行政院國家科學委員會專題研究計畫　成果報告

具相關性資料之統計分析(4/4)

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-009-002-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學統計學研究所

計畫主持人：李昭勝

報告類型：完整報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 92 年 10 月 28 日

行政院國家科學委員會專題研究計畫成果報告
具相關性資料之統計分析(4/4)

計畫編號：NSC 91-2118-M-009-002
執行期限：2002 年 8 月 1 日至 2003 年 7 月 31 日
主持人：李昭勝博士　國立交通大學統計學研究所

本報告含二篇完成之研究成果。

一、 Bayesian Analysis of Mixture Modeling Using the Multivariate *t* Distribution

此乃與博士生林宗儀教授（目前任教於東海大學統計系）及碩士生倪惠芬（目前就讀於台大公衛所博士班）合作的文章。本文已被 Statistics and Computing(an SCI journal) 接受。不久即將出刊。其中、英文之摘要如下。

（一)、中文摘要

使用多變量 t 分佈的有限混合模型已被顯示為常態混合的穩健性延伸。本文中，我們使用貝氏方法來推論 t 分佈的混合模型的參數。微弱訊息的先驗分佈用以避免造成不可積分的後驗分佈。我們使用已觀察到的數據和非完整的未來向量當作樣本，提出兩個後驗分佈最高點的有效 EM 程式。同時也使用馬可夫鏈蒙地卡羅 (MCMC)抽樣策略以得到參數的驗後分佈。我們藉由實際的例子證明貝氏方法優於最大概似法。

關鍵詞：
ECM、ECME、最大後驗值、最大概似估計法、馬卡夫鏈蒙地卡羅、t 分佈的混合模型

（二)、英文摘要

A finite mixture model using the multivariate *t* distribution has been shown as a robust extension of normal mixtures. In this paper, we present a Bayesian approach for inference about parameters of *t*-mixture models. The specifications of prior distributions are weakly informative to avoid causing nonintegrable posterior distributions. We present two efficient EM-type algorithms for computing the joint posterior mode with the observed data and an incomplete future vector as the sample. Markov chain Monte Carlo sampling schemes are also developed to obtain the target posterior distribution of parameters. The advantages of Bayesian approach over the maximum likelihood method are demonstrated via a set of real data.

Keywords:
ECM; ECME; maximum a posteriori; maximum likelihood estimation; MCMC; *t* mixture model

（三)、報告內容

Finite mixture models introduced by Pearson (1894) have been a useful tool for modeling the data that are thought to come from several di®erent groups with varying proportions. In the

past two decades, tremendous improvements and applications have been made in across many research fields. The fundamental idea and usefulness of the mixture models are explained in McLachlan and Basford (1988) and Titterington (1985). A comprehensive introduction to the theory and recent advances can be found in McLachlan and Peel (2000).

Historically, much effort has been devoted to the maximum likelihood (ML) approach for fitting the mixture models. It was first considered by Rao (1948), who used Fisher's scoring method for a mixture of two normal distributions with equal variance. The computation of ML estimates cannot be easily manipulated until the EM algorithm was introduced by Dempster *et al.* (1977). More recently, Peel and McLachlan (2000) considered how to model a mixture of multivariate *t* distributions. They provided the ECM algorithm for parameter estimation and showed the robustness of the model in clustering.

Redner and Walker (1984) pointed out that the ML approach for finite mixture model could encounter unbounded likelihood in some special cases. Hathaway (1985) suggested that using simple constraints in an optimization problem can lead to a strongly consistent and global solution. Hosmer (1973) gave an example of including a portion of labeled observations for each component. However, both solutions are restricted to the univariate case.

In recent developments of computational methods, Bayesian methods are considered an alternative way to deal with mixture models. Diebolt and Robert (1994) used data augmentation and Gibbs sampling as approximation methods for evaluating the posterior distribution and Bayes estimators. They also showed that the duality principle leads to stronger and more general results about the convergence of the simulated Markov chains and of the related moments. Richardson and Green (1997) considered a hierarchical prior that avoid the mathematical pitfalls of using improper priors in mixture model. More recently, Fruhwirth-Schnatter (2001) explored the MCMC output of the random sampler to find suitable identifiability constraints in dealing with label switching problems.

