Bayesian Model Selection of Latent Variable Model

潜在变量模型的贝叶斯模型选择

Li Yunxian
Tang Niansheng

西南交通大学出版社
Http://press.swjtu.edu.cn

Bayesian Model Selection of Latent Variable Model

潜在变量模型的贝叶斯模型选择

Li Yunxian　Tang Niansheng

西南交通大学出版社
·成　都·

图书在版编目（CIP）数据

潜在变量模型的贝叶斯模型选择 = Bayesian model selection of latent variable model : 英文 / 李云仙，唐年胜著. —成都：西南交通大学出版社，2013.7
ISBN 978-7-5643-2429-2

Ⅰ. ①潜… Ⅱ. ①李… ②唐… Ⅲ. ①贝叶斯方法－数学模型－研究－英文 Ⅳ. ①O212.8

中国版本图书馆 CIP 数据核字（2013）第 152636 号

潜在变量模型的贝叶斯模型选择

李云仙　唐年胜　著

责任编辑	罗　旭
封面设计	何东琳设计工作室
出版发行	西南交通大学出版社 （四川省成都市金牛区交大路 146 号）
发行部电话	028-87600564　028-87600533
邮政编码	610031
网　　址	http: //press.swjtu.edu.cn
印　　刷	成都蜀通印务有限责任公司
成品尺寸	146 mm × 208 mm
印　　张	5.437 5
字　　数	195 千字
版　　次	2013 年 7 月第 1 版
印　　次	2013 年 7 月第 1 次
书　　号	ISBN 978-7-5643-2429-2
定　　价	23.80 元

摘 要

潜在变量模型常用于分析潜在变量以及显变量之间的关系。在模型的统计推断中，模型选择是一个非常重要的问题。模型选择的方法非常多，例如 AIC、BIC、DIC、贝叶斯因子等。近年来，贝叶斯方法在模型估计及模型选择问题中的应用比较热门。本书以潜在变量模型为研究对象，并主要研究贝叶斯模型选择方法。本书将引入一种基于贝叶斯准则的模型选择方法，并且与其他贝叶斯方法，如贝叶斯因子及 DIC 相比较。同时，本书将介绍不同的潜在变量模型，主要包括非线性潜在变量模型，含有有序变量的潜在变量模型，两水平潜在变量模型，有限混合潜在变量模型，以及含有不可忽略缺失数据潜在变量模型。对于每一类模型均以模拟研究及实例研究来说明所提出方法在模型选择中的满意表现。

Abstract

Latent variable models are used to analyze the relationship between the latent variable and manifest variable. Model selection is one of the most important issues in statistical inference. There are bunches of different methods that can be used for model selection, like AIC, BIC, DIC, Bayes factor, and so on. Recently, Bayesian approach becomes a popular method in model estimation and model selection. This book focuses on model selection in latent variable models. We introduce a Bayesian criterion method for model selection in latent variables models, including nonlinear latent variable model, latent variable model with ordered categorical variables, two-level latent variable models, finite mixture latent variable model and latent variable model with non-ignorable missing data. In addition, and the results are compared with other Bayesian methods, including Bayes factor and DIC. Different kinds of latent variable models are considered in this book. In each model, a simulation study and real example are presented to show the satisfactory performance of the proposed method.

Preface

Latent variable models are used to analyze the relationship between the latent variable and manifest variable. Model selection is one of the most important issues in statistical inference. Two years ago, we have written a book "A Bayesian Criterion-based Model Selection Method in Structural Equation Models". In that book, four different kinds of structural equation models were discussed. However, we didn't take the other model selection methods and missing data in latent variable models into consideration. In order to make the theory of model selection in latent variable models more complete, we write this book. Compared with the former book, two chapters are added to deal with these problems. Specifically, we give a brief review of the approaches to model selection in Chapter 1. As a matter of fact, Professor Lee, my supervisor in The Chinese University of Hong Kong, has discussed this problem in his book "Structural Equation Modeling: A Bayesian Approach". He reviewed different model selection methods in Chapter 5. With his permission, in Chapter 1 of this book, I quoted part of Chapter 5 from Professor Lee's book. In Chapter 6, we discussed the missing data in latent variable models. Different missing mechanisms are considered in this chapter.

In closing, I'd love to thank everyone for their kind wishes and support.

