariates according to their rescaled least-squares estimates in (5.2), i.e., hdl has the smallest Z-value and bmi has the largest. The last column shows the plausibility function $\text{Pl}_{\mathbf{z},\text{S}}$ evaluated at the assertions A_k in (5.4), for $k = 1, \ldots, 11$. We start at the $k = 11$ row and proceed up the table, selecting variables until the plausibility for A_k exceeds the $\alpha = 0.05$ threshold. In this case, the IM-based approach suggests that we select bmi, map, ltg, sex, the interaction $age \times sex$ and, hence, the lower-order term age. R code for this example is available at www.stat.purdue.edu/ chuanhai.

Order, k	Covariate	Z-value	Plausibility of A_k
1	hdl	0.65	0.3579
2	glu	1.16	0.0446
3	tch	1.21	0.2541
4	ldl	1.49	0.2740
5	tc	1.98	0.0971
6	age	−2.51	0.0796
7	$age \times sex$	3.56	0.0024
8	sex	−4.54	0.0005
9	ltg	4.56	0.0000
10	map	5.13	0.0000
11	bmi	7.98	0.0000

Table 1: IM variable selection results for the diabetes example in Section 5.3.4. Note that only the leading digits are meaningful because the numerical results are obtained via simulations with Monte Carlo sample size 10,000.

The analysis of the data in Efron et al. (2004) was slightly different in that only the main effects were considered. Using Mallows' C_p criterion, LARS selects seven variables including bmi, ltg, map, hdl, sex, glu, and tc, in the order of being entered into the model. When the $age \times sex$ interaction is also considered, both the C_p and ten-fold cross validation criteria select all eleven variables.

5.3.5 A simulation study

Simulations are used to demonstrate the performance of the proposed IM approach. To keep the presentation concise, we limit our comparison to the popular lasso method implemented via the LARS algorithm of Efron et al (2004). In our experiments, we generate data from the basic model (1.1), where each row of $\mathbf{X}$ is sampled from a first-order autoregressive process with standard Gaussian marginal distribution and pairwise correlation $\rho^{|i-j|}$ between the i^{th} and j^{th} components. We consider two scenarios of the regression coefficients: $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$ where effects of important variables are large, and $\boldsymbol{\beta} = (0.85,0.85,0.85,0.85,0,0,0)$ where the important variables have relatively small effects. For each $\rho \in \{0.5,0.8\}$ and each $\boldsymbol{\beta}$ above, we simulated data with sample size $n = 50,100,200,500,1000$, each with 1000 data sets. For the LARS approach, the tuning parameter was chosen according to both the Mallows' C_p and the ten-fold cross validation approaches. The results of the simulation studies (Figures 4-7) are summarized using the percentages of *correct*, *parsimonious*, and *inclusive*. The parsimonious models contain only a subset of important variables and none of the unimportant variables, while the inclusive models contain all the important variables plus at least one unimportant variables. For the inferential model approach, we use $\alpha = 0.05$.

The results show that, in all the four scenarios, the LARS approach selects the true model only about 20% of the time, independent of the sample size. This is consistent to the findings of Leng et al. (2006) that, when the prediction accuracy is used as the criterion to choose tuning parameters, LARS and related procedures select the true set of predictors with a probability that is less than one and does not depend on the sample size. Instead, the LARS approach tends to select more predictors than necessary, which is shown by the high percentage of inclusive models in all scenarios.

Compared to LARS, which almost always gives a larger-than-necessary model, the IM approach has a much better performance in terms of choosing the correct set of variables. When the sample size is small, it tends to choose the parsimonious models with the most important variables. As the sample size increases, more and more important variables are included and the percentage of correctly selected models increases. Eventually with a sufficiently large sample

size, the percentage of correctly selected model stabilizes at about 95%, while the remaining 5% of the time, it selects an inclusive model.

Although direct comparisons of different methods in simulations can be insightful, it is important to keep in mind that the IM-based procedure is a bit more ambitious than LARS or any other variable selection procedure in the following way: to the set of variables selected by the IM procedure, there is an associated uncertainty assessment in the form of a meaningful prior-free probabilistic measure of evidence. In contrast, existing methods rely on an indirect frequentist reasoning to justify their use. But, as our simulations demonstrate, the IM-based procedure can also beat the classical methods at their own game.

