A Bayesian Approach to Variable Selection when the Number of Variables is Very Large

Harri T. Kiiveri

Abstract

In this paper, we present a rapid Bayesian variable selection technique which can be used when the number of variables is *much* greater than the number of samples. The method can handle tens of thousands of variables, such as might be measured using biological array technologies. A general formulation is first given, followed by specific details for the class of generalised linear models.

Keywords: Bayesian; Jeffreys hyperprior; posterior; variable selection; EM algorithm; generalised linear models; survival analysis

1 Introduction

Traditional methods of variable selection for statistical models include backward and forward stepwise procedures, and all subsets calculations using branch and bound algorithms, see for example [19]. Typically some criterion such as LAIC or BICE is used to guide the selection process. These stepwise methods have also been implemented in software packages such as R and Splus for more general models than linear regression, *e.g.* generalised linear models.

These traditional methods were implicitly designed for situations where the number of variables is less than the number of observations, and the number of variables was at most of the order of hundreds. Unfortunately, these methods do not cope well with large numbers of variables, say of the order of ten thousand, or when the number of observations is less than the number of variables. In these circumstance they either fail completely, or, even if they can be modified to work, require such a huge computational effort that they are impractical to use.

More recently, Bayesian variable selection methods based on Markov chain Monte Carlo methods have been developed, see for example [4, 13, 21, 22]. These have some attractive properties; however, aside from other issues, these methods are computationally intensive and do not scale up well to problems with ten thousand variables or more.

With the advent of microarray technologies, variable selection problems with ten thousand variables and hundreds of observations are becoming quite common, with the likelihood that the problem sizes will scale up at least one order of magnitude in the near future. Clearly, new methods are required to handle these large problems.