COMPSTAT

Proceedings
in Computational Statistics

18th Symposium Held in Porto,
Portugal, 2008

Edited by
Paula Brito

With 128 Figures
and 66 Tables

Physica-Verlag
A Springer Company

Professor Dr. Paula Brito
Faculdade de Economia
Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto
Portugal
mpbrito@fep.up.pt

ISBN 978-3-7908-2083-6 e-ISBN 978-3-7908-2084-3

DOI 10.1007/978-3-7908-2084-3

Library of Congress Control Number: 2008932061

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

A Robustified MCMC Sampler – Metropolis Hastings Simulator with Trimming

Veit Köppen[1] and Hans-J. Lenz[1]

Institute of Production, Information Systems and OR, Freie Universität Berlin
Garystr. 21, 14195 Berlin, Germany, {*koeppen;hjlenz*}*@wiwiss.fu-berlin.de*

Abstract. One facet of data quality is the integrity of data. Most main business and economic indicators suffer from statistical discrepancies. Such indicators are modeled as random variables and related by a non-linear stochastic system of equations. In order to check integrity of data with respect to a fully specified model consisting of balance equations or equations due to definitions we need the joint distribution of the right hand side of each single equation, and the distribution of the related left hand side. As the Gaussian distribution is not closed under all four arithmetic operations, we need MCMC simulation to determine the probability distributions. In this paper we use the Metropolis-Hastings (MH) method. Various distributions and moments of indicators are simulated. Using the MH method in a classical way imprecise estimates may be caused by large measurement errors of the variables. Consequently, robust estimation becomes mandatory.

Keywords: particle filter theory, Monte Carlo simulation, trimming, Metropolis-Hastings algorithm

1 Introduction

Business indicators are part of many business reports. The same is true for main economic indicators like Gross Domestic Product, inflation rate or rate of unemployment as collected by the national account group of UNO. A vital question is whether such indicators contradict given balance equations or simple definitions. Business processes are related to services and goods, and deliver indicators which are measured. Of course, some integrated and aggregated indicators may be corrupted by errors or must be estimated because of being not directly observable. The same is true for economic indicators which are characterized by an higher aggregation level. Therefore business and economic indicators can be modeled as random variables. Of course, a special case are crisp data where all variances are zero. The system of equations we consider is a non-linear system with arithmetic operators connecting the variables in each equation. A classical system of business indicator is the DuPont-System, which will be investigated here for the sake of simplicity. Other systems may differ in the equations, but can be handled as well. Markov Chain Monte Carlo (MCMC) simulation is a helpful tool to investigate random variables if it is not possible to analytically determine the proper

distribution functions. The Metropolis-Hastings (MH) algorithm can be used to generate the probability distribution of random variables. This method can be easily implemented for instance in R as we did, and also proves to have reasonable computational performance.

2 The DuPont-system of business indicators

In 1919 the chemical company "DuPont" developed a system of business indicators. The equation system is:

- profit = sales - cost
- profit margin = profit / sales
- return on investment (ROI) = profit / capital
- capital turnover = sales / capital.

The two types of variables are: (1) endogenous (explained or left hand) variables: profit, ROI, profit margin, and capital turnover; (2) exogenous (explaining or right hand) variables: sales, cost, and capital. In some applications some of these variables may have missing values, or can only be estimated or measured with large imprecision. Other variables have a restricted range due to a share holder policy. We assume that all equations considered are mathematical separable. Now, as there exist various ways to compute an indicator, the question arises whether the single equation estimates are "model consistent", and how to compute a combined estimate in the sense of a full information procedure. It is evident that the same problem carries over to the main economic indicators of the national accounts system, i.e. UNO-SNA 2008, which have hundreds of variables.

3 Simulation

Computation of the corresponding joint probability function or marginal distributions of a simultaneous non-linear equation system is usually not a trivial task. We use MCMC methods for the sake of generality. Therefore, the very restrictive assumption of a Gaussian distribution family can be relaxed. Because of special features of the MH algorithm any density function can be used, cf. Hastings (1970), Chib (2004). This is specially true for mixed, skewed and heavy tail distributions.

