SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2010-73

Estimating the effect of a variable in a high-dimensional regression model

Peter Sandholt Jensen and Allan H. Würtz

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Estimating the effect of a variable in a high-dimensional regression model[*]

Peter Sandholt Jensen[†] Allan H. Würtz[‡]

November 24, 2010

Abstract

A problem encountered in some empirical research, e.g. growth empirics, is that the potential number of explanatory variables is large compared to the number of observations. This makes it infeasible to condition on all variables in order to determine whether a particular variable has an effect. We assume that the effect is identified in a high-dimensional linear model specified by unconditional moment restrictions. We consider properties of the following methods, which rely on low-dimensional models to infer the effect: Extreme bounds analysis, the minimum t-statistic over models, Sala-i-Martin's method, BACE, BIC, AIC and general-to-specific. We propose a new method and show that it is well behaved compared to existing methods.

Keywords: AIC, BACE, BIC, extreme bounds analysis, general-to-specific, robustness, sensitivity analysis

Jel: C12, C51, C52

[*] *Acknowledgements*: We are thankful for detailed comments from Tue Gørgens and discussions with Martin Browning, Gernot Doppelhofer, John Geweke, Clive Granger, Joel Horowitz and Gene Savin. Comments from seminar participants at the University of Aarhus, the University of Copenhagen, the University of Helsinki and the NASM 2006 are appreciated. Allan Würtz acknowledges the support from the Centre for Applied Microeconometrics (CAM) and CREATES. The activities of CAM and CREATES are financed from a grant by the Danish National Research Foundation.

[†] Department of Business and Economics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. E-mail: psj@sam.sdu.dk.

[‡] CREATES, School of Economics and Management, University of Aarhus. Building 1322, DK-8000 Aarhus C, Denmark. E-mail: awurtz@econ.au.dk

1 Introduction

The objective of many empirical studies is to determine the effect of an explanatory variable. When there are many other explanatory variables and only a small dataset is available, researchers often resort to low-dimensional models, where only a subset of the explanatory variables are included at a time, instead of a high-dimensional model with all explanatory variables included. In particular, when the number of observations exceeds the number of explanatory variables, it is necessary to use a dimension reduction method. The most common dimension reduction methods applied in economics all build on estimation of many low-dimensional models and on rules for combining the results of the low-dimensional models to an estimate of the effect. The purpose of this paper is to investigate the properties of these methods and, as a result of the findings, suggest an improved method.

Many of the dimension reduction methods are applied in studies of GDP per capita growth. One reason is that there are many potential variables that have been claimed to have an effect on growth, see e.g. Durlauf, Johnson and Temple (2005), who list 145 variables. One of the most extensive growth datasets is the one of Sala-i-Martin (1997). This dataset has around 60 variables, but only 36 complete country observations. Thus, the dataset is high-dimensional, i.e. it has more variables than observations (e.g. Goeman, Van der Geer, Houweiling, 2006; Jensen, 2010). Many of the papers in the empirical growth literature use this dataset (e.g. Fernandez, Ley and Steel, 2001; Hoover and Perez, 2004; Hendry and Krolzig, 2004; Sturm and De Haan, 2005; Acosta-Gonzalez and Fernandez-Rodriguez, 2007). Therefore, standard regression techniques with all variables included cannot be carried out, and as a consequence growth researchers have turned to alternative (dimension reduction) methods (e.g. Sala-i-Martin, 2001; Durlauf et al., 2005).

Our setting is as follows. There is a plethora of potentially important variables for an outcome. The researcher has a particular variable of interest. The effect of the variable of interest is identified as the partial effect in a model with all potential explanatory variables included. This is denoted the high-dimensional model. The high-dimensional model is specified by linear unconditional moment restrictions. This specification includes a linear regression model. For various reasons, for instance an undersized sample with more variables than observations, the researcher relies on models with fewer explanatory variables than the high-dimensional model. These models are the low-dimensional models. They are also specified by linear unconditional moment restrictions. The low-dimensional models can be estimated e.g. by the ordinary least squares estimator. Based on different rules, the results of the low-dimensional models are combined into an (interval) estimate of the effect.

We start our analysis by providing three different assumptions under which the effect is identified by a low-dimensional model. As is well known from the omitted variable bias problem, the effect of a variable in the high-dimensional model cannot, in general, be identified in a low-dimensional model. As is also well known, there are exceptions where a low-dimensional model can identify the effect. Three cases are 1) the variable of interest is conditional mean independent of the excluded variables, 2) the excluded variables do not explain the outcome, and 3) a variable can be used as an instrument for the explanatory variable of interest. Though these three cases constitute the essence of the three assumptions, respectively, the context here is different from the usual omitted variable bias setting. The reason is that all variables are observed (even though they may be perfectly correlated) and that it is not known beforehand which variables, if any, can be excluded such that e.g. the conditional mean independence holds. Hence, the assumptions are more realistic than simply assuming that it is known, say, which explanatory variables do no belong in the model. It is under these assumptions that we investigate the properties of the various methods.

The first set of methods is Bayesian in spirit and builds on Leamer (1983). His approach is to run a set of linear regressions that include the variable of interest and different selections of other explanatory variables. In our terminology, these are the low-dimensional models. If the coefficient on the variable of interest is significant and has the same sign in all the low-dimensional models, he denotes the variable as "robust". The method is known as extreme bounds analysis (EBA) and was first implemented in a growth context by Levine and Renelt (1992). Hansen (2003) develops a variant of Leamer's approach which takes the multiple testing problem into account and uses the bootstrap method proposed by White (2000). Sala-i-Martin (1997) criticizes extreme bounds analysis because a variable is likely to be insignificant in at least one regression if enough regressions are run. As an alternative, Sala-i-Martin suggests a method based on the distribution of the estimates of the effect over different low-dimensional models. Sala-i-Martin, Doppelhofer and Miller (2004) build on the Bayesian Model Averaging technique proposed by Raftery (1995). Their approach is known as Bayesian Averaging of Classical Estimates (BACE). The BACE approach involves a weighted average of the estimates over the different low-dimensional models. In the implementation of Sala-i-Martin et al. (2004), this weighted average is used to assess robustness.

The second set of methods consists of classical model selection methods. Criteria such as AIC and BIC can be employed to choose the subset of variables to be included in the "best" model, and whether a variable has an effect on the outcome is determined by whether it is included in the best model or not. In our terminology, one low-dimensional model is picked among the set of low-dimensional models estimated. The refined general-to-specific procedures suggested by Hoover and Perez (2004), Bleaney and Nishiyama (2002) and Hendry and Krolzig (2004) can be applied to various low-dimensional models to pick a low(er)-dimensional model.

We prove that none of the methods identify the effect under the assumptions based on conditional mean independence, 1), or instrumental variable, 3). The traditional model selection techniques BIC and AIC, together with BACE and general-to-specific identify the effect when a subset of the variables does not explain the outcome, 2); that is, when there is a low-dimensional model which is identical to the high-dimensional model. None of the methods are designed to, nor do they, reveal whether an assumption is, in fact, satisfied. Hence, in terms of identification of the effect without imposing one of the assumptions a priori, none of the methods provide valid inference.

We suggest a new method that provides valid inference without imposing any of the three assumptions. It can, however, only identify the effect if the conditional mean independence assumption is satisfied. If this assumption is not satisfied, the method returns no answer to the effect. A key part of the method is to test whether there is a sufficiently large number of variables that can be excluded because they are conditional mean independent of the variable of interest. It is possible to perform these tests even when the sample is undersized. The possibility of testing with an undersized sample has also been applied by Breusch (1986) and Jensen (2010). Our method first uses tests to reveal if there is a set of variables for which the conditional mean assumption is satisfied and, if so, estimates the low-dimensional model with an appropriate set of explanatory variables.

Using Monte Carlo simulations, we investigate the finite sample properties of the methods in a setting with more variables than observations. The estimators of the effect are substantially biased. As a result, tests of the effect being different from 0 are size-distorted and, in particular, have poor power. Some of the methods have a higher probability of accepting that a variable has an effect in the case where it has none. The Monte Carlo study shows that none of the existing methods work when there are more explanatory variables with non-zero coefficients than observations. It also demonstrates that the new method has good level control in detecting when the conditional mean independence assumption holds, and, in that case, the new method has the correct size and good power against the null of no effect.

We finally apply the new method to Sala-i-Martin's (1997) dataset. The results show that the effects on economic growth of the variables cannot be inferred. While this result is negative, this is not unusual in the literature, see Jensen (2010). Thus, our theoretical and empirical results raise the question of what is actually learnt from the various studies based on dimension reduction methods using for example Sala-i-Martin's (1997) dataset.

The rest of the paper is organized as follows. In Section 2 we state conditions under which the effect of a variable can be identified from a low-dimensional model. In Section 3, we derive properties of the existing methods. Section 4 describes the new method for estimating and testing the effect of a variable, and in Section 5 the different methods are compared in a Monte Carlo study. Section 6 contains the application and Section 7 concludes. All proofs are in the appendix.

2 Identifying an effect using a low-dimensional model

In this section, we give conditions under which the effect of a variable in a high-dimensional model is identified in a low-dimensional model.

We start by a simple example to illustrate circumstances under which the effect in the high-dimensional model can be identified in a low-dimensional model. These circumstances are well-known but worth mentioning here because they cast light on the working of the existing methods discussed in the next section as well as the method we develop in Section 4. Throughout the paper, we assume that the effect of a variable is identified by unconditional moment restrictions in the high-dimensional model. In this simple example, assume that:

$$\begin{aligned} Y &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + U, \\ E(X_1 U) &= E(X_2 U) = E(X_3 U) = E(X_4 U) = 0. \end{aligned} \tag{1}$$

The *effect* of the variable X_1 is defined to be the partial effect β_1.

A low-dimensional model based on linear regression may identify the effect. The linear regression[1] of Y on X_1 and X_2 is:

$$Y = \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon.$$

As is well known, β_1 is not identified by γ_1 due to an omitted variable bias. The omitted variable bias disappears if the included explanatory variables, X_1 and X_2, are conditional mean independent of the excluded variables X_3 and X_4. The omitted variable bias also

[1] In a regression context, Goldberger (1991) denotes the high-dimensional model as the *long* regression and the low-dimensional model as the short regression.

disappears if the excluded variables do not belong in the high-dimensional model, that is, $\beta_3 = \beta_4 = 0$. This circumstance amounts to the low-dimensional model being a *true submodel*.

A low-dimensional model based on instrumental variable regression may also identify the effect. For an instrument, say X_3, to be valid in a low-dimensional model with only the explanatory variable X_1:

$$E(X_1X_3) \neq 0 \text{ and } E(X_2X_3) = E(X_4X_3) = 0.$$

If $\beta_3 = 0$, then X_3 is a valid instrument in the instrumental variable regression of Y on X_1. The coefficient on X_1 identifies β_1.

We finish by stating identifying assumptions in general. Assume that the high-dimensional model consists of K explanatory variables and the low-dimensional model includes at most $K_s < K$ explanatory variables. The effect, β_1, of the variable X_1 is identified in the high-dimensional model by the following unconditional moment restrictions:

$$\begin{aligned} Y &= \sum_{k=1}^{K} \beta_k X_k + U \\ E(X_1U) &= ..... = E(X_KU) = 0. \end{aligned} \tag{2}$$

Without loss of generality, assume that $E(X_k) = 0$, $k = 1,..,K$. The specification of the high-dimensional model is satisfied by a linear regression model. The specification can also be thought of as the minimum mean square linear predictor of a (nonlinear) regression function. Equivalently, we will also write the model as

$$E^*(Y \mid X_1,..,X_K) = \sum_{k=1}^{K} \beta_k X_k$$

where $E^*(Y \mid X_1,..,X_K)$ is the linear projection of Y on $X_1,..,X_K$, see e.g. Wooldridge (2002) for details. A linear projection satisfies

$$E\left(X_k\left[Y - E^*(Y|X_1,..,X_K)\right]\right) = 0,\ k = 1,..,K$$

A low-dimensional model based on linear regression with K_s explanatory variables is

$$\begin{aligned} Y &= \gamma_1X_1 + \gamma_2X_2 + .. + \gamma_{K_s}X_{K_s} + \varepsilon \\ E(X_1\varepsilon) &= ..... = E(X_{K_s}\varepsilon) = 0. \end{aligned}$$

A relationship between the low-dimensional and the high-dimensional model can be derived using linear projection. Let $E^*(X_m|X_1,..,X_{K_s})$ be the linear regression (projection) of X_m on $X_1, .., X_{K_s}$ given by

$$E^*(X_m|X_1, .., X_{K_s}) = \sum_{k=1}^{K_s} \alpha_k^m X_k$$

Then the linear regression of Y on $X_1, .., X_{K_s}$ is

$$\begin{aligned} Y &= E^*(Y|X_1, .., X_{K_s}) + \varepsilon \qquad (3) \\ &= \sum_{k=1}^{K_s} \beta_k X_k + \sum_{m=K_s+1}^{K} \beta_m E^*(X_m|X_1, .., X_{K_s}) + \varepsilon \\ &= \sum_{k=1}^{K_s} \beta_k X_k + \sum_{m=K_s+1}^{K} \beta_m \sum_{k=1}^{K_s} \alpha_k^m X_k + \varepsilon \\ &= \sum_{k=1}^{K_s} \left(\beta_k + \sum_{m=K_s+1}^{K} \beta_m \alpha_k^m \right) X_k + \varepsilon. \end{aligned}$$

The first line defines $\varepsilon \equiv Y - E^*(Y|X_1, .., X_{K_s})$. Hence, by construction $E(X_1\varepsilon) = ... = E(X_{K_s}\varepsilon) = 0$ and the conclusion is that the coefficient on X_k in the low-dimensional model is

$$\gamma_k = \beta_k + \sum_{m=K_s+1}^{K} \beta_m \alpha_k^m, \text{ for } k = 1, .., K_s. \qquad (4)$$

In general, $\gamma_k \neq \beta_k$ and, thus, the effect β_1 is not identified by γ_1.

We can now state two assumptions about identification of the effect based on low-dimensional linear regression models. The first assumption concerns conditional mean independence between included and excluded variables. We will only assume that there exists a set of variables, which are conditional mean independent of X_1, the variable of interest. We will not assume which of the variables belong to this set.

Assumption (O) Conditional mean independence: Let $A \subset \{X_2, .., X_K\}$ with $(K - K_s)$ members and A^c the complement of A. There exists a set A such that for each $X_i \in A$, the coefficient, α_1^i, on X_1 is 0 in the linear regression of X_i on X's in A^c.

The method we propose in Section 4 has a search for the two sets A and A^c as a vital step. It turns out that assumption (O) is testable also when the sample is undersized.

The second assumption based on low-dimensional linear regression models concerns the equivalence of the high-dimensional model and one low-dimensional model. We therefore denote this case the true submodel assumption.

Assumption (S) True submodel: At least $(K - K_s)$ of the coefficients $(\beta_2, .., \beta_K)$ equal 0.

Assumption (S) only says that there is a true low-dimensional model but not which one of them. Jensen (2010) showed that this assumption is testable when the sample size, n, is smaller than K.