CONTENTS

Chapter 1 Introduction to Model Selection

1.1 Introduction

One important statistical inference beyond estimation is model selection. In the field of latent variable modeling, a common approach for hypothesis testing is to use the significance tests on the basis of p-values that are determined by some asymptotic distributions of the test statistics. As pointed out in the statistics literature (see e.g. Berger & Sellke, 1987; Berger & Dalampady, 1987; Kass & Raftery, 1995) there are problems associated with such an approach. Those related to latent variable models are discussed as follows:

(i) Tests on the basis of p-values tend to reject the null hypothesis too frequently with large sample sizes. A dramatic example with a sample size 113,556 was given by Raftery(1986), where a substantively meaningful model (associated with the null hypothesis)n that explained 99.9% of the deviance was rejected by a standard chi-squared test with an extremely small p-value. In the traditional analysis of latent variable models, various descriptive fit indexes, such as the well-known normed or non-normed fit indexes (Bentler & Bonett, 1980) and the comparative fit index (Bentler, 1992) have been proposed as complementary measures for the goodness-of-fit of the model. Very often, the values of the fit indexes are over 0.95, but the p-values of the

χ^2-test are less than 0.01. Under these situations, the conclusions drawn from these two testing methods seem contradictory.

(ii) The p-value of a significance tests in hypothesis testing is a measure of evidence against the null model, not a means of supporting/proving the model. Hence, the conclusion of a significance test can only be used to reject the null hypothesis and cannot offer an assessment of the strength of the evidence in favor of the null hypothesis. As a result, even the chi-square goodness-of-fit test does not reject the null hypothesis, it neither can be used to conclude that the posited model is better than the alternative model, nor to conclude that the given data support the posited model. On the other hand, rejection of the null hypothesis by such a test does not indicate the alternative model is better.

(iii) The significance tests as well as descriptive fit indexes mentioned above cannot be applied to test nonnested hypotheses or to compare nonnested models. Therefore only a hierarchy of nested hypotheses can by assessed, see for example, Bollen (1989). However, we are very often interested in assessing non-nested latent variable models in practical applications.

The well-known statistic in Bayesian model comparison, namely the Bayes factor (Berger, 1985; Kass & Raftery, 1995), is a Bayesian approach for hypothesis testing that does not have the above problems. In the field of latent variable model, we are often interested in comparing a discrete set of competing models. Other methods that emphasize for comparing continuous families of models (see Gelman et al. 2003, and the references therein) are not considered. In general, the computation of Bayes factor is difficult. Various computational methods have been proposed, see Kass and Raftery (1995). A simple but rough approximation, namely the Bayesian Information Criterion (BIC) has been used for model comparison of some latent variable

models. For example, Raftery (1993) applied it to the LISREL model, Lee and Song (2001) applied it to a two-level structural equation model, and Jedidi, Jagpal, and DeSarbo (1997) applied it to finite mixtures of structural equation models with a fixed number of components, among others. Other useful methods for computing the Bayes factor have been established on the basis of posterior simulation, using recently developed MCMC methods. DiCiccio et al. (1997) provided a comparative study on a variety of methods, from Laplace approximation to importance sampling, and bridge sampling, and concluded that bridge sampling is an attractive method. Gelman and Meng (1998) showed that path sampling is a direct extension of the bridge sampling. Naturally, it is expected that path sampling is even better.

Different from estimation, Bayesian model comparison using Bayes factor may by sensitive to prior distributions of the parameters. Hence, these distributions should be selected with care and sensitivity analysis of the prior inputs should be conducted. Another widely used Bayesian model selection statistics is the Deviance Information Criterion (DIC) (Spiegelhalter et al, 2002). It is well known that for complex statistical models, the computation of Bayes factor is difficult (DiCiccio et al, 1997). Like the Bayesian Information Criterion (BIC), DIC takes into account the number of unknown parameters in the model. As the software WinBUGS (Spiegelhalter et al., 2003) provides the DIC values for most latent variable models, the application of DIC is convenient.

While Bayes factor and DIC have some nice features, they have limitations. As mentioned before, Bayes factor requires proper prior distributions of the parameters. In fact, it will favor the competitive model M_0 if the prior of the parameters in model M_1 has a very large spread so as to make it non-informative. This is known as the

"Bartletts Paradox" . Moreover, for competitive models M_0 and M_1, such as multilevel latent variable models with very different structures, it is difficult to find a direct path to link them when applying the path sampling. Under these cases, some auxiliary models may have to be used in computing the Bayes factor (see Lee, 2007). This will increase the computational burden. For DIC, it assumes the posterior mean to be a good estimator; and for some models (for example, the mixture latent variable models), WinBUGS does not give the DIC values. Moreover, if the difference in DIC values is small, only reporting the model with the smallest DIC value may be misleading. In this book, motivated by the above limitations of the Bayes factor and DIC, we propose an attractive Bayesian statistic for model selection for different kinds of latent variable models.