In this article, we extend the ML approach of Peel and McLachlan (2000) to deal with a mixture of *t* distributions from Bayesian viewpoints. Since some observations could be missing in many practical situations, our approach is more general as it allows for some of the observed vectors to be partly known. For the sake of clarity, we only demonstrate one partly known individual and treat it as an incomplete future vector in the model. We compare the prediction and classification results on a real data set between ML and MCMC techniques via cross-validations.

In conclusion, this paper has established efficient EM-type algorithms for calculating the joint posterior mode and provided a workable MCMC algorithm for sampling from the posterior distribution of *t* mixture models. Our algorithms appear quite flexible and have applications in prediction and classification of a partially observed vector. Meanwhile, one can straightforwardly evaluate the predictive distribution using MCMC samples. The techniques can be applied in the presence of missing data and easily generalized to situations in which a set of incomplete future vectors are simultaneously considered.

In our illustrated example, Bayesian MCMC can provide more accurate classification probability and better prediction accuracy than the ML method. It is fair to say that the proposed Bayesian methods should be quite useful for practitioners in dealing with a mixture of *t* distributions.

（四）、參考文獻

Anscombe F.J. 1967. Topics in the investigation of linear relations fitted by themethod of least squares. Journal of the Royal Statistical Soc B 29: 1-52.

Basford K.E., Greenway D.R., McLachlan G.J., and Peel D. 1997. Standard errors of fitted means under normal mixture. Computational Statistics 12: 1-17.

Brooks S.P. and Gelman A. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7: 434-455.

Campbell N.A. and Mahon R.J. 1974. A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. Australian Journal of Zoology 22: 417-425.

van Dyk D.A., Meng X.L. and Rubin D.B. 1995. Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. Statistica Sinica 5: 55-75.

Dempster A.P., Laird N.M. and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39: 1-38.

Diebolt J. and Robert, C.P. 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, B 56: 363-375.

Efron B. and Tibshirani R. 1986. Bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1: 54-77. London: Chapman & Hall.

Fruhwirth-Schnatter S. 2001. Markov

Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96: 194-209.

Geisser S. 1975. The predictive sample reuse method with applications. Journal of the American Statistical Association 70: 320-328.

Gelfand A.E. and Smith A.F.M. 1990. Sampling based approaches to calculate marginal densities. Journal of the American Statistical Association 85: 398-409.

Gelman A.E. and Rubin D.B. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7: 457-511.

Gelman A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 1995. Bayesian Data Analysis.

Champmen & Hall, London. Hathaway R.J. 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Annals of Statistics 13(2): 795-800.

Hosmer D.W. 1973. A comparison of iterative maximum-likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, 29: 761-770.

Liu C.H. and Rubin D.B. 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika 81: 633-648.

Liu C.H. 1995. Missing data imputation using the multivariate *t* distribution. Journal of Multivariate Analysis 53: 139-158.

Liu C.H. and Rubin D.B. 1995. ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. Statistica Sinica 5: 19-39.

Mardia, K.V., Kent, J.T. and Bibby, J. M., 1979. Multivariate analysis. Academic Press, Inc. London.

McLachlan G.J. and Peel D. 1998. Robust cluster analysis via mixtures of multivariate *t*-distribution. In Lecture Notes in Computer Science, 1451, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.). Berlin: Springer-Verlag, pp. 658-666.

McLachlan G.J. and Peel D. 2000. Finite Mixture Model. New York: Wiely.

McLachlan G.J. and Basford, K. E., 1988. Mixture Models: Inference and Application to Clustering. New York: Marcel Dekker.

Meng X.L. and Rubin D.B. 1991. Using EM to obtain asymptotic

variancecovariance
matrices: The SEM algorithm. Journal of the American Statistical
Association 86: 899-909.

Meng X.L. and Rubin D.B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80: 267-278.

Pearson K. 1894. Contributions to the theory of mathematical evolution. Philosophical Transactions of the Royal Society of London A 185: 71-110.

