Simulation of Non-linear Stochastic Equation Systems

Veit Köppen[1], Hans-J. Lenz[2]

(Freie Universität Berlin)

Abstract

In this paper, we combine data generated by Monte-Carlo simulation, which is based on prior knowledge about a fully specified non-linear, stochastic balance equation system with noisy measurements. This two-step estimation procedure strongly improves the precision of the estimation of unobservable quantities of such models.

Keywords: Metropolis-Hastings-Algorithm, Monte-Carlo simulation

1 Introduction

We consider non-linear equation systems with measurement errors in the variables. The variables are simply related by arithmetic operators; however, multiplication and addition imply non-linearity. Various techniques, like general least squares, fuzzy equation solver or Monte-Carlo simulation techniques can be applied to estimate the unobservable variables (”unknown system parameters”), cf. [3, 4]. The Metropolis-Hastings-Algorithm (MH Algorithm) has been shown to be an efficient algorithm to sample from any density function [2, 1]. We introduce a specially tailored algorithm, which fuses the information from the simulation and measurements in order to improve the precision of the estimators of the unobservable variables.

2 Simulation

Simulation of a non-linear stochastic equation system assists data validating and cleansing given a fully specified model of a system. The solution set can be used to

[1] e-mail: koeppen@wiwiss.fu-berlin.de

[2] e-mail: hjlenz@wiwiss.fu-berlin.de

decide whether or not the measured data can be generated from the model, under the premises, that the underlying model is correct. As most of the equations represent balance equations or definitions, this assumption is rather conclusive.

2.1 System Model

Let $\xi = (\xi_1, \ldots, \xi_p)$ be a vector of true but unobservable state variables. The vectors $x = \xi + v$ and $z = H(\xi) + w$ can be observed, whereas the errors ν in the variables and the errors in the model w with dimension q are unobservable (latent), and distributed according to $u = \begin{bmatrix} v \\ w \end{bmatrix} \sim F(\cdot)$, where F is the corresponding distribution function, for example a Gaussian distribution $N(0, \Sigma_{uu})$. The balance equation model is defined by $\zeta = H(\xi)$. We make the assumptions that $u \perp v$, H is measurable, and $q \leqslant p$, cf. [3].

The density functions f_z, f_x of the vectors z, x are given. A necessary condition about the density functions is, that they must fulfil the Lebesgue L2 norm, i.e. $\sqrt{\int |f_x(x)|^2 \, dx} < \infty$. In the following it will become clear, that this is mandatory, because the used transformation of the joint density function should lead to a density function again and should have a well-defined Lebesgue integral as well. A sufficient condition is that the density function f_x is bounded. The same holds true for f_z. Alternatively, the model can be described by a model graph $\mathbf{G} = (\mathrm{N}, \mathrm{A})$, where the set $\boldsymbol{N}$ of nodes represents the set of unobservable or observable variables and the set $\boldsymbol{A}$ represents relations, cf. Fig. 1.

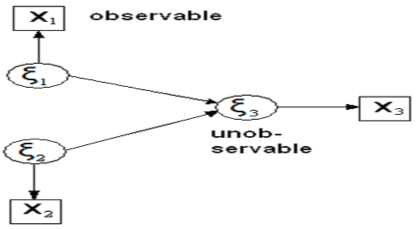

Figure 1. Model Graph G

In Figure 1 the model graph of $\zeta_3 = f_3(\xi_1, \xi_2)$ is shown. In this case ζ_3 is a LHS variable, ξ_1 and ξ_2 are RHS variables. We study equation systems which are separable, i.e. there exist mappings f_1^{-1},f_2^{-1}, such that$\xi_1 = f_1^{-1}(\xi_2, \xi_3)$, $\xi_2 = f_2^{-1}(\xi_1, \xi_3)$.

Our objective is to estimate the vectors ξ and ζ, given data (x, z), and to make inferences about relations involved. To illustrate our approach, we present below the well-known DuPont system. It consists of four equations, which includes seven variables. Note, that all variables are noisy of various degrees. Of course, error-free measurements are a rather special case of our assumptions.