The proposed Bayesian statistic, called the L_v measure, is a criterion-based method that does not require proper prior distributions of the parameters. It will be shown that the computational burden involved is light, and the statistic can be obtained conveniently via observations simulated for the Bayesian estimation. Basically, the L_v measure involves two components. The first component is related to the reliability of the prediction, and the second component measures the discrepancy between the prediction and the observed data. Hence, it can be used to examine the goodness-of-fit of the model to the observed data. We will also consider the calibration distribution of the L_v measure, which will allow us to compare two competing models in more details.

An introduction to the Bayes factor will be presented in Section 1.2, followed by the discussion of Lv measure in Section 1.3. Section 1.4 introduces other methods for model comparison and model checking, and discussion is given in Section 1.6.

1.2 · Bayes Factor

In this section, we introduce an important Bayesian statistic, the Bayes factor (Berger, 1985; Kass & Raftery, 1995), for model comparison/selection. This statistic has a solid logical foundation that offers great flexibility. It has been extensively applied to a lot of statistical models, see the references given in Kass and Rartery (1995). Its applicability is further enhanced by the powerful MCMC methods that are recently developed in statistical computing.

Suppose the given data Y with a sample size n have arisen under one of the two competing models M_1 and M_0 according to a probability density $p(Y|M_1)$ or $p(Y|M_0)$, respectively. Let $p(M_0)$ be the prior probability of M_0 and $P(M_1)=1-P(M_0)$, and let $P(M_k|Y)$ be the posterior probability, for $k=0,1$. From the Bayes theorem, we obtain

$$p(M_k|Y)=\frac{p(Y|M_k)p(M_k)}{p(Y|M_1)p(M_1)+p(Y|M_0)p(M_0)},k=0,1$$

Hence,
$$\frac{p(M_1|Y)}{p(M_0|Y)}=\frac{p(Y|M_1)p(M_1)}{p(Y|M_0)p(M_0)} \tag{1.1}$$

The Bayes facter for comparing M_1 and M_0 is defined as

$$B_{10}=\frac{p(Y|M_1)}{p(Y|M_0)} \tag{1.2}$$

From (1.1), we see that

Posterior odds = Bayes factor × prior odds.

In the special case where the competitive models M_1 and M_0 are equally probable a prior so that $p(M_1)=p(M_0)=0.5$, the Bayes factor is equal to the posterior odds in favor of M_1. In general, it is a

summary of evidence provided by the data in favor of M_1 as oppose to M_0, or in favor of M_0 to M_1. It may reject a null hypothesis associated with M_0, or may equally provide evidence in favor of the null hypothesis or the alternative hypothesis associated with M_1. Moreover, unlike the significance test approach that is based on the likelihood ratio criterion and its asymptotic chi-square statistic, the comparison does not depend on the assumption that either model is "true" . Moreover, it can be seen from (1.2) that the same data set is used in the comparison. Hence, it does not favor the alternative hypothesis (or M_1) in extremely large samples. Finally, it can be applied to compare nonnested models M_0 and M_1.

According to the suggestion given in Kass and Raftery (1995), the criterion that is given in Table 1.1 is used for interpreting B_{10} and $2\log B_{10}$. Kass and Raftery(1995) pointed out that these categaories furnish appropriate guidelines for practical applications of the Bayes fator. Depending on the competing models M_0 and M_1 for fitting a given data set. If the Bayes factor (or 2log Bayes factor) reject the null hypothesis H_0 that is associated with M_0, we can conclude that the data give evidence to support the alternative hypothesis H_1, a more definite conclusion of supporting H_0 can be attained.

Table 1.1 Interpretation of Bayes factor

B10	2 logB10	Evidence against H0 (M0)
< 1	< 0	Negative (supports H0 (M0))
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive (supports H1 (M1))
20 to 150	6 to 10	Strong
> 150	> 10	Decisive

The interpretation of evidence provided by Table 1.1 depends on

the specific context. For two nonnested competitive models, say M_0 and M_1, we should select M_0 if $2\log B_{10}$ is negative. If $2\log B_{10}$ is in (0,2), we may interpret M_1 is slightly better than M_0 and hence,

It may be better to select M_1. The choice of M_1 is more definite if $2\log B_{10}$ is larger than 6. For two nested competitive models, say M_0 is nested in the more complicated model M_1, $2\log B_{10}$ is most likely larger than zero. If M_1 is significantly better than M_0, it can be much larger than 6. Then the above criterion will suggest a decisive conclusion to select M_1. However, if $2\log B_{10}$ is in (0,2), then the difference between M_0 and M_1 is not worth more than a bare mention . Under this situation, great caution should be taken in drawing a definite conclusion. According to the "parsimonious" guildline in practical applications, it may be desirable to select M_0 if it is much simpler than M_1. The criterion given in Table 1.1 is a suggestion, and it is not necessary to regard it as a strict rule. Similarly in frequentist hypothesis testing, one may take the type I error to be 0.05 or 0.10, and the choice is decided with other factors in the substantive situation, Similar to other data analyses, for conclusions drawn from the marginal cases, it is always helpful to conduct other analysis, for example residual analysis, to cross-validate the results. Generally speaking, model selection should be approached on a problem-by-problem basis. It is also desirable to take the opinions from experts into account if no clear conclusion can be drawn.