Peel D. and McLachlan G.J. 2000. Robust mixture modeling using the *t* distribution. Statistics and Computing 10: 339-348.

Raftery A.E. 1996. Hypothesis testing and model selection via posterior simulation. In practice Markov Chain Monte Carlo (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 163-188. Chapman & Hall, London.

Rao C.R. 1948. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society B 10: 159-203.

Redner R.A. and Walker H.F. 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26: 195-239.

Relles D.A., and Rogers W.H. 1977. Statistics are fairly robust estimators of location. Journal of the American Statistical Association 72: 107-111.

Richardson S. and Green P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society B 59: 731-792.

Stephens M.A. 1997. Bayesian method for mixtures of normal distributions. Ph.D. thesis, University of Oxford.
Stone, M. 1974. Cross-validatory choice and assessment of statistical prediction (with discussion). Journal of the Royal Statistical Society B 36: 111-147.

Tiao G.C. 1967. Discussion on “Topics in the investigation of linear relations fitted by the method of least squares.” Journal of the Royal Statistical Society, series B 29:

Titterington, D.M., Smith, A.F.M., and Markov, U.E., 1985. Statistical analysis of finite mixture distributions. New York: Wiely.

Vounatsou P. and Smith A.F.M. 1997. Simulation-based Bayesian inferences for two-variance components linear models. Journal of Statistical Planning Inferences 29: 139-161.

（五）、計畫成果自評

本研究成果乃計畫所提研究的一部份，其被接受的期刊是個 Impact

Factor 1.00的國際刊物，相當值得。

二、Bayesian Estimation for Time Series Regressions with Applications

此乃與逢甲大學陳婉淑教授、博士生牛維方、碩士生李向宇合作的文章。是一篇有關 time series regression的貝氏分析。此與所提之計畫有關。本文已被 Journal of Statistical Computing and Simulation (an SCI journal) 所接受。其中、英文摘要如下。

(一)、中文摘要

此計畫中，我們在貝氏的推演架構下，為時間序列迴歸模型提供一個估計程序。估計時間序列中，主要的障礙包含探討其起始的觀察值。我們可藉由 Wise(1955)的準確方法，得到一個準確概似函數，用以估計參數。我們也提出了一個不會和時間序列之穩定性相衝突的重新參數化，這是 Chib(1993) 以及 Chib 和 Greenberg(1994)的研究中未考慮到的。模擬研究顯示我們的方法可獲得更準確的推演。

關鍵詞：
自回歸過程、概似、馬卡夫鏈蒙地卡羅

(二)、英文摘要

We propose an estimation procedure for the time series regression models under the Bayesian inference framework. The major obstacle for estimating a time series involves treating its initial terms. With the exact method of Wise (1955), an exact likelihood function can be obtained, which can be used to estimate the parameters. We also propose a reparameterization that does not conflict with the stationarity of the time series, which was not taken into consideration by Chib(1993) and Chib and Greenberg(1994). Simulation studies show that our method leads to more accurate inferences.

Key Words:
autoregressive process; exact likelihood; MCMC.

(三)、報告內容

Consider a regression model possessing error terms with mean zero and unknown variance. The assumption that the errors are uncorrelated is generally unrealistic. Violations of independent assumption can be checked by using residual plots, run test, Durbin-Watson test and so on. In most analysis of time series, it is often shown that the covariance matrix of the regression model disturbance terms has a Markov pattern. In this paper, we develop an exact method to analyze time series regression models in a Bayesian framework. To avoid contradicting the stationarity of the time series, reparameterization is required. Detailed description is presented in Section 2. Parameter estimation will be done using the Gibbs sampling and Metropolis-Hastings algorithm. Compared with a simple regression model, the major obstacle for estimating

a time series regression model involves treating its initial terms. Bayesian inference for time series regression regarding autoregressive processes conditional on initial observations has been considered by Chib (1993), McCulloch and Tsay (1994), and Albert and Chib (1993), among others. However, conditioning on the initial terms, the problem setting loses its time series feature completely and 3 becomes a pure regression problem. In this paper, we consider the likelihood that is not conditional on the initial observations. We employ the results of Wise (1955) to obtain an inverse autocovariance matrix in the exact likelihood function. An alternative expression of the exact likelihood function for a pth order autoregressive process can also be found in Box and Jenkins (1976). However, the exact likelihood function in regression form as presented in the paper is more appealing. Chib and Greenberg (1994) developed exact methods to analyze time series model in a Bayesian framework which expressed the time series regression model in state space form. By using the same time series regression models, we will compare our results with those of Chib (1993) and Chib and Greenberg (1994) in a simulation study.