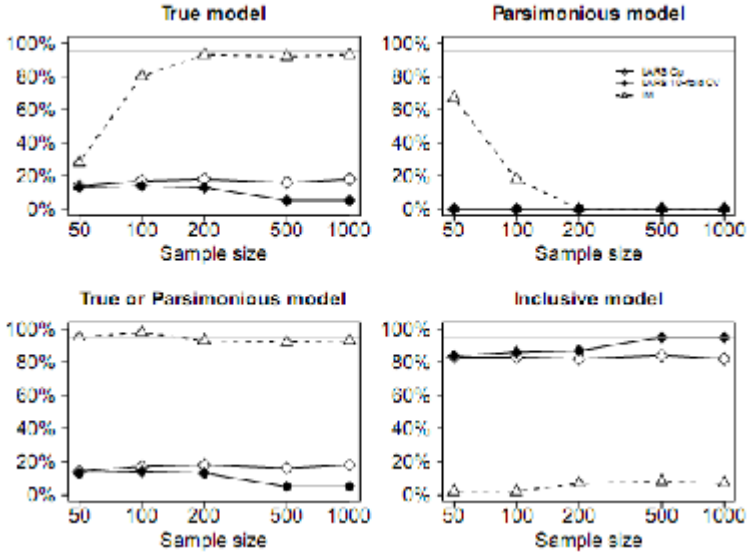

Figure 4: Percentage of selected true model, parsimonious models, and inclusive models when $\rho = 0.5$ and $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$

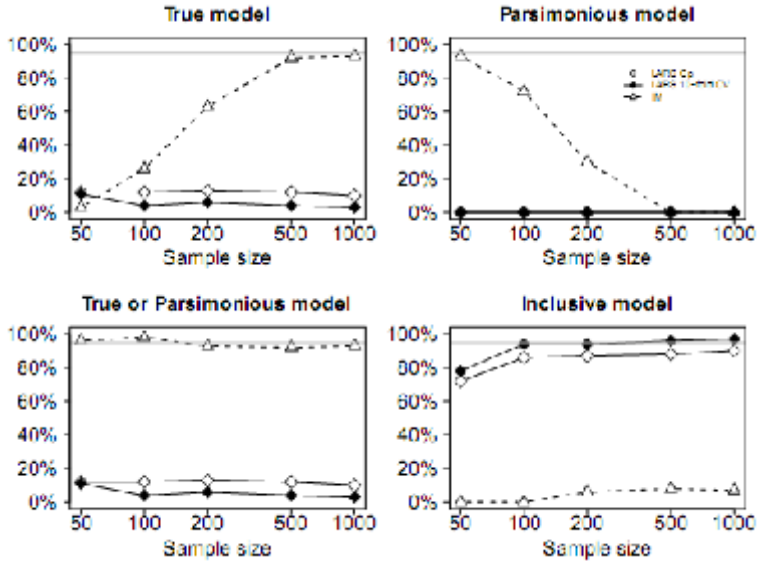

Figure 5: Percentage of selected true model, parsimonious models, and inclusive models when $\rho = 0.8$ and $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$