3.1 Extending the Metropolis Hastings algorithm

In the first phase of the MH algorithm the probability functions of the given variables are determined. Furthermore, a proposal distribution is chosen for each exogenous variable. To reduce the sampling cost the shape of proposal should be as close as possible to the desired probability function. If this is

not a priori possible, the sampling size has to be extended caused by the so called burn-in phase. The size of the sample depends upon two parameters: the number of particles and the number of repetitions. The second parameter describes how many simulated means of particles per run will be used to estimate the distribution function of the exogenous variables. In the second phase all exogenous variables are simulated using MH algorithm. In the third phase distributions of the endogenous variables are estimated using only those equations where the corresponding exogenous variables are sampled. The number of iterations for these two phases is kept as a parametric constant. At the end of a MCMC simulation experiment a sample for all variables is generated. Therefore estimates of moments or densities can be derived, tests can be performed and given indicators can be checked.

Extended MH algorithm:

Experimental set-up: Fix the repetition size and the number of particles.

1. Initialize the exogenous variables with proposal distributions and target probability functions.

2. Draw samples from exogenous variables using MH algorithm.

3. Derive distribution of endogenous variables from equation system.

4. Compute the means and variances of all variables.

5. If the repetition size is not reached, go to 2.

3.2 Evaluation of the algorithm

An evaluation of the extended MH algorithm requires to analyze whether or not the quality of the simulation depends upon artifacts. We consider this step as a kind of "calibration". This primarily concerns non linearity due to products and quotients as well as non-normality. The following estimators are used for the mean and the variance of the simulated data, where T describes the sample size of the simulated data:

$$\hat{\mu} = 1/T \sum_T X_i \qquad \hat{\sigma}^2 = 1/(T-1) \sum_T (X_i - \hat{\mu})^2$$

As estimators for the triangular distribution parameters we use:

$$\hat{l} = 1/T \sum_T min(X_i) \qquad \hat{p} = 1/T \sum_T \hat{\mu} \qquad \hat{u} = 1/T \sum_T max(X_i)$$

In all of our simulation experiments the simulation uses 1000 particles (sampled values) per experiment and each experiment is repeated 5000 times.

3.3 Single linear equation with Gaussian and non-Gaussian distributed variables

We use the single linear equation $sales = profit + cost$. The exogenous random variables are profit and cost, and the endogenous random variable is

sales. Firstly, we consider the joint Gaussian distribution mainly for comparison purposes. Computation of theoretical mean is done by $E[X_1 \pm X_2] = E[X_1] \pm E[X_2]$ and variance by $Var[X_1 \pm X_2] = Var[X_1] + Var[X_2]$, of course, under the independence assumption. The results for the theoretical and simulated estimates are given in Table 1.

distribution	μ	σ^2	$\hat{\mu}$	$\hat{\sigma}^2$	l	p	u	$\hat{l}$	$\hat{p}$	$\hat{u}$
(i)	100	17	100	17						
(ii)	100	61.85	100	67.40	95	100	105	95.12	100	104.81
(iii)	100	67.45	99.62	67.42						

Table 1. Results of endogenous variable in linear equation case.

(i) Gaussian distribution

We assume Gaussian probability functions and specify the parameters as follows: $profit \sim N(20, 1^2)$ and $cost \sim N(80, 4^2)$. The proposal distribution is the Gaussian distribution with the same parameters, too. By running the algorithm estimated means and standard deviations of the exogenous variables do not differ from the desired levels. As the variables are Gaussian distributed, the sum of two Gaussian random variables is also a Gaussian distribution. Thus a Kolmogorov-Smirnov (KS) goodness of fit test can be used based on the simulated data of *sales*. The H_0-hypothesis of all tests is that the simulated data is corresponding to a Gaussian distribution with mean 100 and variance 17. Mean of p-values from simulations delivers $\overline{p} \approx 50.4\%$. We infer that the simulation at least for uncorrelated Gaussian distributed variables in the linear case is a good choice.