As indicated above in the simple example, an instrumental variables approach may be possible. In the following assumption, it is assumed that there exists one variable which is correlated with X_1 and uncorrelated with the other variables.[2]

Assumption (I) Instrumental variable: There exists a variable, $X_k, k \neq 1$, such that $\beta_k = 0$, $E(X_k X_1) \neq 0$ and $E(X_k X_m) = 0$ for $m = 2, .., K$, $m \neq k$.

Similar to the two other assumptions, assumption (I) does not indicate which of the variables satisfy the restrictions in the assumption, only that such variables exist.

In the next three sections, we analyze existing methods and propose a new method for identifying the effect β_1 in view of the assumptions stated in this section.

3 Identifying the effect with existing methods

In this section we investigate existing Bayesian and classical methods that build on low-dimensional models to infer the effect in a high-dimensional model. None of the methods have been analyzed before under all the identifying assumptions in Section 2 and, thus, the analysis helps fill a void, see Durlauf (2001) and Durlauf et al. (2005). Under each of the three assumptions in Section 2, we prove whether or not a method identifies the presence of the effect and, if so, the size of the effect.

Subsections 3.1-3.6 analyze different methods: EBA, Minimum t-statistic over models, Sala-i-Martin's method, BACE, BIC, AIC, general-to-specific. A preview of the results is given in Table 1**.** The table shows whether or not a method correctly determines if X_1 has an effect under each of the three assumptions.

[2] The methods applied in the empirical growth literature almost solely buid on regressions, there are exceptions. For example, Durlauf, Kourtellos and Tan (2008) use the Bayesian Model Averaging approach based on instrumental variables.

Table 1: Correctly identifying the presence of an effect: $\beta_1 = 0$ or $\beta_1 \neq 0$.

Method	Assumption (O)	Assumption (S)	Assumption (I)
EBA	Partly	No	No
Minimum t	Almost	No	No
Sala-i-Martin	No	No	No
BIC, AIC	No	Yes	No
BACE	No	Yes	No
GSP	No	Yes	No

Notes: Minimum t=minimum t-statistic over models, Sala-i-Martin=Sala-i-Martin's method, GSP=general-to-specific.

It is notable that no method works under assumptions (O) and (I). None of the methods include a test of the assumption being satisfied. Hence, even when some of the methods work under assumption (S), it is necessary a priori to know that the assumption is satisfied. In practice, it is unlikely that it is known whether one of the assumptions holds. In contrast, the method we develop in Section 4 includes a test of the assumption being satisfied.

In the analysis, we use asymptotic techniques to derive the identification properties of the various methods. To preserve the defining property of the problem of using low-dimensional models to infer the effect in the high-dimensional model, we keep K_s and K fixed as $n \to \infty$. Hence, the main aspect of e.g. an undersized sample does not disappear in the asymptotic analysis. In Section 5 we investigate the finite sample behavior of the methods, and we relate the finite sample behavior to the identification results in this section.

In the next subsections, we focus on the properties of the various methods to estimate the effect β_1 under the assumptions (O) and (S). We do not analyze assumption (I) further because none of the methods considered are based on instrumental variable estimators. For simplicity, we assume throughout a random sample of $\{Y, X_1, .., X_K\}$ is available.

3.1 EBA

The EBA of Leamer (1983), Leamer and Leonard (1983), and Levine and Renelt (1992) defines the variable X_1 as robust if the estimates of its coefficient are significantly different from 0 and have the same sign in all the low-dimensional models. Other authors have slightly different definitions of robustness, see the next subsections. All authors agree, however, that the idea of robustness is to determine whether or not the variable has an effect, i.e. whether or not $\beta_1 = 0$, see the discussions in Sala-i-Martin (2001) and Durlauf, et al. (2005). Therefore, we treat robustness as an estimator of whether a variable has an effect. Most of the other methods provide a point estimator of β_1, whereas the EBA provides an interval.

EBA has been criticized by various authors. Despite this criticism, the extreme bounds analysis continues to enjoy widespread popularity.[3] McAleer, Pagan and Volker (1985) give necessary and sufficient conditions under which a variable is robust. Breusch (1990) calculates the extreme bounds based on the high-dimensional model, and notes that the bounds are narrow when exclusion of variables do not result in a loss of fit. Granger and Uhlig (1990) derive the extreme bounds over the low-dimensional models that have a reasonable fit (in terms of R^2) relative to the best and worst fitting models. McAleer (1994) criticizes Levine and Renelt (1992) for not reporting diagnostic tests. A recent review of the different results is given by Ericsson (2008:pp. 897-898) who also derives new results regarding the EBA. He obtains his result for the case in which there is an omitted variable which is never included in the linear regressions used to calculate the extreme bounds. In our analysis, all variables are included in some linear regressions.

Let γ_1^i be the population coefficient on X_1 in the i^{th} linear regression of Y on X_1 and at most $(K_s - 1)$ other variables, and let the set of all such linear regressions be $\mathcal{F}$. The next proposition concerns the population properties of extreme bounds under assumptions (O) and (S) from Section 2.

Proposition 1 (EBA) *The EBA selects the interval* $I_{EBA} \equiv \left[\min_{i \in \mathcal{F}} \gamma_1^i \ , \ \max_{i \in \mathcal{F}} \gamma_1^i\right]$ *for the population effect of* X_1*.*

Under assumption (O)*, the EBA identifies an interval containing the effect:* $\beta_1 \in I_{EBA}$*. It can determine that* X_1 *has an effect* $(\beta_1 \neq 0)$ *on* Y *if* $0 \notin I_{EBA}$*.*

Under assumption (S)*, the EBA does not identify an interval containing the effect* β_1*, nor can it determine whether* X_1 *has an effect on* Y*.*

The proposition shows that EBA is not a consistent procedure for determining whether a variable has an effect for Y in the high-dimensional model. Under the conditional mean independence assumption (O), EBA identifies an interval which contains the effect. If the interval only has one member then β_1 is identified. In any other case, the coefficient on X_1 can change sign across linear regressions such that the interval contains both positive and negative values. This result is in line with proposition 4 in Ericsson (2008), which

[3] For recent applications, see e.g. Ghosh and Yamarik (2004) and Berggren, Elinder and Jordahl (2008).

states that a result may be non-robust when the set of models for EBA excludes the true model. In our case, this means that $\beta_1 \neq 0$, but the interval includes zero. Under the true submodel assumption (S) there is no guarantee that the extreme bounds contain 0 in cases where the true submodel has K_s explanatory variables and does not include X_1. This result is in line with proposition 2 in Ericsson (2008), which states that a result may be robust when the set of models for EBA excludes the true model, i.e. robustness is possible with $\beta_1 = 0$.

3.2 Minimum t-statistic over models test

The minimum t-statistic over models test is carried out as follows. First, all linear regression models in $\mathcal{F}$ are run and the t-statistics for testing for inclusion of X_1 are carried out. Second, the minimum t-statistic (in absolute value) is found. There is an effect if the coefficient with the smallest t-statistic is significantly different from zero. This is equivalent to the t-statistic, t_i, in linear regression, i, exceeding the appropriate critical value since $P(|t_i| \ > c, \forall i) = P\left(\underset{i}{Min}|t_i| \ > c\right)$. White (2000) and Hansen (2003) have shown that the bootstrap can be applied to approximate the distribution of the minimum t-statistic under different regularity conditions. Just as with EBA, the approach does not provide an estimator of the partial effect. The following proposition provides the population properties of the minimum t-statistic over models test.

Proposition 2 (Minimum t-statistic over models test) *Under assumptions (O) or (S), the minimum t-statistic over models test does not identify whether X_1 has an effect on Y.*

The method almost works under assumption (O). It works when $\beta_1 = 0$, and it works when $\beta_1 \neq 0$ except when an omitted variable bias exactly offsets β_1. The set of β_1's, $\beta_1 \neq 0$, for which the omitted variable bias cancels the effect of X_1 has Lebesgue measure 0. The minimum t-statistic does not work under assumption (S), because the true submodel may not include X_1, but X_1 is always included in the low-dimensional models, and the estimators of β_1 are biased.

3.3 Sala-i-Martin's method

Sala-i-Martin (1997) motivates his approach as an alternative to the EBA, which better takes sampling uncertainty into account.[4] He considers a setup in which all the low-

[4] Sala-i-Martin's article has been very influential and has 240 citations in the Social Sciences Citation Index. Recent applications are in e.g. Sturm and de Haan (2005), de Haan (2007), Dreher, Sturm and

dimensional models have the same number of explanatory variables and always include the variable of interest X_1. Among the different versions of the method he presents, we focus on the one from his general setup:

$$CDF(0) = \sum_{i=1}^{m} w_i CDF_i(0),$$

where w_i is the weight of linear regression i, $CDF_i\,(0) = Max(\Phi(\widehat{\gamma}_1^i/\widehat{\sigma}_{\widehat{\gamma}_1^i}), 1 - \Phi(\widehat{\gamma}_1^i/\widehat{\sigma}_{\widehat{\gamma}_1^i}))$, $\widehat{\gamma}_1^i$ is the OLS estimator, and $\widehat{\sigma}_{\widehat{\gamma}_1^i}$ is the standard error. The quantity $CDF_i(0)$ can be interpreted as the largest of the two following p-values: The p-value from the one-sided tests of the coefficient on X_1 being 0 against larger than 0 and the p-value from the one-sided test against the coefficient being below 0. A variable is robust in Sala-i-Martin's terminology if $CDF\,(0)$ is larger than 0.95. Sala-i-Martin assumes conditional normality of Y in all the linear regressions. The weight of model j is then defined as:

$$w_j = \frac{SSE_j^{-n/2}}{\sum_{i=1}^{m} SSE_i^{-n/2}},$$

where SSE_j is the sum of squared errors in model j. Sala-i-Martin uses $\hat{\gamma}_1^{SiM} = \sum_{i=1}^{m} w_j \hat{\gamma}_1^i$ as an estimator of β_1 in another of his setups. We also use it for the general setup.

The next proposition shows that Sala-i-Martin's method cannot determine if X_1 has an effect because it does not identify β_1. The properties of the Sala-i-Martin's method depends on the best fitting low-dimensional linear regression model. The measure of fit of a model is $E(V^*(Y \mid C))$, where $C \subset \{X_1, .., X_K\}$ and

$$V^*(Y \mid C) \equiv E((Y - E^*(Y \mid C))^2 \mid C)$$

We will denote the lower the value of $E(V^*(Y \mid C))$, as the better fit.

Proposition 3 (Sala-i-Martin's method) *Let $\underline{Z}$ be a subset of $\{X_2, .., X_K\}$ with $(K_s - 1)$ members. Sala-i-Martin's method selects the coefficient of X_1 from the linear regression in $\mathcal{F}$ with minimum $E(V^*\,(Y|X_1, \underline{Z}))$ as the population effect of X_1. In case several linear regressions achieve the minimum $E(V^*\,(Y|X_1, \underline{Z}))$, the effect is a weighted average of the coefficients of X_1 in those linear regressions.*

Under assumptions (O) or (S), Sala-i-Martin's method does not identify the effect nor can it determine whether X_1 has an effect on Y.

Vreeland (2009).

Sala-i-Martin's method chooses the best fitting (in terms of minimum $E(V^*(Y|X_1, \underline{Z}))$) population low-dimensional linear regression model with X_1. As a consequence, the method cannot determine whether X_1 has an effect on Y under any of the assumptions. It fails to work under assumption (O), because the only low-dimensional linear regression with an unbiased estimator of β_1 is the one with the conditional mean independent regressors and that low-dimensional model may not be the best fitting. The method does not work under the true submodel assumption (S) because the true submodel may not include X_1, and therefore the true submodel is not among the models contributing to the estimator, $\hat{\gamma}_1^{SiM}$, of β_1. The true submodel with X_1 added, however, is the only regression guaranteed to provide an unbiased estimator of β_1. The method would be consistent under assumption (S) if it is modified to a search over all low-dimensional linear regression models or if, in addition, it is assumed that at most $K_s - 1$ coefficients are different from 0.

3.4 Model selection criteria: BIC and AIC

Model selection criteria are usually based on a penalized likelihood value, see e.g. Burnham and Anderson (2002). Whether a given variable has an effect is determined by whether or not it is included in the selected model. If it is included, its effect is estimated by the coefficient in the selected model. To analyze the AIC- and BIC-based procedures it is necessary to make an assumption about the conditional distribution of Y. We assume a normal distribution for comparability with BACE analyzed in the next subsection.

One model selection criterion is BIC (Schwarz information criterion). The BIC for model j is:

$$BIC_j = n \log \frac{1}{n} SSE_j + \log(n) K_j,$$

where σ_j^2 is the maximum likelihood estimate of the variance of the error associated with model j, and K_j is the number of parameters in model j. The next proposition gives the results for BIC.[5]

Proposition 4 (BIC) *Let $\underline{Z}$ be a subset of $\{X_1, X_2, .., X_K\}$ with at most K_s members. Assume conditional normality of Y. BIC selects the coefficient of X_1 in the linear regression model with minimum $E(V^*(Y|\underline{Z}))$ as the population effect of X_1.*

Under assumption (O), BIC does not identify the effect, nor whether X_1 has an effect on Y.

Under assumption (S), BIC identifies the effect and, thus, whether X_1 has an effect on Y.

[5] For example, BIC is the basis of an algorithm used to search for important growth regressors by Acosta-Gonzalez and Fernandez-Rodriguez (2007) using the dataset by Sala-i-Martin (1997).

The BIC method identifies the effect under assumption (S), because the true submodel minimizes $E(V^*(Y|\underline{Z}))$. Under assumption (O), there is no guarantee that X_1 is included in the best fitting linear regression even if it has an effect. If this is the case, the partial effect of X_1 is estimated to be 0. It may also happen that the best fitting linear regression includes X_1 when $\beta_1 = 0$ due to omitted variable bias. The conclusion under assumption (S) confirms the well-known fact that BIC is a consistent model selection criterion, see e.g. McQuarrie and Tsai (1998:p.41).

Another model selection criterion is the Akaike information criterion, AIC, and its corrected version, AICC. The AIC and AICC for model j are given by:

$$\begin{aligned} AIC_j &= n \log \frac{1}{n} SSE_j + 2K_j \\ AICC_j &= AIC_j + \frac{2K_j (K_j + 1)}{n - K_j - 1}. \end{aligned}$$

The next proposition shows that AIC and AICC have properties similar to BIC.

Proposition 5 (AIC and AICC) *Let $\underline{Z}$ be a subset of $\{X_1, X_2, .., X_K\}$ with at most K_s members. Assume conditional normality of Y. AIC and AICC select the coefficient of X_1 in the linear regression model with minimum $E(V^*(Y|\underline{Z}))$ as the population effect of X_1.*

Under assumption (O), AIC and AICC do not identify the effect, nor do they determine whether X_1 has an effect on Y.

Under assumption (S), AIC and AICC identify the effect and, thus, they determine whether X_1 has an effect on Y.