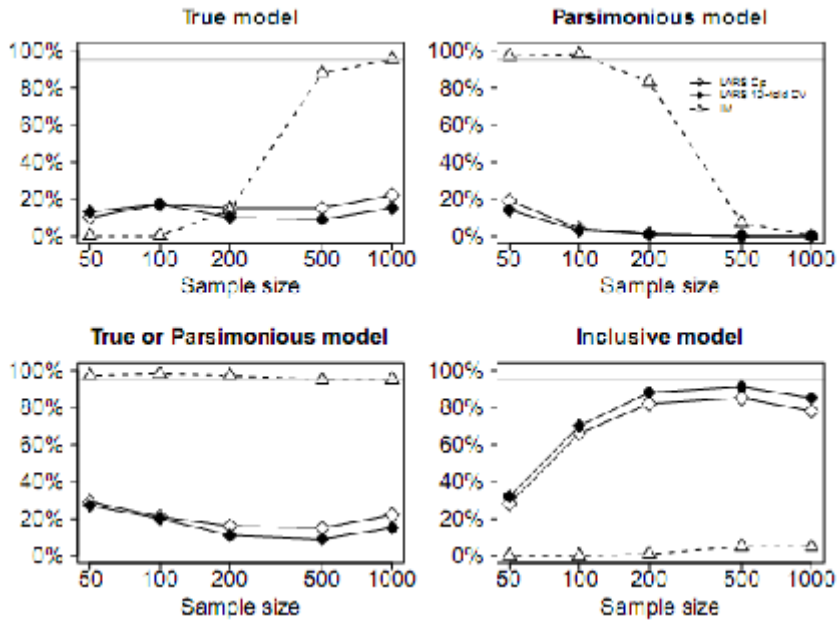

Figure 6: Percentage of selected true model, parsimonious models, and inclusive models when $\rho = 0.5$ and $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0, 0, 0)$

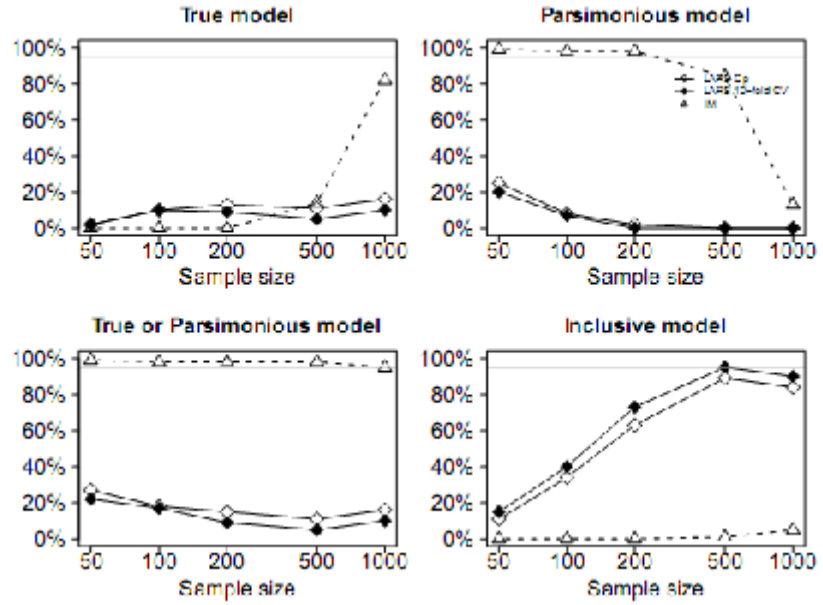

Figure 7: Percentage of selected true model, parsimonious models, and inclusive models when $\rho = 0.8$ and $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0, 0, 0)$

6. Prediction

Here we briefly discuss an IM-based procedure for prediction of future observations. Suppose that model (1.1) is to be used for prediction and that it is of interest to predict $\mathbf{Y}_{\text{new}} \in \mathrm{R}^m$, the response variables associated with m new

covariate measurements aligned as columns of the $m \times p$ matrix $\mathbf{X}_{\text{new}}$. Then the full statistical model becomes

$$\mathbf{Y} = \mathbf{X}^{\prime} + \sigma \quad , \qquad _{\text{new}} = \quad _{\text{new}}^{\prime} + \sigma \quad _{\text{new}}, \quad (\quad , \quad _{\text{new}}) \sim N_{n+m}(0, \quad).$$

By using the conditional IM specified in (3.4) and (3.5) and the usual probability calculus, we find a conditional IM for inference about $\mathbf{Y}_{\text{new}}$, given by

$$\mathbf{Y}_{\text{new}} = \mathbf{X}_{\text{new}}\mathbf{B} + S\mathbf{T}_{\text{new}}, \quad \text{with} \quad \mathbf{T}_{\text{new}} \sim t_m(0, \mathbf{I} + \mathbf{X}_{\text{new}}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}_{\text{new}}^{\text{T}}, \nu).$$

That is, inference about $\mathbf{Y}_{\text{new}}$ is based on the conditional IM obtained by conditioning on the observed direction of the residuals $\mathbf{P}^{\perp}\mathbf{U}$ in the original regression setup. This conditional IM has the m-variate Student-t vector $\mathbf{T}_{\text{new}}$ as the a-variable to be predicted. Then efficient inference about $\mathbf{Y}_{\text{new}}$ can be obtained by specifying PRSs for predicting $\mathbf{t}_{\text{new}}^{*}$ according to relevant assertions about $\mathbf{Y}_{\text{new}}$.