(ii) Triangular distribution

The next example uses the same equation with a triangular probability function. $Profit$ is distributed in the interval [19, 21] with peak (mode) at 20. *Cost* are distributed in the interval [76, 84] with peak at 80. The proposal distributions are now a uniform distribution of $profit$ in the interval [19, 21] and of *cost* in the interval [76, 84]. Testing of the simulated exogenous variables against the theoretical distribution the p-value of a two-sided KS test has a value of about 0.

Fig. 1 left upper part makes clear that simulations are insufficient near to the upper and lower bounds. The simulated densities and distribution functions are compared with their theoretical functions in Fig. 1.

(iii) Contaminated Gaussian distributions

In our third example $profit$ and *cost* are now corresponding to an ϵ-contaminated Gaussian distribution. This probability function is described by:

$$(1 - \epsilon) \cdot N(\mu_1, \sigma_1^2) + \epsilon \cdot N(\mu_2, \sigma_2^2) \text{ with } \epsilon \in [0, 1]. \quad (1)$$

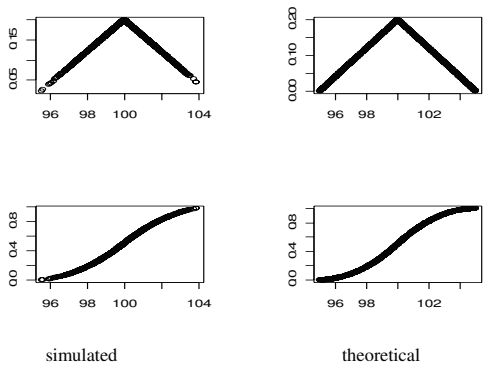

Fig. 1. Simulated and theoretical triangular distributions.

profit: $\mu_1 = 21\ \sigma_1^2 = 2\ \mu_2 = 16\ \sigma_2^2 = 1\ \epsilon = 0.1$
cost: $\mu_1 = 75\ \sigma_1^2 = 4\ \mu_2 = 90\ \sigma_2^2 = 3\ \epsilon = 0.3$

Our proposal distribution is a Gaussian distribution and the starting values for the MH algorithm are uniformly distributed in the interval [10, 30] for profit and [70, 100] for *cost*. The theoretical value of sales is well supported by the simulated variance as easily seen in Tab. 1. The KS test can not be applied, because the distribution function of the sum is not known.

As a first conclusion we note that simulation of a linear equation system using the MH algorithm leads to good results not only for endogenous variables, but also for exogenous variables. Some problems may arise if the variance is very large, since it affects the simulated data. A proposal of tackling this problem by robustification is given in the last section.

3.4 Single nonlinear equation with Gaussian and non-Gaussian distributed variables

As an example of a non-linear equation we take from the DuPont system: $profit = ROI \cdot capital$. The exogenous variables are ROI and $capital$. $Profit$ is here the endogenous variable. The theoretical estimates for the endogenous variable can be computed under independence assumption as (Mood (1973)):

$$E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2],$$
$$Var[X_1 \cdot X_2] \approx E^2[X_1] \cdot Var[X_2] + E^2[X_2] \cdot Var[X_1] + Var[X_1] \cdot Var[X_2]$$

The theoretical and simulated results for the endogenous variable $profit$ are shown in Table 2.