Both AIC and BIC identify the effect under assumption (S), but they do so by different linear regressions. To see this, consider an example in which $K_s = 2$ and X_2 is the only variable with an effect ($\beta_1 = 0$, $\beta_2 \neq 0$, $\beta_3 = .. = \beta_K = 0$). In this case, BIC selects the regression of Y on X_2 with probability 1, whereas AIC selects any linear regression which includes X_2 with positive probability. In those linear regressions, the coefficient on the other variable equals 0. The reason is that AIC has a positive probability of selecting models that nest the true model. This result reflects the known fact that AIC is inconsistent for model selection, see McQuarrie and Tsai (1998:p.41) when the size of the model is fixed as the sample size increases in the asymptotic analysis.

3.5 BACE

A simplified version of Bayesian Model Averaging is implemented by Sala-i-Martin et al. (2004). Their version is called BACE. A closely related application of Bayesian Model

Averaging is considered by Fernandez et al. (2001). The implementation described here is also used by Jones and Schneider (2006) and Jensen (2010). It is necessary to assume a distribution of Y. Following Sala-i-Martin et al. (2004) we assume conditional normality of Y.

All linear regressions are included in the averaging, also the ones without the variable of interest. Let C^* be the total number of low-dimensional linear regressions. The posterior probability of the j'th linear regression, M_j, is:

$$P\left(M_j \mid y\right) = \frac{P\left(M_j\right) n^{-K_j/2} SSE_j^{-n/2}}{\sum_{i=1}^{C^*} P\left(M_i\right) n^{-K_i/2} SSE_i^{-n/2}}, \tag{5}$$

where $P\left(M_i\right)$ is the prior probability of model i. Sala-i-Martin et al. suggest using $\bar{k}/K$ as prior probability for each variable, where $\bar{k}$ is the average model size. The BACE estimator, $\widehat{\gamma}_1^{SDM}$, of β_1 is the weighted average of the estimators from each model with posterior model probabilities as the weights:

$$\widehat{\gamma}_1^{SDM} = \sum_{i=1}^{C^*} \widehat{\gamma}_1^i P(M_i \mid y),$$

where $\widehat{\gamma}_1^i$ is the estimator of β_1 in model i.

The next proposition states properties of BACE.

Proposition 6 (BACE) *Let $\underline{Z}$ be a subset of $\{X_1, X_2, .., X_K\}$ with at most K_s members. Assume conditional normality of Y. BACE selects the coefficient of X_1 in the linear regression with minimum $E(V^*(Y|\underline{Z}))$ as the population effect of X_1. In case several linear regressions achieve the minimum $E(V^*(Y|\underline{Z}))$, the partial effect is a weighted average of the coefficients on X_1 in those linear regressions.*

Under assumption (O), BACE does not identify the effect and fails to determine whether X_1 has an effect on Y.

Under assumption (S), BACE identifies the effect and, thus, determines whether the variable X_1 has an effect on Y.

The proposition shows that BACE is similar to BIC. In fact, it can be shown that the posterior probability of a model in the Bayesian averaging approach by Sala-i-Martin et al. (2004) is a function of BIC when the conditional distribution of Y is normal. In the appendix, we show that the true model will get a posterior model probability approaching 1 as $n \to \infty$. This is similar to the consistency result for BIC.

3.6 General-to-specific

The basic general-to-specific procedure has been refined by Hendry and Krolzig (1999, 2004) and Hoover and Perez (1999, 2004).[6] In a sufficiently large sample case, the procedure begins with a "general" unrestricted model (called GUM) that cannot be rejected by a host of misspecification tests. Then the procedure searches over different paths where the model is restricted until all variables are significant. The restricted models are also subjected to misspecification tests and a path may be abandoned if models do not pass the tests. In the end, a model is chosen that cannot be rejected by misspecification tests nor by encompassing tests against candidate models from other paths. Hendry (1995) calls this a congruent model.

When the dataset is high-dimensional a general unrestricted model cannot be estimated. Therefore, we perform general-to-specific on each low-dimensional linear regression with the maximum number of regressors, K_s. Among the models selected by the general-to-specific procedure for each of these linear regressions, we choose the best. The procedure is similar to the one described by Hansen (1999) in a time series context. The procedure is:

1. Select a subset of K_s regressors.

2. Delete the variable with the lowest insignificant t-statistic. Reestimate and continue until all coefficients are significant.

3. Repeat 1 and 2 for all combinations of the regressors.

4. Among the candidate models, choose the one with the lowest standard error, $E(V^*(Y \mid \underline{Z}))$.

Usually, there are steps involving misspecification testing. In our setup, we simplify the setup of the general-to-specific method. We do not make assumptions on conditional heteroskedasticity in the high-dimensional model and, therefore, we do not consider heteroskedasticity as misspecification. By assumption, there is no autocorrelation. The next proposition shows the population properties of the general-to-specific procedure.

Proposition 7 (General-to-specific) *Let $\underline{Z}$ be a subset of $\{X_1, X_2, .., X_K\}$ with at most K_s members. General-to-specific selects the coefficient of X_1 in the linear regression model with minimum $E(V^*(Y|\underline{Z}))$ as the population effect of X_1.*

[6] The general-to-specific approach examined in Hoover and Perez (1999) inspired Hendry and Krolzig (1999) to develop their PC Gets algorithm.

Under assumption (O), general-to-specific does not identify the effect nor whether X_1 has an effect on Y.

Under assumption (S), general-to-specific identifies the effect and, thus, whether X_1 has an effect on Y.

The result is similar to that of BACE and the model selection criteria. The general-to-specific procedure works under the true submodel assumption (S) because it relies on a measure of model fit that identifies the true submodel.

4 New method based on conditional mean independence

In general, when the data is high-dimensional the effect in a high-dimensional model cannot be inferred from low-dimensional models without further identifying assumptions. In Section 2 we provide three different identifying assumptions. We considered existing methods in Section 3, and showed that only a few of them work, and only under the true submodel (S) assumption. None of the methods test the validity of the identifying assumptions. In practice, it is usually not known whether one of the assumptions hold. In this section, we develop a new method that tests the validity of the conditional mean independence assumption (O) and provides a consistent estimator of the effect when the assumption is satisfied.

We consider assumption (O) because it does not impose restrictions on the regression coefficients β_1,..,β_K unlike assumption (S). Thus, assumption (O) has the advantage in practice that it allows the true model to contain more variables than available observations.

To build a method on assumption (O), the method must reveal if the assumption is valid and, in case it is, estimate the effect under the assumption. According to assumption (O) and (3) it is necessary to find a set $A \subset \{X_2, .., X_K\}$ with $(K - K_s)$ members such that the coefficient on X_1 is 0 in the linear regression of any $X_i \in A$ on the variables in A^c; that is, for all $X_i \in A$, the coefficient α_1^i in the linear regression

$$X_i = \alpha_1^i X_1 + \sum_{X_k \in A^c \setminus X_1} \alpha_k^i X_k + u \tag{6}$$

must be 0. A direct implementation would be to make a joint test of $\alpha_1^i = 0$ over all the regressions $X_i \in A$ for all choices of A. Computational aspects lead us to reformulate the problem.

We propose the following method. The problem to find a valid set A is broken into two parts. In the first part we search for a set A, which is likely to satisfy assumption (O) and in the second part we check if A in fact satisfies assumption (O). Searching for the set A is done by investigating which set of variables best explain the variation in X_1. The relationship to (6) follows because $\alpha_1^i = 0$ for all i in (6) is equivalent to $\lambda_i^1 = 0$ in the reverse linear regression (see Appendix)

$$X_1 = \lambda_i^1 X_i + \sum_{X_k \in A^c \backslash X_1} \lambda_k^i X_k + v \qquad (7)$$

for all $X_i \in A$. If a variable explains some of the variation in X_1, it cannot be conditional mean independent of X_1. Hence, to find the set A, we search for the X's which maximize the fit in the linear regression (7) of X_1 on all $X_k \in A^c \backslash X_1$. Since the fit is maximized by including as many variables in $A^c \backslash X_1$ as possible, we will always include $(K_s - 1)$ variables in A^c. Therefore, since the dependent variable is X_1 in all the linear regressions, and all linear regressions have $(K_s - 1)$ explanatory variables, we can compare the fit of the various choices of $A^c \backslash X_1$ using R^2. The choice of A is the one for which the variables in $A^c \backslash X_1$ give the highest R^2.[7]

After finding a candidate set for A, we next test whether X_1 is conditional mean independent of the variables in A conditional on the variables in $A^c \backslash X_1$. We do this by testing for joint significance of the λ_i^1 in (7) over all the $(K - K_s)$ linear regressions with $X_i \in A$ as dependent variable. In each regression, calculate the t-statistic for the hypothesis $\lambda_i^1 = 0$ against $\lambda_i^1 \neq 0$. The maximal absolute t-statistic over the $(K - K_s)$ regressions is then computed. The hypothesis of conditional mean independence is then tested using the maximal absolute t-statistics and a significance level based on the Bonferroni correction.

In case the validity of A cannot be rejected, we proceed to estimate the effect by linearly regressing Y on the variables in A^c. The effect is the estimate of the coefficient on X_1. The next theorem summarizes the main steps in the method, which we denote the CMI method due to its foundation on conditional mean independence. The proposition shows that the method provides a consistent estimator of the effect if assumption (O) is satisfied.

[7] With a fixed number of explanatory variables in all the regressions, maximizing R^2 is the same as e.g. maximizing the F-test statistics of joint significance or maximizing BIC assuming conditional normality of X_1.

Theorem 8 (CMI method) *The CMI method is*

1. *Let* $A \subset \{X_2, .., X_K\}$ *with* $(K - K_s)$ *members. Find the set* $\widehat{A}$ *that maximizes* R^2 *over all linear regressions of* X_1 *on* $A^c \backslash X_1$ *for all possible choices of* A*.*

2. *Let* t_i *be the t-test statistic for testing whether* $\lambda_i^1 = 0$*.* λ_i^1 *is the coefficient on* $X_i \in \widehat{A}$ *in the linear regression of* X_1 *on* X_i *and the variables in* $\widehat{A}^c \backslash X_1$*. Test whether* $\lambda_i^1 = 0$ *for all* i *using the maximal absolute t-statistic*

$$t_{\max} = \max_i \{t_i\}.$$

3. *If 2 leads to non-rejection, estimate the effect of* X_1 *in* (2) *as the coefficient on* X_1 *in the linear low-dimensional regression of* Y *on the variables in* $\widehat{A}^c$*.*

Assume a random sample of $\{Y, X_1, .., X_K\}$*. If assumption* (O) *holds, then the CMI method provides a consistent estimator of the effect of* X_1 *in the high-dimensional model (2).*

Steps 1) and 2) provide a consistent test of assumption (O)*.*

The CMI method consistently rejects assumption (O) when it is false. When either assumption (S) or (I) is true, but not assumption (O), the CMI method consistently indicates that inference cannot be based on assumption (O). Hence, one is not misled by getting an estimate of the effect when, in fact, the assumption for obtaining the estimate is not satisfied. This is a clear contrast to existing methods considered in Section 3. The ones that are consistent under one of the assumptions are only reliable if the assumption is imposed a priori. The CMI method has a built-in test for reliability due to step 2.

The application of the maximal absolute t-statistics in step 2 can be viewed as solving a multiple testing problem. Rejecting the hypothesis if the maximum of the absolute value of all the t-statistics is larger than a critical value is the same as rejecting the hypothesis if one of the individual t-statistics is larger than the same critical value. In choosing a critical value for the maximal absolute t-statistics, conservative or liberal critical values can be used. A conservative critical value can be selected by ignoring the multiple testing problem and using the overall nominal level for each test of zero correlation. A liberal critical value can be based on the Bonferroni correction. Usually, applications of the Bonferroni correction lead to conservative tests, but here the Bonferroni correction is used in a two-step procedure. This implies that the smaller the critical value, the more likely conditional mean independent regressors are rejected. Step 1 can be considered a pretest. Based on Monte Carlo results (see Section 5 below), we have found that using the desired overall nominal level as the nominal level in each of the steps together with the Bonferroni correction in step 2 works satisfactorily.

There exist other tests applicable for step 2 in the CMI method. Instead of using the Bonferroni correction, it is possible to apply the bootstrap to find the critical value for the maximal absolute t-statistics. This is a more computer intensive task. In principle, it is also possible to apply the bootstrap to steps 1 and 2 simultaneously. Our Monte Carlo simulations in Section 5 suggest that the implementation with the Bonferroni correction works well.

The number of variables, K_s, to include in the linear regression must be chosen such that there are at least $(K-N+1)$ regressors which are conditionally mean independent of X_1. In practice, steps 1 and 2 can be repeated for different values of K_s to find a feasible value of K_s. When a feasible K_s is found, it does not influence the identification of β_1 to include extra regressors but it changes the variance of the estimator of β_1. Whether or not including extra regressors increases or decreases this variance depends both on the distribution of the regressors and on the unknown values of the corresponding β's in the high-dimensional model.

The CMI method has similarities with the recent method by Crainiceanu, Dominici and Parmigiani (2008). They develop a method for $n > K$ with the goal of finding the effect of X_1 on Y. They suggest a two-stage method in which a first step finds the best set of predictors for the variable of interest X_1. This step resembles our step 1. Their step 2 is different from ours, as they find good predictors for Y. In the final model they include good predictors for both X_1 and Y. They describe the situation as "adjustment uncertainty in effect estimation" with the idea being that the variables that are correlated with X_1 should be included in order to obtain a good estimator of β_1.

Another recent approach is the Weighted Average Least Squares (WALS) approach by Magnus, Powell and Prüfer (2010). WALS is conceptually close to Bayesian Model Averaging, but builds on different model priors. Its implementation requires $n > K$, as parameter calculation requires the estimated variance of the error term from the high-dimensional linear regression (Magnus et al., 2010:p.145). Thus, it is designed for a different situation than the CMI method which can be carried out when $n < K$.

5 Finite sample properties of the methods

In this section we investigate the finite sample properties of the CMI method presented in Section 4 and compare it with some of the methods considered in Section 3. We report Monte Carlo results on the estimation of the effect and power properties for testing significance of the variable. The designs focus on the conditional mean independence assumption (O) since the new method is based on this assumption.[8] We limit ourselves to report three Monte Carlo designs which illustrate the main effects of facing a high-dimensional dataset.

The Monte Carlo designs have $K = 30$ variables. We consider two sample sizes, $n = 25$ and $n = 50$. The sample with $n = 25$ is undersized. We assume that the low-dimensional linear regressions include $K_s = 3$ variables and an intercept. With the undersized sample, this implies 21 degrees of freedom or about 7 degrees of freedom per parameter in the linear low-dimensional regressions. There is a total of 4,525 combinations of choosing 3 out of 30 variables. The variable of interest is X_1. It is correlated with X_2 and X_3. These three variables are independent of the other 27 variables. This implies that assumption (O) is satisfied with $A = \{X_4, .., X_{30}\}$. Assumptions (S) and (I) are not satisfied. All the variables have zero mean and unit variance. The realizations of $\{X_1, X_2, X_3\}$ are drawn from a multivariate normal distribution with $Corr(X_1, X_2) = Corr(X_1, X_3) = 0.5$ and $Corr(X_2, X_3) = -0.25$. The other variables are independent and identically standard normally distributed. The regressand, Y, is generated by

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + 5X_5 + 4.5X_6 + 1X_7 + .. + 1X_{30} + 5 + U, \qquad (8)$$

where $U \sim N(0, 0.25)$ and U is independent of $X_1, .., X_{30}$. The number of Monte Carlo replications is 1,000.