7. Discussion

In this paper we have elaborated on the recently proposed inferential model framework, which produces prior-free probabilistic summaries of evidence for/against assertions of interest and, moreover, these summaries have a desirable frequency-calibration property. We have demonstrated here that IMs are promising for data analysis with linear regression models by showing how they can be used for model building and checking, variable selection, and prediction. In particular, we have shown via real- and simulated-data examples that one relatively simple IM-based strategy for variable selection can outperform a popular and powerful existing method like lasso (via the LARS algorithm). We expect that this new IM framework will be met with some initial skeptisism, and foundational work to overturn this skepticism is ongoing. But we believe that the positive numerical results alone make a strong enough case to pursue IMs further.

Throughout we have assumed that the predictor matrix $\mathbf{X}$ is of full rank, but in applications it may happen that $\mathbf{X}^{\text{T}}\mathbf{X}$ is close to singular. This is not a problem theoretically, but it may lead to some computational instability in the reparametrization (5.1). However, the proposed model checking method in Section 4 and the corresponding modeling process in Section 5.2 can be used as an initial screening process whereby the set of candidate variables can be pruned prior to using a formal variable selection process.

Although Gaussian linear regression is a fundamental problem, some applications require other kinds of linear and non-linear models, such as Poisson or logistic regression. The methodology presented here is tailored specifically to the Gaussian case, so the calculations may not immediately apply in a problem with a different model. But work is underway to extend both the theory and methods described here to more general cases.

Acknowledgements

The authors thank the two referees for their helpful comments and suggestions. A portion of this work was completed while R. Martin was affiliated with the Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis. C. Liu was partially supported by the U.S. National Science Foundation, grant DMS-1007678.

References

1. Clyde, M. and George, E. I. (2004). "Model uncertainty," Statist. Sci., 19, 81–94.
2. Dempster, A. P. (2008). "Dempster-Shafer calculus for statisticians," Internat. J. of Approx. Reason., 48, 265–277.
3. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression," Ann. Statist., 32, 407–499, with discussion, and a rejoinder by the authors.
4. Ermini Leaf, D. (2011). "Inferential models and restricted spaces," Ph.D. thesis, Purdue University, West Lafayette, IN.
5. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning, New York: Springer-Verlag, 2nd ed.
6. Heaton, M. J. and Scott, J. G. (2010). "Bayesian computation and the linear model," in Frontiers of Statistical Decision Making and Bayesian Analysis, eds. Cheh, M.-H., Dey, D., M¨uller, P., Sun, D., and Ye, K., Springer, pp. 527–545.
7. Leng, C., Lin, Y., and Wahba, G. (2006). "A note on the lasso and related procedures in model selection," Statist. Sinica, 16, 1273–1284.
8. Martin, R., Hwang, J.-S., and Liu, C. (2011a). "Conditional inferential models," Submitted manuscript. www.stat.purdue.edu/∨chuanhai.— (2011b), "Marginal inferential models," Submitted manuscript. www.stat.purdue.edu/∨chuanhai.
9. Martin, R. and Liu, C. (2011). "Inferential models," Submitted manuscript, www.stat. purdue.edu/∨chuanhai.
10. Martin, R., Zhang, J., and Liu, C. (2010). "Dempster-Shafer theory and statistical inference with weak beliefs," Statist. Sci., 25, 72–87.
11. Shafer, G. (1976). A mathematical theory of evidence, Princeton, N.J.: Princeton University Press.
12. Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso," J. Roy. Statist. Soc. Ser. B, 58, 267–288.
13. Zhang, J. (2010). "Statistical inference with weak beliefs," Ph.D. thesis, Purdue University, West Lafayette, IN.
14. Zhang, J. and Liu, C. (2011). "Dempster-Shafer inference with weak beliefs," Statist. Sinica, 21, 475–494.