(i) Gaussian distribution

In the case of a Gaussian probability function the variables are distributed as: $ROI \sim N(0.25, 0.025^2)$ and $capital \sim N(80, 0.4^2)$. To reduce computation

distribution	μ	σ^2	$\hat{\mu}$	$\hat{\sigma}^2$	l	p	u	$\hat{l}$	$\hat{p}$	$\hat{u}$
(i)	20	5.01	19.99	4.99						
(ii)	20	0.275	20	0.215	18.24	20	21.84	18.29	20	21.77
(iii)	20	6.19	20.03	6.78						

Table 2. Results for endogenous variable in case of a non-linear equation.

effort the proposal distributions are Gaussian distributions with same parameters. A KS test for all simulated particles per run has a mean of $\overline{p} = 0$. The underlying distribution is a normal distribution with the theoretical mean and variance. Because of the zero values of these two-sided tests, the H_0-hypothesis that the simulated data have a Gaussian distribution must be rejected.

(ii) Triangular distribution

In case of a triangular distribution, ROI is symmetrically distributed in the interval [0.24, 0.26] and capital is also symmetrically distributed, however, in the interval [76, 84]. The proposal distribution is a uniform distribution. As before the simulated data differs clearly from the theoretical distribution at the boundaries. If the exogenous variables are described by a triangular probability function, the simulated data for the endogenous variable should follow a triangular distribution. The two-sided KS test has a mean in the p-values of 0. Thus the Hypothesis, that $profit$ is described by a symmetrically triangular distribution in the interval [18.24, 21.84] is rejected.

(iii) Contaminated Gaussian distribution

In the next example variables ROI and $capital$ are distributed as a two-peak Gaussian which is described by equation 1:

ROI: $\mu_1 = 0.21\ \sigma_1^2 = 0.05^2\ \mu_2 = 0.31\ \sigma_2^2 = 0.06^2\ \epsilon = 0.4$

capital: $\mu_1 = 71\ \sigma_1^2 = 10^2\ \mu_2 = 86\ \sigma_2^2 = 8^2\ \epsilon = 0.6$

The proposal distribution for each variable is a Gaussian distribution. Due to the unknown distribution function of the product a KS test can not be applied.

4 Variance-reduction techniques by trimming

The problem of too large deviations caused by MCMC simulation can be reduced by a robust estimation procedure. To achieve robustness one possible solution is to eliminate the extreme values at both ends from the sample. A solution is the γ-trim mean and γ-trim variance, for instance cf. Büning (1991). This implies to drop the $\gamma \cdot R$ upper and lower sampled values, where $0 \leq \gamma < 0.5$. The estimators change to:

$$\hat{\mu}_\gamma = \frac{1}{R(1-2\cdot\gamma)} \sum_{i=\gamma\cdot R+1}^{R(1-\gamma)} X_i \qquad \hat{\sigma}^2 = \frac{1}{R(1-2\cdot\gamma)-1} \sum_{i=\gamma\cdot R+1}^{R(1-\gamma)} (X_i - \hat{\mu}_\gamma).$$

Because the sample size is reduced, sampling should be extended. The amount of post sampled particles depends on the target sampled size and the γ-trim. The relationship can be described by: sampling particles = demanded particles / $(1 - 2 \cdot \gamma)$. The γ parameter is dependent on the variance of the variable. If the variance is high, γ should be increased. On the other hand, the parameter should be set to 0 for a low variance. This will reduce the sampling effort and, consequently, speed-up the algorithm.

5 Simulation of a non-linear stochastic system of equations

The following example illustrates, that not only a single equation but a full system of equations can be simulated by our algorithm with trimming. The exogenous variables are capital, cost and sales and they are distributed as follows: $capital \sim N(80, 4^2)$, $cost \sim N(80, 4^2)$ and $sales \sim N(100, 5^2)$. To keep the simulation size small, the proposal distribution for the endogenous variables is also a Gaussian distribution. For simulation the DuPont system is used. Fig. 2 shows the distributions of mean and standard deviation of all variables. The sampling is done with γ equals to 0 (no trimming). Fig. 3 shows the results for γ-trimmed estimation. The value of γ is set to 0.3. The shrinkage of the spread of all distributions becomes evident, cf. Fig. 2 and 3.