The three Monte Carlo designs reported below only differ in their values of β_2, β_3, and β_4 in (8). Table 2 shows the values of β_2, β_3, and β_4 that define the designs denoted A, B, and C. To facilitate the interpretation of the Monte Carlo results, Table 2 also reports properties of the designs. The table shows that the best fitting linear regression in terms of minimum mean square linear prediction error, $E(V^*(Y \mid X_i, X_j, X_k))$, depends on the value of β_1, where

$$V^*(Y \mid X_i, X_j, X_k) \equiv E((Y - E^*(Y \mid X_i, X_j, X_k)) \mid X_i, X_j, X_k)^2$$

If there were no sampling uncertainty, the best fitting low-dimensional linear regression determines the properties of many of the methods discussed in Section 3. In particular, the bias of the estimator of β_1 in the best fitting low-dimensional linear regression is important. The biases in the linear regressions are reported in the bottom part of Table 2.

[8] For Monte Carlo studies in which assumption (S) holds, see Jensen (2006, 2010).

Table 2: Properties of the three Monte Carlo designs based on (8).

	Design		
	A	B	C
$\{\beta_2, \beta_3, \beta_4\}$	$\{-4, -4, 3\}$	$\{-2, -2, 4\}$	$\{10, 12, 3\}$
Best fit linear reg.			
$\beta_1 \in$	$(-\infty, 1) \cup (7, \infty)$	$(-2, 6)$	$(-8.7, 8.7)$
Regressors	X_1, X_5, X_6	X_4, X_5, X_6	X_2, X_3, X_5
$\beta_1 \in$	$(1, 7)$	$(-\infty, -2) \cup (6, \infty)$	$(-\infty, -8.7) \cup (8.7, \infty)$
Regressors	X_4, X_5, X_6	X_1, X_5, X_6	X_1, X_2, X_3
$Bias(\beta_1)$ with reg.			
X_1, X_2, X_3	0	0	0
$X_1, X_2, X_{k\geq 4}$	$-3\frac{1}{3}$	$-1\frac{2}{3}$	10
$X_1, X_3, X_{k\geq 4}$	$-3\frac{1}{3}$	$-1\frac{2}{3}$	$8\frac{1}{3}$
$X_1, X_{k\geq 4}, X_{j\geq 4}$	-4	-2	11
$X_{k\geq 2}, X_{j\geq 2}, X_{i\geq 2}$	$-\beta_1$	$-\beta_1$	$-\beta_1$

Note: Best fit. linear reg. is the linear regression ($K_s = 3$) with the lowest $E(V^*(Y \mid X_k, X_i, X_j))$. $Bias(\beta_1)$ with reg. is the bias of the estimator of β_1 in the linear regression. If X_1 is not included in the linear regression, then the estimator of β_1 is 0. All linear regressions include an intercept.

In design A, X_1 is included in the best fitting low-dimensional linear regression along with X_5 and X_6 when X_1 has *no* effect ($\beta_1 = 0$). The variables X_5 and X_6 are included because they have relatively large coefficients. The reason why X_1 is included despite $\beta_1 = 0$ is that X_1 is correlated with X_2 and X_3 in such a way that it provides a better fit than including either X_2 or X_3.

In design B the best fitting linear regression does not include X_1 when $\beta_1 = 0$. When β_1 is sufficiently large, X_1 is included in the best fitting linear regression of Y on X_1, X_5, and X_6. The main difference from design A is that X_1 is not included in the best fitting linear regression when X_1 has no effect.

Design C has the property that the best fitting linear regression is X_1, X_2, X_3 for β_1 sufficiently large. This is the only linear regression that provides an unbiased estimator of β_1.

The EBA, Sala-i-Martin's method and general-to-specific are calculated as described in Section 3 with the exception that the final model selection step in the general-to-specific procedure is done using BIC. For the Bayesian test, we apply a t-test of $\widehat{\gamma}_1^{SDM}$ using the standard error suggested by Sala-i-Martin et al. (2004).[9] In the implementation of the CMI method, step 2 is not used as a stopping rule. Instead we choose the linear regression that minimizes the partial correlation between X_1 and the excluded variables.

We have also investigated how the maximal absolute t-statistic in Step 2 of the CMI method performs in terms of type I error. When the null of conditional mean independence is true, the maximal absolute t-statistic has a rejection probability of 0.043 when the Bonferroni correction is used to bound the size of the test to 0.05. Thus, the Bonferroni correction leads to a rejection probability close to the nominal level. The test also has power for $n = 25$ when there is one regressor extra correlated with X_1. For example, when the correlation coefficient is 0.40, the rejection probability is 0.725 and almost one when the correlation coefficient is above 0.50. Relatively low power is seen when the correlation coefficient is 0.25 with a rejection probability of 0.157. The power is, as expected, increasing in sample size. The maximal absolute t-statistic has also been shown to perform well when many more variables are partially correlated with X_1. The simulations reported in Jensen (2010) show that it has good power in many of these cases.

5.1 Estimating the effect

We first investigate the properties of the methods in estimating the effect, β_1, of X_1. The comparisons between methods are made in terms of bias and standard deviation of the estimators of β_1. The bias for each linear regression is known beforehand, see Table 2. In different samples, however, the methods either select different linear regressions or a combination of linear regressions, and the estimators are therefore pretest estimators.

The biases for various estimators of the effect, β_1, in design A with $n = 25$ are shown in Table 3. The biases are shown as a function of β_1. The CMI method has a low bias compared to the other methods. The bias in Sala-i-Martin's method is constant for different values of β_1. The reason is that Sala-i-Martin's method includes X_1 in all the linear regressions and only β_1 varies. The bias in the BACE method varies with the value of β_1. The bias is large and negative, and substantially larger (in absolute terms) at e.g. $\beta_1 = 6$ than at $\beta_1 = 2$. In the absence of sampling uncertainty, the bias of the estimator is $-\beta_1$ for $\beta_1 \in (1, 7)$, see Table 2. General-to-specific, AIC and BIC have properties similar to BACE when there is no sampling uncertainty, see Section 3. With sampling uncertainty, however, there are differences due to the fact that BACE combines all linear

[9] In our simulation study, we include all models when calculating $\widehat{\gamma}_1^{SDM}$. Thus, we abstain from making any approximations. In practice, we note that models with low weights are given zero weights by some researchers, e.g. Raftery et al. (1997).

regressions whereas the other three methods select only one linear regression.

Table 3: Bias and standard deviation of estimator of β_1 in design A with $n = 25$.

		β_1					
		0	2	4	6	8	10
Sala-i-Martin	Bias	-3.78	-3.78	-3.78	-3.78	-3.78	-3.78
	Std	1.70	1.70	1.70	1.70	1.70	1.70
BACE	Bias	-1.39	-2.25	-3.97	-5.56	-5.93	-5.10
	Std	1.78	0.71	0.36	1.07	2.24	2.70
BIC	Bias	-1.65	-2.33	-3.95	-5.40	-5.38	-4.45
	Std	2.42	1.11	0.69	1.63	2.82	2.82
AIC	Bias	-1.65	-2.33	-3.95	-5.40	-5.38	-4.45
	Std	2.42	1.11	0.69	1.63	2.82	2.82
GSP	Bias	-1.64	-2.32	-3.95	-5.40	-5.39	-4.45
	Std	2.43	1.11	0.69	1.63	2.82	2.82
CMI method	Bias	0.07	0.07	0.07	0.07	0.07	0.07
	Std	3.52	3.52	3.52	3.52	3.52	3.52
Benchmark	Bias	0.13	0.13	0.13	0.13	0.13	0.13
	Std	3.47	3.47	3.47	3.47	3.47	3.47

Notes: Sala-i-Martin = Sala-i-Martin's method; AIC=Akaike Information Criterion, GSP=general-to-specific; Std=standard deviation; Other abbreviations are self-explanatory.

The properties of all the methods can be compared to the only linear regression which provides an unbiased estimator of β_1. This is the linear regression of Y on X_1, X_2, and X_3. This regression is denoted the benchmark regression. In practice, it is not known if a regression providing an unbiased estimator exists. The CMI method selects this regression if it exists. The CMI method and the benchmark regression have about the same bias. The bias in the benchmark regression is solely due to Monte Carlo sampling error. The reason for the similarity of the benchmark regression and the CMI method is that the CMI method selects the benchmark model with probability around 0.95.[10]

The effect of increasing the sample size from $n = 25$ to $n = 50$ is seen by comparing

[10] While our main focus is on reducing bias, it is of interest to note that the estimator of β_1 in the CMI-method has a higher standard deviation than some of the other estimators. This is mainly a result of the other methods often selecting low-dmensional linear regressions where X_1 is not included. This lowers the variation of the estimator of β_1. It is worth noting that if one adopts a mean square error loss criterion, then BACE, BIC, AIC and general-to-specific have a lower mean square error loss than the CMI-method (and the benchmark regression). This result is reversed as the sample size grows since the mean square error loss of the CMI-method approaches 0, whereas for the other methods the mean square error loss converges to the bias squared.

Table 3 with Table 4. Table 4 shows the results for design A with $n = 50$. The standard deviation decreases with the larger sample size for all methods. For the CMI method, the bias also decreases. Not all of the biases of the other methods, however, decrease. For example, for BACE and $\beta_1 = 0$ the bias changes from -1.39 to -2.19. This is consistent with the results in Section 3 and Table 1, which show that the bias eventually (for $n \to \infty$) approaches -4. The larger sampling uncertainty for $n = 25$ compared to $n = 50$ reduces the bias because BACE puts lower probability on linear regressions with X_1 and they induce bias when $\beta_1 = 0$.

Table 4: Bias and standard deviation of estimator of β_1 in design A with $n = 50$.

		β_1					
		0	2	4	6	8	10
Sala-i-Martin	Bias	-3.97	-3.97	-3.97	-3.97	-3.97	-3.97
	Std	0.99	0.99	0.99	0.99	0.99	0.99
BACE	Bias	-2.19	-2.24	-4.00	-5.68	-5.04	-4.10
	Std	1.89	0.67	0.05	0.74	1.84	1.26
BIC	Bias	-2.41	-2.25	-4.00	-5.63	-4.77	-4.05
	Std	2.28	0.93	0.10	1.04	2.00	1.22
AIC	Bias	-2.41	-2.25	-4.00	-5.63	-4.77	-4.05
	Std	2.28	0.93	0.10	1.04	2.00	1.22
GSP	Bias	-2.41	-2.25	-4.00	-5.63	-4.77	-4.05
	Std	2.28	0.93	0.10	1.04	2.00	1.22
CMI method	Bias	0.08	0.08	0.08	0.08	0.08	0.08
	Std	2.25	2.25	2.25	2.25	2.25	2.25
Benchmark	Bias	0.08	0.08	0.08	0.08	0.08	0.08
	Std	2.25	2.25	2.25	2.25	2.25	2.25

Notes: See notes to Table 3.

With a sample size of $n = 50$, the high-dimensional linear regression is possible. This linear regression identifies β_1. The variance of the estimator of β_1 in the high-dimensional linear regression cannot be uniformly ranked against the variance of the estimator of β_1 in the benchmark regression. There are two extremes which do not depend on the distribution of the regressors. If the β's of the excluded variables in the Benchmark regression are 0, then the variance is lower in the Benchmark regression. Conversely, if the β's of the excluded variables are sufficiently large, then the variance is lower in the high-dimensional linear regression. The Benchmark regression is not known in practice, but the properties of the CMI method are similar. This means that it is not possible to say whether it is better to run the high-dimensional linear regression compared to the CMI method when the sample is not high-dimensional. Asymptotically, they are equivalent.

Table 5 presents the results for design B. The bias with the CMI method is low and about the same as for the benchmark regression. When $\beta_1 = 0$, Sala-i-Martin's method has the highest bias. This is because Sala-i-Martin's method assigns most weight to the best fitting linear regression with X_1, which results in a bias equal to -2. According to Table 1, if there were no sampling uncertainty, then BACE, general-to-specific, AIC, and BIC would not choose a linear regression with X_1 when $\beta_1 = 0$. In finite samples, this is reflected in a lower bias of the latter methods compared to Sala-i-Martin's method when $\beta_1 = 0$.

Table 5: Bias and standard deviation of estimator of β_1 in design B with $n = 25$.

		β_1					
		0	2	4	6	8	10
Sala-i-Martin	Bias	-1.91	-1.91	-1.91	-1.91	-1.91	-1.91
	Std	1.60	1.60	1.60	1.60	1.60	1.60
BACE	Bias	-0.27	-1.99	-3.65	-4.12	-3.30	-2.44
	Std	0.74	0.22	0.90	2.14	2.65	2.31
AIC	Bias	-0.33	-2.00	-3.60	-3.68	-2.75	-2.12
	Std	1.15	0.31	1.31	2.68	2.82	2.20
BIC	Bias	-0.33	-2.00	-3.60	-3.68	-2.75	-2.12
	Std	1.15	0.31	1.31	2.68	2.82	2.20
GSP	Bias	-0.33	-2.00	-3.60	-3.68	-2.75	-2.11
	Std	1.15	0.31	1.31	2.69	2.82	2.21
CMI method	Bias	0.15	0.15	0.15	0.15	0.15	0.15
	Std	3.64	3.64	3.64	3.64	3.64	3.64
Benchmark	Bias	0.15	0.15	0.15	0.15	0.15	0.15
	Std	3.62	3.62	3.62	3.62	3.62	3.62

Notes: See notes to Table 3.

Results for design C are reported in Table 6. In this design the benchmark regression is the best fitting linear regression if there is no sampling uncertainty and β_1 is sufficiently large. This, however, is not obvious for sample size $n = 25$. The reason is that the benchmark regression is rarely selected by the methods based on model fit. For example, general-to-specific selects the benchmark regression in just 16.5% of the samples generated when $\beta_1 = 10$. In contrast, the bias of the CMI method is low compared to the other methods.

Table 6: Bias and standard deviation of estimator of β_1 in design C with $n = 25$.

		β_1					
		0	2	4	6	8	10
Sala-i-Martin	Bias	7.05	7.05	7.05	7.05	7.05	7.05
	Std	4.99	4.99	4.99	4.99	4.99	4.99
BACE	Bias	1.64	0.50	-0.09	0.02	0.92	2.36
	Std	3.13	4.16	5.41	6.66	7.65	8.06
BIC	Bias	0.95	-0.44	-1.28	-1.55	-0.63	1.13
	Std	3.23	4.32	5.74	7.29	8.65	9.34
AIC	Bias	0.95	-0.44	-1.28	-1.55	-0.63	1.13
	Std	3.23	4.32	5.74	7.29	8.65	9.34
GSP	Bias	0.94	-0.43	-1.26	-1.54	-0.61	1.15
	Std	3.22	4.33	5.76	7.30	8.66	9.35
CMI method	Bias	0.52	0.52	0.52	0.52	0.52	0.52
	Std	3.86	3.86	3.86	3.86	3.86	3.86
Benchmark	Bias	0.14	0.14	0.14	0.14	0.14	0.14
	Std	3.47	3.47	3.47	3.47	3.47	3.47

Notes: See notes to Table 3.