From (1.2) ,we see that the density $p(Y \mid M_k)$ is involved in the Bayes factor. This function is obtained by integrating $p(Y \mid \theta_k, M_k)p(\theta_k \mid M_k)$ over the parameter space. That is

$$p(Y \mid M_k) = \int p(Y \mid \theta_k, M_k)p(\theta_k, M_k)\mathrm{d}\theta_k \qquad (1.3)$$

Where θ_k is the parameter vector in $M_k, p(\theta_k \mid M_k)$ is its prior

density, and $p(Y \mid \theta_k, M_k)$ is the probability density of Y given θ_k. The dimension of this integral is equal to the dimension of θ_k. This quanity can be interpreted as the marginal likelihood of the data, obtained by integrating the joint density of (Y, θ_k) over θ_k. It can also be interpreted as the predictive probability of the data; this is , the probability of seeing the data that actually were observed, calculated before any data become available. Sometimes, it is also called an integrated likelihood. Note that, as in the computation of the likelihood ratio statistic but unlike in some other applications of likelihood, all constants appearing in the definition of the likelihood $p(Y \mid \theta_k, M_k)$ must be retained when computing B_{10}. In fact, B_{10} is closely related to the likelihood ratio statistic, in which the parameters θ_k are eliminated by maximinzation rather than by integration. Very often, it is very difficult to obtain B_{10} analytically, and various analytic and numerical approximations have been proposed in the literature. For example, Chib (1995), and Chib and Jeliazkov (2001) respectively developed efficient algorithms for computing the marginal likelihood through MCMC chains produced by the Gibbs sampler and the MH algorithm. Based on the results of DiCiccip et al.(1997); and the recommendation of Gelman and Meng (1998), we will apply path sampling to compute the Bayes factor for model comparison.

A procedure based on path sampling (Gelman & Meng,1998) is introduced in this section for computing the Bayes factor. The key feature of path sampling is to compute the ratio of normalizing constants of probability densities (or equivalently difference of the logarithm of them). Hence it can be applied to compute the Bayes factor. Following Gelman and Meng (1998), we motivate this computing tool from importance sampling (see for example Gelfand & Dey, 1994) and bridge sampling (Meng & Wong, 1996).

In the context of latent variable models, we consider two

competitive models M_0 and M_1 with the matrix Ω of latent variables. Let θ be the vector of unknown parameters in both M_1 and M_0. Usually, it is rather difficult to apply path sampling to evaluate $p(Y \mid M_k)$ under complicated situations. Hence, we first use the idea of data arguementation to include the latent variables in the computation, by considering the complete-data sea (Y, Ω). Basically, we compute

$$z(k) = p(Y \mid M_k) = \int p(Y, \Omega, \theta \mid M_k) \mathrm{d}\Omega \mathrm{d}\theta, k = 0,1$$

and obtain B_{10} as the ratio of $z(1)/z(0)$.

The key idea of importance sampling is the following equality:

$$z(1) = p(Y \mid M_k) = \int \frac{p(Y, \Omega, \theta \mid M_1)}{p(\Omega, \theta \mid Y, M_1)} p(\Omega, \theta \mid Y, M_1) \mathrm{d}\Omega \mathrm{d}\theta$$

Viewing $p(\Omega, \theta \mid Y, M_1)$ as a probability density function, $z(1)$ is the expectation of $p(Y, \Omega, \theta \mid M_1) / p(\Omega, \theta \mid Y, M_1)$ with respect to the joint distribution if Ω and θ under M_1. On the basis of the simple idea that approximating the expectation by the sample mean of observations in a sufficiently large sample, it can be approximated as below,

$$z(1) \cong \frac{1}{j} \sum_{j=1}^{J} \frac{p(Y, \Omega^j, \theta^j \mid M_1)}{P(\Omega^j, \theta^j \mid Y, M_1)},$$

where $\{\Omega^j, \theta^j, j = 1, \cdots, J\}$ is a sample of observations drawn from the target distribution $p(\Omega, \theta \mid Y, M_1)$ via some posterior simulation methods. Similarly we can obtain z(0). Finally, the Bayes factor can be estimated via z(1)/z(0). In this method, we require to simulate observation from both $p(\Omega, \theta \mid Y, M_0)$ and $p(\Omega, \theta \mid Y, M_1)$. The following slight modification may be preferable. It follows from an identity in Meng and Wrong (1966, equation 1.4) that the ratio of normalizing constants can be expressed as,