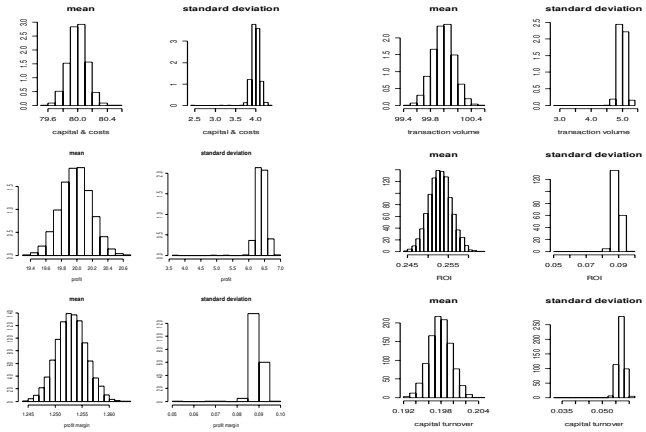

Fig. 2. Simulated means and standard deviations of the DuPont system ($\gamma = 0$).

6 Results and future work

We showed that simulation based upon the MH algorithm is a sound method to simulate simultaneous equation systems. The inputs are the system of

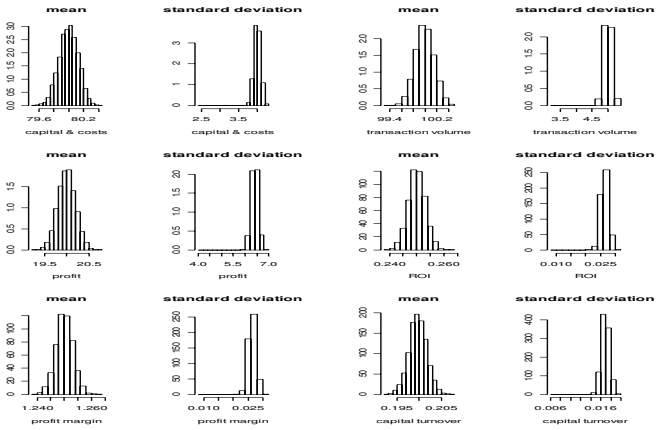

Fig. 3. Simulated means and standard deviations of γ-trimmed DuPont System ($\gamma = 0.3$.)

equations and the probability functions of the exogenous variables. All other variables can be simulated. If the sampling size is adequate, the sampling distributions are similar to the theoretical distributions. Most of those theoretical distributions are not analytically derivable, thus sampling is the only way to solve such systems. The improvement of using random variables instead of crisp quantities for main business and economic indicator systems is obvious. The often published "statistical discrepancies" related to variables become crucial if measurement errors are effective.

The results of our simulation study of non-linear equation systems are: (1) MH algorithm is a very flexible method to sample from any joint probability function. (2) A critical drawback is the sampling from a probability density function defined on a finite domain. (3) For large variances of variables the classic MH algorithm should be modified to a MH with trimming. The simulation size must accordingly be adapted. The pay-offs for increased simulation are improved estimators. However, further problems exist like missing values as well as stochastic dependencies between variables.

References

BÜNING, H. (1991): *Robuste und adaptive Tests*. Walter de Gruyter, Berlin.

CHIB, S. (2004): Markov Chain Monte Carlo Technology. In: J.E. Gentle, W. Härdle and Y. Mori (Eds.): *Handbook of Computational Statistics, Concepts and Methods*. Springer, Berlin, 71–102.

HASTINGS, W.K. (1970): Monte Carlo sampling method using Markov Chains and their applications. *Biometrica 57 (1), 97–109*.

KÖPPEN, V., HAUSMANN, A., and LENZ, H.-J. (2005): Simulation - A Support for Controllers Decision Process. In: *Proceedings of ICTM 2005: "Challenges and Prospects"*. Multimedia University, Melakka, 1155–1170.

MOOD, A.M., GRAYBILL, F.A., and BOES, D.C. (1973): *Introduction to the theory of statistics*. 3rd edition. McGraw-Hill, Tokyo.