The results in Tables 3-6 also point to the relevance of propositions 3-7 and Theorem 8. The bias is substantial for Sala-i-Martin's method, BACE, BIC, AIC, and general-to-specific. This is in line with the result in the propositions that the methods do not identify β_1 under assumption (O). It should also be noted that Crainiceanu et al. (2008) show that when the goal is to estimate an effect of a variable, Bayesian Model Averaging may be biased. Our Monte Carlo study shows that this may also happen for the BACE version of Bayesian Model Averaging when assumption (O) holds.

5.2 Test of no effect $(\beta_1 = 0)$

In this subsection, we investigate the ability of the different methods to determine whether $\beta_1 = 0$. As noted earlier, this is similar in spirit to the question addressed in the literature on sensitivity analysis and robustness of a variable, namely whether a variable is robust or not. We do not show the results for AIC and BIC because they are similar to the results for general-to-specific, and we do not show the results for the benchmark regression because they are similar to those for the CMI method.

The control of the Type I error for determining if X_1 has an effect is shown in Figure

1 for design A with $n = 25$. The figure shows the true value of β_1 on the first axis and the rejection probabilities of testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ on the second axis. The nominal significance level is 0.05 (marked with a horizontal line). The CMI method has a probability of a Type I error close to the nominal level. The EBA has a low probability of a Type I error whereas the probability for the other methods is substantially above the nominal level. For example, general-to-specific and Sala-i-Martin's method have probabilities of Type I errors of about 0.34 and 0.66, respectively. In the terminology of sensitivity analysis, these methods accept well above the nominal significance level that X_1 is "robust" when, in fact, it is not.

Figure 1 also shows the power functions. The power of the CMI method is monotonically rising for β_1 values further away from 0. The powers of the other methods, however, decrease as β_1 moves from 0 to 4. That is, as the effect of X_1 becomes larger, the less likely is is that the other methods will accept that X_1 has an effect. For β_1 close to 4, EBA, BACE and general-to-specific have powers close to 0. A reason may be found in Table 1, which shows that the best fitting linear regression is Y on X_4, X_5, and X_6 for $\beta_1 \in (1, 7)$. Hence, it is likely that these methods often select the linear regression without X_1 and consequently conclude that X_1 has no effect.

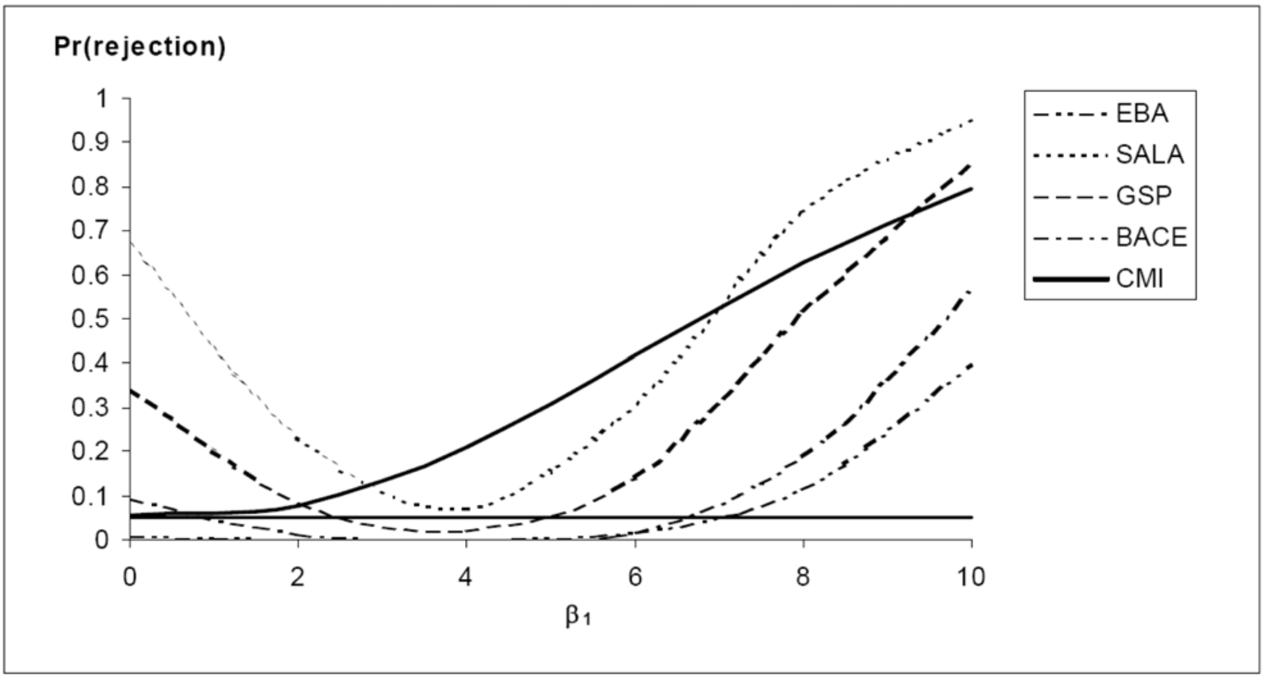

Figure 1. Power of testing $\beta_1 = 0$ against $\beta_1 \neq 0$ using a 0.05 nominal significance level in design A with $n = 25$.

Notes: SALA =Sala-i-Martin's method, GSP = General-to-Specific., CMI=CMI method

Figure 2 shows the powers for design A with $n = 50$. Compared to Figure 1 it is seen that the control of the type I error worsens for many of the methods. The reason is that it becomes more likely that the methods based on best fit select the linear regression Y on X_1, X_5, and X_6 when X_1 has no effect. The estimator of β_1 in this linear regression has a bias equal to -4 and thus the test indicates that X_1 has an effect. The power of the CMI method increases with the sample size. Since the sample is not high-dimensional when $n = 50$, a two-sided t-test in the high-dimensional linear regression is feasible. In this regression, this test is an invariant uniformly most powerful test. This does not imply, however, that the test is more powerful than the two-sided t-test performed on the linear regression found using the CMI method. The reason is similar to the one discussed in subsection 5.1 regarding the ranking of the efficiency of the estimators of β_1 in the benchmark regression versus the high-dimensional linear regression. The ranking depends on the distribution of the regressors and the values of those β's in the high-dimensional linear regression that are excluded from the benchmark regression.

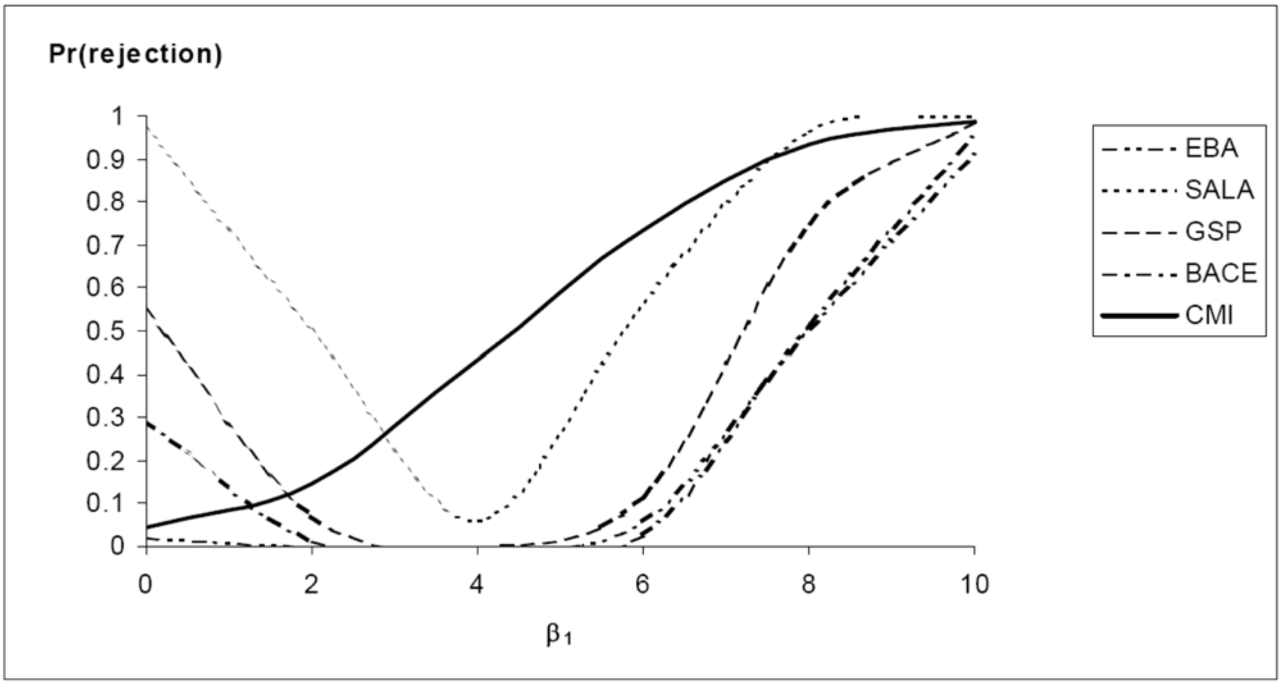

Figure 2. Power of testing $\beta_1 = 0$ against $\beta_1 \neq 0$ using a 0.05 nominal significance level in design A with $n = 50$.

Notes: SALA =Sala-i-Martin's method, GSP = General-to-Specific.

Figure 3 shows the power functions for design B. Contrary to design A, when X_1 has no effect, the best fitting linear regression does not include X_1, see Table 1. This explains why AIC, BIC, general-to-specific and BACE methods control the Type I error much better than in design A. Their powers, however, still decrease as β_1 increases. The powers only increase for $\beta_1 > 2$. As can be seen in Table 2, the reason is that X_1 is only included in the best fitting population linear regression when $\beta_1 \geq 6$. The CMI method controls the Type I error and has monotonically rising power in β_1.

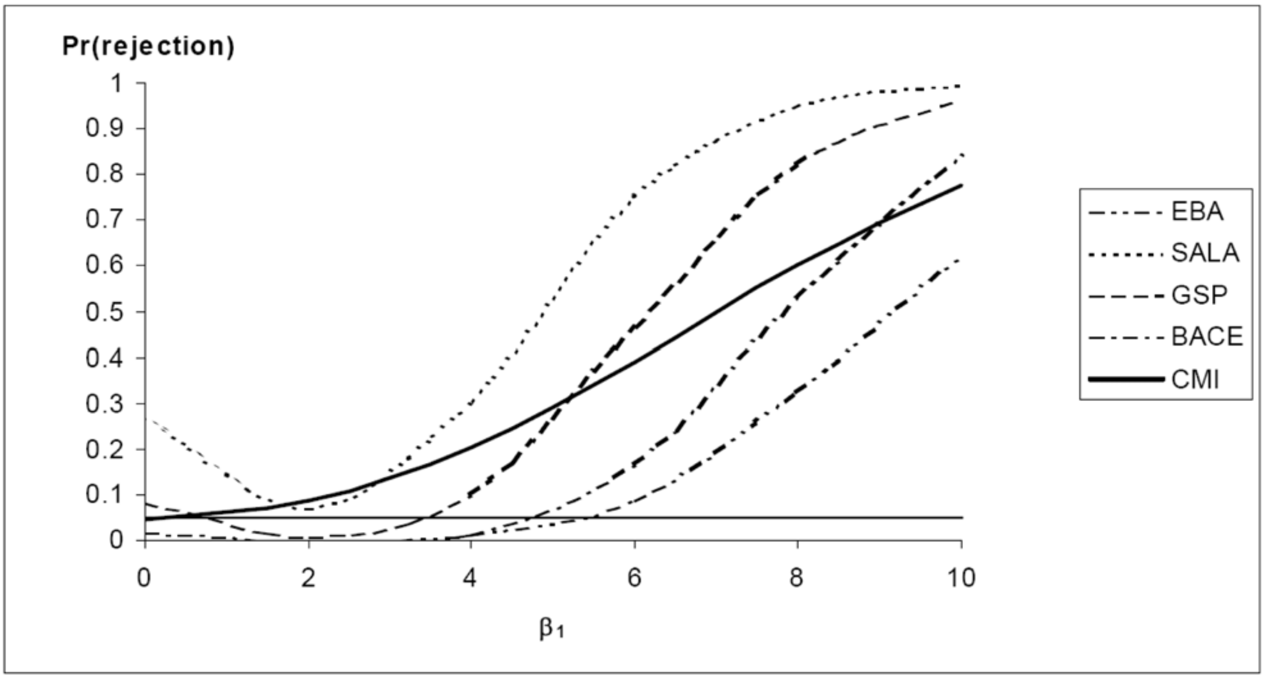

Figure 3. Power of testing $\beta_1 = 0$ against $\beta_1 \neq 0$ using a 0.05 nominal significance level in design B with $n = 25$.

Notes: SALA =Sala-i-Martin's method, GSP = General-to-Specific.

The power results for design C are shown in Figure 4. Contrary to designs A and B, all methods have monotonically increasing powers in β_1. An explanation can be found in the fact that the best fitting population linear regression only includes X_1 when $\beta_1 \neq 0$ and this linear regression provides an unbiased estimator of β_1. The power of general-to-specific (and AIC and BIC) and BACE is below the power of the CMI method. Only EBA has a power as high as the CMI method. A reason why EBA is performing well in this particular case is that the bias is positive in all linear regressions which include X_1 and this lowers the probability of getting both positive and negative estimates of β in these linear regressions. The power for negative values of β_1, which we calculated but do not show, is low.