In conclusion, we propose a Bayesian estimation procedure for a simple but most frequently used model in practice, namely the time series regression models. We use the exact likelihood instead of the likelihood conditional on initial observations based on Wise (1955). With the application of the transformation from $\boldsymbol{\varphi}$ to the partial autocorrelations (PACF) $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_p)^T$, the stationarity condition of $\boldsymbol{\varphi}$ becomes $|\eta_i| < 1$, $i = 1, \cdots, p$. Consequently, the proposed procedure will be valid for any order of the AR(p) process. Therefore, there is no difficulty in dealing with the AR(p) model with $p > 3$. We obtained better inferential results on simulations when compared with those results in Chib (1993) and Chib and Greenberg (1994). The results for real data sets show that the fitted models are adequate for the data sets. The proposed methodology can be extended in the following directions.

1. The estimation procedure can be extended to regression with ARMA errors. The exact closed form of the likelihood function of a general ARMA model can be found in Newbold (1974) and Hillmer and Tiao (1979) among others.

2. To allow heteroscedasticity in ε_t, we can assume GARCH-type conditional variance. Time series regression model with GARCH errors has become very common in econometric and nancial research.

3. When we deal with financial data, typical empirical evidence in the literature indicates that ε_t is usually

fat-tailed. In future work, we could consider leptokurtic distributions as distributions for ε_t, such as student t-distribution or generalized error distribution.

（四）、參考文獻

Albert, J. and S. Chib, 1993, Bayesian inference for autoregressive time series with mean and variance subject to Markov jumps, *Journal of Business and Economic Statistics*, **11**,1-15.

Barndor.-Nielsen, O. E., Schou, G., 1973. On the parameterisation of autoregreeive models by partial autocorrelations. J. Multivariate. Anal. 3, 408-419.

Box, G.E.P. and G.M. Jenkins, 1976, Time Series Analysis Forecasting and Control, 2nd ed. (Holden-Day, San Francisco, CA).

Chen, C. W. S. and Wen, Yu-Wen, 2001, On goodness of fit for time series regression models. *Journal of Statistical Computation and Simulation*, **6**, 239-256.

Chib, S., 1993, Bayes regression with autogressive errors: A Gibbs sampling approach, *Journal of Econometrics*, **58**, 275-294.

Chib, S., and E. Greenberg, 1994, Bayes inference in regression model with ARMA(p,q) Errors, *Journal of Econometrics*, **64**, 183-206.

Chib, S. and Greenberg, E., 1995, Understanding the Metropolis-Hastings algorithm, *American Statistician*, **49**, 327-335.

Hastings, W. K., 1970, Monte-Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97-109.

Hillmer, S. C. and G.C. Tiao, 1979, Likelihood function of stationary multiple autoregressive moving average models, *Journal American Statistical Association*, **74**, 652-660.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H., 1953, Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

McCulloch, R. E. and R.S. Tsay, 1994, Bayesian analysis of autoregressive time series via the Gibbs sampler, *Journal of Time Series Analysis*, **15**, 235-250.

Newbold, P., 1974, The exact likelihood function for a mixed autoregressive moving average process, *Biometrika*, **61**, 423-426.

Raftery, A. E. and S.M. Lewis, 1992, How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid

and A. F. M. Smith), 765-776. (Oxford University Press, Oxford).

So, M. K. P. and Chen, C. W. S., 2003, Subset threshold autoregression. *Journal of Forecasting*, **22**, 49-66.

Wise, J.,1955, The autocorrelation function and spectral density function, *Biomertrics*, **42**, 151-159.

（五）、計畫成果自評

本研究成果是貝氏分析的一部份，也與所提之計畫有關。應是篇有價值的產品。