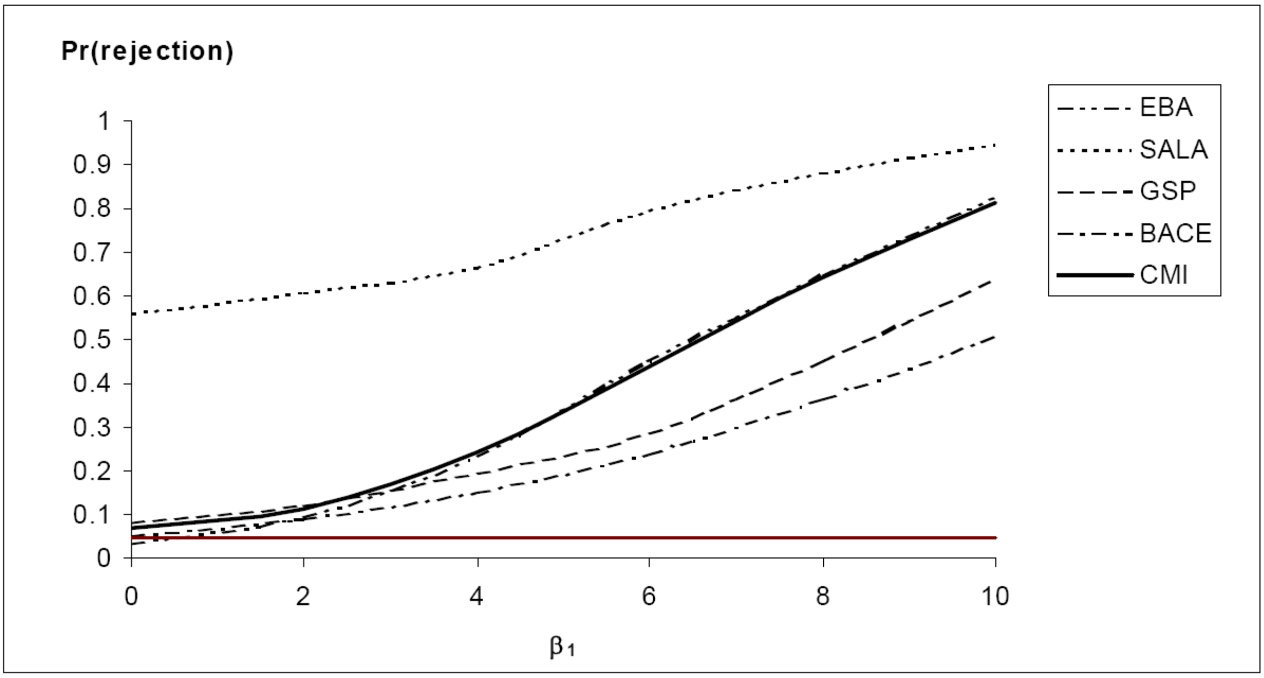

Figure 4. Power of testing $\beta_1 = 0$ against $\beta_1 \neq 0$ using a 0.05 nominal significance level in design C with $n = 25$.

Notes: SALA =Sala-i-Martin's method, GSP = General-to-Specific.

Figures 1-4 also point to both the relevance and limitations of the results in Section 2. For example, Sala-i-Martin's method always has a high type I error. We noted that a variable may be chosen even if it has no effect under assumption (O). The asymptotic results shows that this will happen with probability 1, but in the finite samples it varies from around 0.3 to almost 1. For the BACE method the asymptotic results say that X_1 will always be selected when it is in the best fitting model even with $\beta_1 = 0$. In design A, this only happens with a probability equal to 0.10 with $n = 25$ and increases to about 0.30 with $n = 50$. The asymptotic result therefore overstates the selection probability, but is still right in suggesting that X_1 will be chosen often, even when it is irrelevant.

It should be noted that some of the methods have also been investigated by Hoover and Perez (2004). They consider EBA, Sala-i-Martin's method, and general-to-specific. They perform a Monte Carlo study for the case in which the true model can be estimated with $K_S < K < n$. First, they find that the EBA has low power. This is also the case in our simulations, especially for low values of β_1. Further, they find that Sala-i-Martin's method often selects variables that are excluded from the true model. This result is in line with Proposition 3 which shows that Sala-i-Martin's method could label a variable robust even when $\beta_1 = 0$ and assumption (S) is true. Our simulations show that this is also true for assumption (O). Finally, they show that general-to-specific works well in their setup. This is in line with proposition 7, which shows that when assumption (S) is true, general-to-specific should work well. Under assumption (O), this is no longer the case.

6 Application

In this section, we apply the CMI method described in Section 4 to the Sala-i-Martin (1997) dataset mentioned in the introduction. In addition to the growth rate of GDP per capita for the period 1960-92, the dataset contains 61 potential growth determinants, but the data are complete for only 36 countries. Thus the high-dimensional linear regression cannot be carried out. Sala-i-Martin (1997) applied the method described in subsection 3.3 to these 61 variables,[11] and found that 22 were "robust" by his criterion. As shown in this paper, Sala-i-Martin's method cannot determine whether a variable has an effect correctly under any of the assumptions (O) or (S).

Fernandez et al. (2001) also used Sala-i-Martin's dataset, but restricted the dataset such that the number of observations was 72 and that only 41 out 61 regressors were included in the analysis. This is the approach often taken in the literature. Restricting the set of regressor implies that the authors assume that the true model includes 41 variables at most. The advantage of this approach is that the regression with these variables can be carried out. The disadvantage is that 20 variables are excluded from the analysis and assumed to have zero coefficients at the outset due to limited data. Excluding variables is a problem as it potentially leads to bias. We note that if variables nevertheless are dropped a priori, this suggests that one believes that they are unimportant and that the true model contains a set of variables from the included ones.

Hoover and Perez (2004) make the assumption that all variables can be transformed into variables following a normal distribution and draw from these distributions to assign values for missing observations to get a dataset that is no longer high-dimensional. Thus, multiple imputation allows the high-dimensional linear regression to be carried out. While this approach is intuitively appealing, it is problematic. The reason is that the imputation procedure induces measurement error in the regressors and bias in regression coefficients (Hendry and Krolzig, 2004:p.806). The exact bias for the coefficient on a regressor in a

[11] Panel data would seem a possible solution, but the difficulty is that time series data **are** not available for all 61 variables, leaving us with the cross section only.

bivariate linear regression is calculated by Little and Rubin (2002:p.65-66). Their calculation also affects the correlation between imputed variables. Jensen (2010) shows that F-tests of exclusion restrictions on the imputed data have size distortions in a Monte Carlo study. If data is imputed and the measurement error ignored, it becomes plausible to assume that the true model is in the set.

While a respectable case can be made for the two previous options, we have decided on a more 'puritanical' approach and only use the data available. Yet we acknowledge that losing many countries is not unproblematic, as data absences in a context like economic growth will rarely be random. Nevertheless, we notice that dropping countries is not unusual in the literature, as the Fernandez et al. (2001) application illustrates. In the original Sala-i-Martin dataset, there are growth rates for 120 countries, meaning that 48 countries are dropped.

We carried out the CMI method for all 61 regressors. We carried out the search for regressors in three different ways. First, we used $K_s = 7$ and searched for six variables in the first step. Second, following Sala-i-Martin (1997) we included three fixed regressors[12] and searched for three additional regressors in the first step. The regressors found in the first step are candidate sets for the set A^C in assumption (O). The search strategy implies that we have run more than 50 million regressions for each of the 61 regressors.

For each of the 61 regressors, we carry out step 2 of the CMI method based on the absolute maximal t-statistic. Both relatively conservative and liberal critical values are investigated. The conservative critical value is selected by ignoring the multiple testing problem, and using a 5% nominal level for each test of zero correlation. With this critical value there are not sufficiently many regressors that are conditional mean independent of the variable of interest.[13] The liberal critical value is based on the Bonferroni bound.[14] Given the model with $K_s = 7$, the null is rejected whenever the p-value associated with the test is less than $0.05/54$ corresponding to a critical value of about 3.67 in absolute value.[15] The result is that based on the Bonferroni bound based critical value, assumption (O) is not satisfied. We therefore conclude that inference cannot be made on the coefficients on the 61 regressors using the CMI method.

[12] These three regressors are the log of GDP per capita in 1960, the primary school enrollment rate in 1960 and the life expectancy in 1960.

[13] Using the regular 5% cut-off corresponding to a critical value of 1.97.

[14] Note, usually the Bonferroni bound is conservative, but here the Bonferroni bound is used in a two-step procedure; the smaller the critical value, the less likely that conditional mean independence is found to be satisfied.

[15] Under the null there are 54 possible tests in which rejection could happen erroneously. Using a critical value of $0.05/54$ puts a bound on the type I error.

Though it is difficult to make inference on *individual* coefficients, Jensen (2010) shows that the hypothesis that all 61 variables have zero coefficients is rejected by the maximal absolute t-statistic test. Thus, at least one of the regressors has an effect on growth. Applying the BACE method, he finds that a candidate for the true submodel contains "fraction with Confucian religion". However, a test rejects that this is the only regressor with a non-zero coefficient. Thus, our findings complement those of Jensen (2010) who finds that making inference on the coefficients on individual regressors in the long growth regression is difficult using assumption (S). Our results are negative in the sense that they suggest that drawing a sound conclusion about the effect of a particular growth regressor is very difficult with a dataset with more variables than observations.

Our theoretical and empirical results, complemented by those of Jensen (2010), cast some doubts on the validity of the results obtained in the other papers using the Sala-i-Martin dataset, even though many authors make strong claims about which variables are determinants of growth, see e.g. section 5.1 in Durlauf et al. (2005) for a summary of findings across studies.

7 Discussion and conclusion

In this paper we have examined methods known from the empirical growth literature. We have analyzed them under two assumptions that make it possible to identify the effect of a variable from a linear regression. If any one of the conditions holds, there will be no omitted variable bias in the coefficient on the variable of interest when the right linear regression is performed. The majority of these methods are based on a measure of model fit and many of the methods work only under assumption (S).[16] Importantly, Sala-i-Martin's method and the EBA do not work under any of the assumptions considered, when the task is to determine whether a variable has any effect.

We have derived population properties of the methods. Our results are comparable with those of Ericsson (2008) who discusses the relevance of the encompassing principle for robustness analysis. For example, EBA focuses on how coefficient estimates change as the conditioning set alters. Our proposition 2 and Ericsson's result show that this approach may lead to the wrong results. The basic problem is that information may wrongly be excluded. Ericsson points to encompassing tests as the basic remedy. They test whether

[16] It is worth emphasizing that our results also hold when there are more observations than variables. Whether the fit based methods, such as BACE and general-to-specific, work is determined by the number of regressors included in the low-dimensional regressions. If this number is less than the number of variables in the high-dimensional regression, the same properties regarding biases and powers result.

information is validly excluded. In this sense, they are tests of robustness. When the high-dimensional model can be estimated, the usual encompassing test is the F-test applied in general-to-specific modeling (Ericsson, 2008:p.906). This test is not available with $n < K$. In this case, tests like those described in Section 4 and in the companion paper by Jensen (2010) are available to test whether variables are validly excluded.

Our CMI method includes step 2 which tests whether variables can be excluded when the goal of the analysis is to learn about β_1. Step 2 implies that the CMI method has some affinity to the encompassing principle. The CMI method, however, searches for valid restrictions on the linear regression with X_1 as the dependent variable, and not for restrictions on (2). Model averaging may be viewed as an alternative approach to account for excluded information. Like Ericsson (2008), we prefer an approach that tests assumptions.

The paper also points to assumption (O) as an alternative to assumption (S) in the linear regression context. The basic CMI method tells us to search for variables that make X_1 conditional mean independent of the remaining. Future research should improve the CMI method by incorporating more advanced methods for selecting variables in high-dimensional regression models, see e.g. Huang, Horowitz and Wei (2010).[17]

Finally, our theoretical and empirical results suggest being cautious in drawing conclusions about the effect of a variable when the dataset is high-dimensional.

8 Appendix

Proof of Proposition 1 (EBA). Let $\mathcal{F}$ be the set of all linear regressions with X_1 and at most $(K_s - 1)$ other variables. In the sample, X_1 is robust if the estimates of the coefficient, γ_1^i, to X_1 in all the linear regressions, i, are significant and have the same sign. In the population without sampling uncertainty, the extreme bounds for the partial effect of X_1 are $\left[\min\limits_{i\in\mathcal{F}} \gamma_1^i \ , \ \max\limits_{i\in\mathcal{F}} \gamma_1^i\right]$.[18]

To prove when an assumption (in this and the following proofs) is not sufficient for identification, it suffices to consider the following example with four regressors. Let the

[17] Their approach also allows for non-linear effects of individual variables which could be important in a growth context. Our focus on linear models has been dictated by the methods analyzed.

[18] In terms of the t-statistics and asymptotics, the decision rule can be determined the following way. The t-statistics used for testing $\gamma_1^{[1k]} = 0$ is given by $\widehat{t}_1^{[1k]} = \widehat{\gamma}_1^{[1k]}/\sqrt{V(\widehat{\gamma}_1^{[1k]})}$, where ^indicates the estimator, for instance the OLS estimator. The probability limit of the t-statistics is degenerate at $+\infty$ or $-\infty$ when $\gamma_1^{[1k]}$ is positive or negative, respectively (consistency of t-test). For $\gamma_1^{[1k]} = 0$ the distribution of the t-statistics is $N(0,1)$ under regularity conditions. When the sample size approaches ∞, the significance probability should approach 0 and, thus, the probability of accepting approaches 1.

high-dimensional linear regression be $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + U$ as defined by (1) and suppose $K_s = 2$. Let $\gamma_1^{[1k]}$ be the coefficient on X_1 in the linear regression of Y on X_1 and X_k, and $\gamma_1^{[10]}$ the coefficient on X_1 in the regression of Y on X_1. Then

$$\begin{aligned}
\gamma_1^{[12]} &= \beta_1 + \frac{\rho_{13} - \rho_{12}\rho_{23}}{1-\rho_{12}^2}\beta_3 + \frac{\rho_{14} - \rho_{12}\rho_{24}}{1-\rho_{12}^2}\beta_4, \\
\gamma_1^{[13]} &= \beta_1 + \frac{\rho_{12} - \rho_{13}\rho_{23}}{1-\rho_{13}^2}\beta_2 + \frac{\rho_{14} - \rho_{13}\rho_{34}}{1-\rho_{13}^2}\beta_4, \\
\gamma_1^{[14]} &= \beta_1 + \frac{\rho_{12} - \rho_{14}\rho_{24}}{1-\rho_{14}^2}\beta_2 + \frac{\rho_{13} - \rho_{14}\rho_{34}}{1-\rho_{14}^2}\beta_3, \\
\gamma_1^{[10]} &= \beta_1 + \rho_{12}\beta_2 + \rho_{13}\beta_3 + \rho_{14}\beta_4.
\end{aligned} \tag{9}$$

where it is assumed that $E(X_k) = 0$, $V(X_k) = 1$ and $Corr(X_k, X_m) = \rho_{km}$, $m \neq k$.

Under assumption (O), the linear regression with the conditionally mean independent regressors excluded provides an unbiased estimator of β_1. This implies that $\beta_1 \in \left[\min_{i\in\mathcal{F}} \gamma_1^i\ ,\ \max_{i\in\mathcal{F}} \gamma_1^i\right]$. Whether X_1 has an effect, however, cannot be determined correctly. This can be seen from the example (9). Suppose $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$. Then assumption (O) is satisfied. The extreme bounds are:

$$[\min(\beta_1, \beta_1 + \rho_{12}\beta_2), \max(\beta_1, \beta_1 + \rho_{12}\beta_2)]$$

If $\beta_1 > 0$ (and, thus, has an effect) and $\beta_1 < -\rho_{12}\beta_2$, then the extreme bounds contain 0 and the lower boundary is negative and the upper bound positive. Then X_1 is not robust and, thus, X_1 is incorrectly labeled as not having an effect on Y.

Under assumption (S), the extreme bounds may not contain β_1. This can be seen using the example with four regressors from above. Suppose $\beta_1 = \beta_2 = 0$ (and, thus, X_1 has no effect) and $\rho_{12} = \rho_{13} = 0$. If $\rho_{14} > 0$, $\rho_{34} < 0$, $\beta_3, \beta_4 > 0$, then the lower bound of the extreme bounds is positive. Hence, β_1 is not in the interval, and X_1 is denoted robust when it has no effect on Y. ■

Proof: Proposition 2 (Minimum t-statistic over models test).

The test will accept that X_1 has an effect if none of the coefficients γ_1^i to X_1 in the linear regressions equal 0. The test accepts that X_1 is unimportant if at least one coefficient on X_1 in a linear regression equals 0.

Under assumption (O), the linear regression, j, with all the conditionally mean independent regressors excluded gives $\gamma_1^j = \beta_1$. Therefore, the test is correct when X_1 is unimportant because $\gamma_1^j = 0$. If $\beta_1 \neq 0$, then $\gamma_1^i \neq 0$ except when an omitted variable bias exactly cancels the effect of β_1. Hence, the test cannot correctly determine when X_1 has an effect. This can also be seen in the four regressor example, (9), in the proof of proposition 1 with $\beta_1 \neq 0$, $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$ and the other parameters being non-zero. If $\beta_1 = -\rho_{12}\beta_2$, then $\gamma_1 = 0$ in the regression of Y on X_1.

Under assumption (S), the four regressor example (9) can be used to show that the test is not consistent. Suppose the true submodel is Y on X_2 and X_3. Since the test always includes X_1 as a regressor, the regression of Y on X_1, X_2 and X_3 is not performed. Hence, the coefficient on X_1 in the linear regressions may be biased. Note, if the true model has fewer than K_s variables, then the test is correct when X_1 has no effect on Y. When X_1 has an effect, the test may imply that X_1 is denoted not robust if an omitted variable bias cancels the effect of β_1 in the same manner as under assumption (O). ■

Proof of Proposition 3 (Sala-i-Martin's method). The robustness of X_1 is determined by $CDF(0) = \sum_{i=1}^{m} w_i CDF_i(0)$ being above or below $1-\alpha$, where α resembles a significance level. Sala-i-Martin's method does not have an obvious analogue in the population, and therefore the population version is derived as a probability limit for $n \to \infty$ keeping K_s and K fixed.

Consider first $CDF_i(0) = Max\left(\Phi(\widehat{\gamma}_1^i/\widehat{\sigma}_{\widehat{\gamma}_1^i}), 1-\Phi(\widehat{\gamma}_1^i/\widehat{\sigma}_{\widehat{\gamma}_1^i})\right)$. In the population, $\widehat{\gamma}_1^i$ is replaced by γ_1^i and there is no uncertainty. If $\gamma_1^i \neq 0$, then $CDF_i(0) = 1$. If $\gamma_1^i = 0$, then both the numerator and the denominator equal 0. Under suitable regularity conditions $\widehat{\gamma}_1^i/\widehat{\sigma}_{\widehat{\gamma}_1^i} \to^p Z$, $Z \sim N(0,1)$. Since $\Phi(Z) \sim U$, $U \sim Uniform[0,1]$,

$$P(CDF_i(0) < a \mid \gamma_1^i = 0) = P(Max(U, 1-U) < a) = 2a - 1, \ 0.5 \leq a \leq 1. \tag{10}$$

Therefore, if $\gamma_1^i = 0$, then the test accepts that $\gamma_1^i = 0$.

Secondly, the weight can be rewritten as

$$w_j = \frac{SSE_j^{-n/2}}{\sum_{i=1}^{m} SSE_i^{-n/2}} = \frac{1}{\sum_{i=1}^{m} \left(\frac{\frac{1}{n}SSE_i}{\frac{1}{n}SSE_j}\right)^{-\frac{n}{2}}}, \tag{11}$$

where SSE_j is the sum of squared residuals in regression j. Let $\underline{Z}$ be a subset of $\{X_2, .., X_K\}$ with at most $(K_s - 1)$ members and γ_Z^i the corresponding parameter vector in linear regression i. Then

$$\frac{1}{n}SSE_i \to^p E_{x_1,\underline{Z}}(V^*(Y \mid X_1, \underline{Z})) \equiv \sigma_i^2$$

under suitable regularity conditions, where $V^*(Y \mid X_1, \underline{Z}) = V((Y - E^*(Y \mid X_1, \underline{Z})) \mid X_1, \underline{Z})$

The convergence of the terms $\left(\frac{1}{n}SSE_j / \frac{1}{n}SSE_i\right)^{\frac{n}{2}}$ depends on the probability limits of the numerator and denominator:

$$\left(\frac{\frac{1}{n}SSE_j}{\frac{1}{n}SSE_i}\right)^{\frac{n}{2}} \rightarrow^p \begin{cases} \infty & \text{if } \sigma_j > \sigma_i \\ 0 & \text{if } \sigma_j < \sigma_i \\ W & \text{if } \sigma_j = \sigma_i \end{cases},$$

where W is a random variable with support on the unit interval.

The weight in the population for regression j is

$$w_j = \underset{N\rightarrow\infty}{plim} \frac{1}{1 + \sum_{i \neq j} \left(\frac{\frac{1}{N}SSE_j}{\frac{1}{N}SSE_i}\right)^{\frac{N}{2}}}.$$

If $\sigma_j^2 < \sigma_i^2$ for all $i \neq j$, then the weight on regression j equals 1 and $\gamma_1^{sim} = \gamma_1^j$. If two (or more) linear regressions achieve the lowest σ, then the weight is between 0 and 1 with probability 1.

The lack of identification of β_1 under the assumptions (O) and (S) can be demonstrated in the four regressor example (9) used in the proof of proposition 1.

Under assumption (O), suppose $\beta_1 = 0$ (and, thus, X_1 has no effect) and $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$. Suppose the regression of Y on X_1 and X_3 has the lowest expected conditional variance, $\sigma_{[13]}^2$. Then $\gamma_1^{SiM} = \rho_{12}\beta_2$ and $CDF(0) = 1$ if $\rho_{12}, \beta_2 \neq 0$. Hence, X_1 is denoted robust when it has no effect.

Under assumption (S), suppose $\beta_1 = 0$ (and, thus, has no effect), $\beta_2 = 0$ and $\rho_{13} = \rho_{34} = 0$. Suppose the regression of Y on X_1 and X_3 has the lowest expected conditional variance, $\sigma_{[13]}^2$. Then $\gamma_1^{SiM} = \rho_{14}\beta_4$ and $CDF(0) = 1$ if $\rho_{14}, \beta_4 \neq 0$. Hence, X_1 is denoted robust when it has no effect on Y. ■

Proof: Proposition 4 (BIC). The choice of model can be determined by the differences in BIC. A model i is chosen over a model j if and only if

$$n(\log\frac{1}{n}SSE_i - \log\frac{1}{n}SSE_j) + \log(n)(K_i - K_j) < 0$$

for all $j \neq i$. The population equivalent or probability limit of $\frac{1}{n}SSE_j$ is σ_j^2. The first term diverges to infinity unless $\sigma_i = \sigma_j$. If $\sigma_i = \sigma_j$, then

$$n(\log\frac{1}{n}SSE_j - \log\frac{1}{n}SSE_i) \rightarrow^d e^W,$$

where W has a non-degenerate distribution. Then

$$BIC_j - BIC_i = \begin{cases} \infty & \text{if } \sigma_j > \sigma_i \\ -\infty & \text{if } \sigma_j < \sigma_i \\ \infty & \text{if } \sigma_j = \sigma_i \text{ and } K_j > K_j \\ -\infty & \text{if } \sigma_j = \sigma_i \text{ and } K_j < K_i \\ e^W & \text{if } \sigma_j = \sigma_i \text{ and } K_j = K_i \end{cases}.$$

Hence, BIC selects the model with the lowest σ with the fewest parameters. In case several models with the same number of variables achieve the lowest σ, a tie-breaker is necessary.

Under assumption (S), the lowest σ is achieved by the true model. This is seen by a generalization of e.g. Wooldridge (2002), p. 31, property CV.3. Let $B \subset \{X_1, .., X_K\}$ be the explanatory variables of the true submodel:

$$E_B^*(Y \mid B) = E^*(Y \mid X_1, .., X_K)$$

For any set $C \subset \{X_1, .., X_K\}$,

$$E\left(V^*(Y \mid X_1, .., X_K)\right) = E_C\left(V_C^*(Y \mid C)\right) - E\left(E^*(Y \mid X_1, .., X_K) - E_C(Y^* \mid C)\right)^2. \quad (12)$$

where $V_C^*(Y \mid C) = V_C((Y - E^*(Y \mid C)) \mid C)$. It follows that if $C \not\supseteq B$, then

$$E\left(V_B^*(Y \mid B)\right) < E_C\left(V^*(Y \mid C)\right)$$

because $E(Y \mid X_1, .., X_K) = E_B^*(Y \mid B) \neq E_C^*(Y \mid C)$ for some values of $x_1, ., x_K$.

Under assumption (O), the linear regression with the lowest σ may not include X_1. This can be seen using the same example as in the proof of BACE stated below. Hence, BIC denotes X_1 as having no effect on Y when it does. ■

Proof: Proposition 5 (AIC and AICC). The choice of model can be determined by the differences in AIC. A model i is chosen over a model j if and only if

$$n(\log\frac{1}{n}SSE_i - \log\frac{1}{n}SSE_j) + 2(K_i - K_j) < 0$$

for all $j \neq i$. The population equivalent or probability limit of $\frac{1}{n}SSE_j$ is σ_j^2. The first term diverges to infinity unless $\sigma_i = \sigma_j$. If $\sigma_i = \sigma_j$, then

$$n(\log\frac{1}{n}SSE_j - \log\frac{1}{n}SSE_i) \to^d e^W,$$

where W has a non-degenerate distribution. Then

$$AIC_j - AIC_i = \begin{cases} \infty & \text{if } \sigma_j > \sigma_i \\ -\infty & \text{if } \sigma_j < \sigma_i \\ e^W + 2(K_j - K_i) & \text{if } \sigma_j = \sigma_i \text{ and } K_j > K_i \\ e^W + 2(K_j - K_i) & \text{if } \sigma_j = \sigma_i \text{ and } K_j < K_i \\ e^W & \text{if } \sigma_j = \sigma_i \text{ and } K_j = K_i \end{cases}.$$

The corrected AIC is the same as AIC in the population since the correction term is 0 in the population. AIC selects the model with the lowest σ. In case several models with the same number of variables achieve the lowest σ, a tie-breaker is necessary.

Under assumption (S), the lowest σ is achieved by the true model according to (12). The coefficient on X_1 in that model equals β_1.

Under assumption (O), the linear regression with the lowest σ may not include X_1. This can be seen using the same example as in the proof of BACE stated below. Hence, AIC denotes X_1 as having no effect when it does. ■

Proof of Proposition 6 (BACE). The BACE estimator of the partial effect, β_1, of X_1 is $\widehat{\gamma}_1^{SDM} = \sum_i \widehat{\gamma}_1^i P(M_i \mid y)$. The posterior model probability, (5), can be rewritten as

$$P(M_j \mid y) = \frac{1}{1 + \sum\limits_{i \neq j} \frac{P(M_i)}{P(M_j)} n^{(K_j - K_i)/2} \left(\frac{1}{n}SSE_i / \frac{1}{n}SSE_j\right)^{-\frac{n}{2}}}.$$

The population analog of $\widehat{\gamma}_1^i$ is the regression coefficient, γ_1^i, on X_1 in linear regression i. The population analog of the posterior probability can be derived as the probability limit for $n \to \infty$. Assume that regression j has at most K_s regressors, $\underline{Z}$. Then $\frac{1}{n}SSE_j \to^p E_{\underline{Z}}(V(y \mid \underline{Z})) \equiv \sigma_j^2$, see the proof of proposition 4. Therefore,

$$\left(\frac{\frac{1}{n}SSE_j}{\frac{1}{n}SSE_i}\right)^{\frac{n}{2}} \to^p \begin{cases} 0 & \text{if } \sigma_i > \sigma_j \\ \infty & \text{if } \sigma_j < \sigma_i \end{cases},$$

and

$$n^{(K_j - K_i)/2} \left(\frac{\sigma_j^2}{\sigma_i^2}\right)^{\frac{n}{2}} \to^p \begin{cases} \infty & \text{if } \sigma_i > \sigma_j \\ 0 & \text{if } \sigma_i < \sigma_j \\ \infty & \text{if } \sigma_i = \sigma_j \text{ and } K_i < K_j \\ 0 & \text{if } \sigma_i = \sigma_j \text{ and } K_i > K_j \\ W & \text{if } \sigma_i = \sigma_j \text{ and } K_i = K_j \end{cases},$$

where W is a random variable with support on the unit interval. Let $\mathcal{S}$ be the set of indexes of the linear regressions with the minimum expected conditional variance: $\mathcal{S} = \arg\min\limits_i \sigma_i$. Then the probability limit of the posterior probability is:

$$P(M_j \mid y) = \begin{cases} 0 & \text{if } \sigma_j > \min\limits_i \sigma_i \\ 1 & \text{if } \sigma_j < \min\limits_{i \neq j} \sigma_i \\ 0 & \text{if } \sigma_j = \min\limits_{i \neq j} \sigma_i \text{ and } K_j > \min\limits_{i \in \mathcal{S}} K_i \\ 1 & \text{if } \sigma_j = \min\limits_{i \neq j} \sigma_i \text{ and } K_j < \min\limits_{i \in \mathcal{S}, i \neq j} K_i \\ W_1 & \text{if } \sigma_j = \min\limits_{i \neq j} \sigma_i \text{ and } K_j = \min\limits_{i \in \mathcal{S}, i \neq j} K_i \end{cases}, \qquad (13)$$

where W_1 is a random variable with support on the unit interval. Hence, the value of γ_1^{SDM} is determined by γ_1^i in the linear regression with the smallest σ.

Under assumption (S), the true model is among the linear regressions. Since the expected conditional variance is smallest for the true model according to (12), this model is chosen by BACE with probability 1 according to (13). For the true model, $\gamma_1 = \beta_1$ and, thus, $\gamma_1^{SDM} = \beta_1$.

Under assumption (O), the four regressor example (9) is used to show that BACE does not identify β_1. Suppose $\beta_1 \neq 0$ and $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$. Suppose that the linear regression of Y on X_2 and X_3 has the lowest expected conditional variance, σ. This is possible if β_2 and β_3 are sufficiently large. Then $\gamma_1^{SDM} = 0$ because X_1 is not in the model with the posterior probability equal to 1. Hence, X_1 is denoted non-robust when it does have an effect. ■

Proof of Proposition 7 (General-to-specific). The general-to-specific procedure selects the models with the smallest expected conditional variance, $E(V^*(Y \mid \underline{Z}))$, among the linear regressions with all $\gamma_k^i \neq 0$, where γ_k^i is the coefficient on a regressor X_k in regression i. The reason is that the procedure first eliminates all the linear regressions with $\gamma_k^i = 0$. Hence, if only one linear regression achieves the lowest $E(V^*(Y \mid \underline{Z}))$, say in regression j, then the procedure selects γ_1^j as the partial effect of X_1. In case several linear regressions achieve the lowest $E(V^*(Y \mid \underline{Z}))$, it is necessary with a tie-breaker.

Under assumption (S), the true submodel has the lowest $E(V^*(Y \mid \underline{Z}))$, see the proof of proposition 4, and $\gamma_1^j = \beta_1$ in that model.

Under assumption (O), the procedure does not identify β_1. This can be proved by using the same example as used in the proof for the BACE procedure. ■

Proof of Theorem 8 (CMI method). Regarding step 1). Consider the two linear regressions (projections) in (6) and (7):

$$X_i = \alpha_1^i X_1 + \sum_{X_k \in A^c \backslash X_1} \alpha_k^i X_k + u, \;\; E(X_k u) = 0$$

and

$$X_1 = \lambda_i^1 X_i + \sum_{X_k \in A^c \backslash X_1} \lambda_k^i X_k + v, \;\; E(X_k v) = 0$$

Then $\alpha_1^i = 0$ if and only if $\lambda_i^1 = 0$. This can be seen from the following projection. Let $X_{A^c} \equiv \{X_k \in A^c \backslash X_1\}$, and α_{A^c} and λ_{A^c} the corresponding parameter vectors. Then

$$\begin{bmatrix} E(X_i^2) & E(X_i X_{A^c})' \\ E(X_i X_{A^c}) & E(X_{A^c} X_{A^c}') \end{bmatrix} \begin{pmatrix} \lambda_i^1 \\ \lambda_{A^c} \end{pmatrix} = \begin{pmatrix} E(X_i X_1) \\ E(X_{A^c} X_1) \end{pmatrix}$$

Insert for X_i and use that $\alpha_1^i = 0$ by assumption, and $E(X_{A^c} u) = 0$ and $E(X_1 u) = 0$ by construction:

$$\begin{bmatrix} \alpha_{A^c}' E(X_{A^c} X_{A^c}') \alpha_{A^c} + \sigma_u^2 & \alpha_{A^c}' E(X_{A^c} X_{A^c}') \\ E(X_{A^c} X_{A^c}') \alpha_{A^c} & E(X_{A^c} X_{A^c}') \end{bmatrix} \begin{pmatrix} \lambda_i^1 \\ \lambda_{A^c} \end{pmatrix} = \begin{pmatrix} \alpha_{A^c}' E(X_{A^c} X_1) \\ E(X_{A^c} X_1) \end{pmatrix}$$

Multiply the last rows with α'_{A^c} and subtract from first row to get:

$$\lambda_i^1 \sigma_u^2 = 0$$

Hence, $\lambda_i^1 = 0$. The opposite direction can be established by similar arguments.

The best linear projection of X_1 on $(X_2, .., X_K)$ involves only a subset of $(K_s - 1)$ variables. Suppose the set A^* satisfies assumption (O). Then the explained variance of X_1 is the largest when projected on $(A^*)^c$ than on any other set of variables not containing $(A^*)^c$. Hence, the set A^* can be found by maximizing R^2 over all linear projections of X_1 on sets of variables with $(K_s - 1)$ members. In practice, let $\widehat{A}_n$ be the set for which R^2 is maximized in the linear regression of X_1 on variables in $\widehat{A}_n^c$. For any $\delta > 0$, $\exists n_1$ s.t.

$$\Pr(\widehat{A}_n = A^*) > 1 - \delta, \, \forall n \geq n_1$$

In step 2, a test for the validity of assumption (O) is performed using A^* from step 1 in (7). For the i^{th} variabel in A^*, the estimate of λ_i^1 in (7) should be 0 apart from natural variation due to sampling uncertainty. Hence, for all the corresponding t-test statistics t_i, the hypothesis of conditional mean independence is not rejected if

$$t_i \notin R_i, \, i = 1, .., (K - K_s)$$

where R_i is the rejection region of the i^{th} t-test statistics. Suppose κ_i is the level of the i^{th} t-test statistics. Then the probability of non-rejection is

$$\Pr(\textit{non-reject} \mid \widehat{A}_n = A^*) \geq 1 - \sum_{i=1}^{K-K_s} \kappa_i$$

Let κ_i^* be the nominal significance level. Each t-test is a consistent test. This implies that for any $\varepsilon > 0$, there exists an n_2 s.t. $\kappa_i < \kappa_i^* + \varepsilon/(K - K_s)$ for all i and $\forall n \geq n_2$. Let $\kappa_i^* = \kappa^*/(K - K_s)$, where κ^* can be the desired nominal significance level. This is the Bonferroni corrected significance level when testing multiple hypotheses. This implies that

$$\Pr(\textit{non-reject} \mid \widehat{A}_n = A^*) \geq 1 - (\kappa^* + \varepsilon)$$

For any $\delta > 0$, let $\kappa^* = \delta - \varepsilon$. Then

$$\Pr(\textit{non-reject} \mid \widehat{A}_n = A^*) \geq 1 - \delta, \, \forall n \geq n_2$$

This result translate directly to using the maximal absolute t-test with nominal significance level $\kappa^*/(K - K_s)$ and the same rejection region as for the individual t-test statistics. The correct level rejection region is no larger than this rejection region.

Step 3 is performed if step 2 leads to a non-rejection. Consistency of the estimator is the case when considered the correct selection of variables. For any $\delta > 0$ and $\varepsilon > 0$, $\exists n_3$ s.t.

$$\Pr(|\widehat{\beta}_1 - \beta_1| < \varepsilon \mid \textit{non-reject}, \ \widehat{A}_n = A^*) > 1 - \delta, \ \forall n \geq n_3$$

All together, let $n^* = \max(n_1, n_2, n_3)$. Then for any $\varepsilon > 0$ and $\delta > 0$,

$$\begin{aligned} \Pr(|\widehat{\beta}_1 - \beta_1| &< \varepsilon) \\ &\geq \Pr(|\widehat{\beta}_1 - \beta_1| < \varepsilon, \ \textit{non-reject}, \ \widehat{A}_n = A^*) \\ &= \Pr(|\widehat{\beta}_1 - \beta_1| < \varepsilon \mid \textit{non-reject}, \ \widehat{A}_n = A^*) \cdot \Pr(\textit{non-reject}|\widehat{A}_n = A^*) \cdot \Pr(\widehat{A}_n = A^*) \\ &> (1-\delta)(1-\delta)(1-\delta) = (1-\delta)^3, \ \forall n \geq n^* \end{aligned}$$

Hence, under assumption (O), the CMI method provides a consistent estimator of β_1.

If assumption (O) is not satisfied, then the CMI method consistently rejects the possibility of infering β_1 based on assumption (O). This can be seen from step 2. If assumption (O) is not satisfied, then at least one t-test statistics asymptotically will reject the hypothesis since the t-test is a consistent test, and, thus, so will the maximal absolute t-test. ■

References

[1] Acosta-González, E., Fernández-Rodrigues, F., 2007 Model Selection via Genetic Algorithms Illustrated with Cross-Country Growth Data. Empirical Economics 33, 313-337.

[2] Berggren, N., Elinder, M., Jordahl, H., 2008. Trust and Growth: a Shaky Relationship. Empirical Economics 35, 251-274.

[3] Bleaney, M., Nishiyama, A., 2002. Explaining Growth: A Contest Between Models. Journal of Economic Growth 7(1), 43-56.

[4] Breusch, T., 1986. Hypothesis Testing in Unidentified Models. Review of Economic Studies 53(4), 635-651.

[5] Breusch, T., 1990. Simplified Extreme Bounds. In Modelling Economic Series. Ed. C. Granger. Clarendon Press, Oxford.

[6] Burnham, K.P., D.R. Anderson, 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, USA.

[7] Crainiceanu, C., Dominici, F., Parmigiani, G., 2008. Adjustment Uncertainty in Effect Estimation. Biometrika 95(3), 635–651.

[8] Dreher, A., Sturm J.E., Vreeland, J.R., 2009. Development Aid and International Politics: Does Membership on the UN security Council influence World Bank Decisions? Journal of Development Economics 88(1), 1-18.

[9] Durlauf, S.N., 2001, Manifesto for a Growth Econometrics. Journal of Econometrics 100(1), 65-69.

[10] Durlauf, S.N., Johnson, P., Temple, J., 2005. Growth Econometrics, Handbook of Economic Growth. Edt. Aghion, P. and S. N. Durlauf, Elsevier.

[11] Durlauf S.N., Kourtellos, A., Tan C.M., 2008. Are any Growth Theories Robust? Economic Journal 118(527), 329-346.

[12] Ericsson, N., 2008. The Fragility of Sensitivity Analysis: An Encompassing Perspective. Oxford Bulletin of Economics and Statistics 70, SUPPLEMENT 0305-9049, 895-914.

[13] Fernandez, C., Ley, E., Steel, M.F.J., 2001. Model Uncertainty in Cross-country Growth Regressions, Journal of Applied Econometrics 16, 563–576.

[14] Ghosh, S., Yamarik, S., 2004. Are Regional Trading Arrangements Trade Creating? An Application of Extreme Bounds Analysis. Journal of International Economics 63, 369-395.

[15] Goeman, J., Van de Geer, S., Houweiling, H., 2006. Testing Against a high-dimensional Alternative. Journal of the Royal Statistical Society B 68(3), 477-493.

[16] Goldberger, A.S., 1991. A Course in Econometrics. Harvard University Press.

[17] Granger, C., Uhlig, H., 1990. Reasonable Extreme Bound Analysis. Journal of Econometrics 44(1-2), 159-170.

[18] de Haan, J., 2007. Political Institutions and Economic Growth Reconsidered. Public Choice 131(3-4), 281-292.

[19] Hansen, B.E., 1999. Discussion of 'Data Mining Reconsidered.' Econometrics Journal 2(2), 192-201.

[20] Hansen, P.R., 2003. Regression Analysis with Many Specifications: A Bootstrap Method for Robust Inference. Working Paper.

[21] Hendry, D. F., 1995. Dynamic Econometrics. Oxford: Oxford University Press.

[22] Hendry, D.F., Krolzig, H-M., 1999. Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez. Econometrics Journal 2, 202-219.

[23] Hendry, D.F., Krolzig, H-M., 2004. We Ran One Regression. Oxford Bulletin of Economics and Statistics 66(5), 799-810.

[24] Hoover, K., Perez, S., 1999. Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search. Econometrics Journal 2, 167-191.

[25] Hoover, K., Perez, S., 2004. Truth and Robustness in Cross-country Growth Regressions. Oxford Bulletin of Economics and Statistics 66(5), 765-798.

[26] Huang, J., Horowitz, J.L., Wei, F., 2010. Variable Selection in Nonparametric Additive Models. Annals of Statistics 38(4), 2282-2313.

[27] Jensen, P.S., 2006. Essays on Growth Empirics and Economic Development. PhD thesis, Department of Economcis, University of Aarhus.

[28] Jensen, P.S., 2010. Testing the Null of a low-dimensional Growth Model. Empirical Economics 38, 193-215.

[29] Jones, G., Schneider, W.J., 2006. Intelligence, Human Capital and Economic Growth: A Bayesian Averaging of Classical Estimates (BACE) approach. Journal of Economic Growth 11(1), 71-93.

[30] Leamer, E., 1983. Let's Take the Con Out of Econometrics. American Economic Review 73(1), 31-43.

[31] Leamer, E., Leonard, H., 1983. Reporting the Fragility of Regression Estimates. The Review of Economics and Statistics 65(2), 306-317.

[32] Levine, R., Renelt, D., 1992. A Sensitivity Analysis of Cross-Country Growth Regressions. American Economic Review 82(2), 942-963.

[33] Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley-interscience, USA.

[34] Magnus, J.R., Powell, O., Prüfer, P. 2010. A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics. Journal of Econometrics 154, 139-153.

[35] McAleer, M., Pagan, A.R., Volker, P.A., 1985. What Will Take the Con Out of Econometrics? The American Economic Review 75(3), 293-307.

[36] McAleer, M., 1994. Sherlock Holmes and the Search for Truth: A Diagnostic Tale. Journal of Economic Surveys 8(4), 317-370.

[37] McQuarrie, A.D.R, Tsai, CL, 1998. Regression and Time Series Model Selection. World Scientific, Singapore.

[38] Raftery, A., 1995. Bayesian Model Selection in Social Research. Sociological Methodology 25, 111-116.

[39] Raftery, A., Madigan, D., Hoeting, J.A., 1997. Bayesian Model Averaging for Linear Regression Models Journal of the American Statistical Association 92(437), 179-191.

[40] Sala-i-Martin, X., 1997. I Just Ran Two Million Regressions. American Economic Review 87(2), 178-183.

[41] Sala-i-Martin, X., 2001. Comment on "Growth Empirics and Reality." The World Bank Economic Review 15(2), 277-282.

[42] Sala-i-Martin, X., Doppelhofer, G., Miller, R., 2004. Determinants of Long-Term growth: A Bayesian Averaging of Classical Estimates (BACE) approach. American Economic Review 94(4), 813-835.

[43] Sturm J.E., de Haan J., 2005. Determinants of Long-Term Growth: New Results Applying Robust Estimation and Extreme Bounds Analysis. Empirical Economics 30(3), 597-617.

[44] White, H., 2000. A Reality Check for Data Snooping. Econometrica 68, 1097-1127.

[45] Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, USA.

Research Papers 2010

2010-61: Christian Bach and Bent Jesper Christensen: Latent Integrated Stochastic Volatility, Realized Volatility, and Implied Volatility: A State Space Approach

2010-62: Bent Jesper Christensen and Malene Kallestrup Lamb: The Impact of Health Changes on Labor Supply: Evidence from Merged Data on Individual Objective Medical Diagnosis Codes and Early Retirement Behavior

2010-63: Martin M. Andreasen: How Non-Gaussian Shocks Affect Risk Premia in Non-Linear DSGE Models

2010-64: Tim Bollerslev and Viktor Todorov: Jump Tails, Extreme Dependencies, and the Distribution of Stock Returns

2010-65: Almut E. D. Veraart: How precise is the finite sample approximation of the asymptotic distribution of realised variation measures in the presence of jumps?

2010-66: Ole E. Barndorff-Nielsen, David G. Pollard and Neil Shephard: Integer-valued Lévy processes and low latency financial econometrics

2010-67: Shin Kanaya and Dennis Kristensen: Estimation of Stochastic Volatility Models by Nonparametric Filtering

2010-68: Dennis Kristensen and Anders Rahbek: Testing and Inference in Nonlinear Cointegrating Vector Error Correction Models

2010-69: Søren Johansen: The analysis of nonstationary time series using regression, correlation and cointegration –with an application to annual mean temperature and sea level

2010-70: Søren Johansen and Morten Ørregaard Nielsen: A necessary moment condition for the fractional functional central limit theorem

2010-71: Nektarios Aslanidis and Isabel Casas : Modelling asset correlations during the recent financial crisis: A semiparametric approach

2010-72: Søren Johansen and Katarina Juselius: An invariance property of the common trends under linear transformations of the data

2010-73: Peter Sandholt Jensen and Allan H. Würtz: Estimating the effect of a variable in a high-dimensional regression model