The DuPont System is described below:

P	$=$	TV	$-$	CO	(1)	profit $= P$	transaction volume $= TV$
ROI	$=$	P	$/$	CA	(2)	costs $= CO$	capital turnover $= CT$
PM	$=$	P	$/$	TV	(3)	capital $= CA$	return on investment $= ROI$
CT	$=$	TV	$/$	CA	(4)		profit margin $= PM$

Evidently, the DuPont system is separable. For example, equation (1) can be rearranged as $TV = P + CO$ or $CO = TV - P$.

2.2 Sampling

In the first step we are sampling from each distribution. The MHA can be used to generate samples from any density function involved.

```
Algorithm $MH$
  input:      $f$ target function,
              $q(\cdot,\cdot)$ transition kernel
              $t$ iteration index; $T$ maximum number of iterations
  output:     $s$ statistic, from which the estimates $\hat{f}$ are derived

1. initialise $s_0$ and set $t := 1$

2. repeat
     incr($t$)
     sample $\varphi$ from $q(\theta_{t-1}, \cdot)$
     evaluate $\alpha(s_{t-1}, \varphi) := \min\left(1, \frac{f(\varphi)\cdot q(\varphi, s_{t-1})}{f(s_{t-1})\cdot q(s_{t-1}, \varphi)}\right)$,
     if $\alpha$ accepted then $s_t := \varphi$ else $s_t := s_{t-1}$.

3. until $t = T$.
```

In a second step, the full equation system is used in order to simulate the (joint) distribution of the vector z. This step produces an estimate of the unobservable variables.

2.3 Consistency Check

When $n > 1$ estimates exist for a given variable, it can be checked whether or not the measurement fulfils the balance equation system.
If the variable shows up in $n \leqslant q$ equations, α-quantiles for all n simulated density functions of a variable are computed.

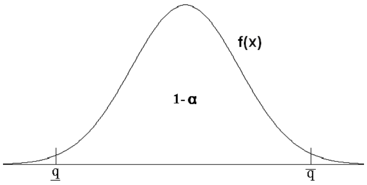

Figure 2. α-quantile

If the maximum of the lower quantilesmax $\left\{\underline{q_1}, \underline{q_2}, \ldots, \underline{q_n}\right\} = \underline{q_{\max}}$is greater than the minimum of the upper quantilesmin $\{\overline{q_1}, \overline{q_2}, \ldots, \overline{q_n}\} = \overline{q_{\min}}$, the data set should be marked as inconsistent. Otherwise go to the next step (”projection ”).

2.4 Projection and Distribution on Subspace

If at least one simulated distribution and a prior distribution per variable is at hand, the two-dimensional space range $(x_{obs}) \times$ range (ξ_{sim}) is projected on the subspace $x_{obs} - \xi_{sim} = 0$. We assume that the samples are independently generated. Therefore the joint distribution function can be easily derived. This assumption is plausible, since the simulation produces the estimates from different equations. The range of interest of the projection is given by $I_q = \left[\overline{q_{\min}}, \underline{q_{\max}}\right]$. The joint distribution is given by:

$\hat{f}_{x\overset{\wedge}{\xi}}(y,y) = \text{c } \hat{f}_x(y) \cdot \hat{f}_{\overset{\wedge}{\xi}}(y) \quad \forall y \in I_q, c \in \boldsymbol{R}_+.$ (5)

$c \in \boldsymbol{R}_+$ is a normalising constant. As each of the density functions in (5) are assumed to be $L2$ normal, the product has again a Lebesgue integral and normalisation is feasible.

3 Sampling and Projection - Algorithm SamPro

```
Input:     model $\mathbf{G} = (\mathrm{N}, \mathrm{A})$, $f_x$, $f_z$ and data set$(x, z)$
Output:    $\left(\hat{\xi}, \hat{\Sigma}_{uu}\right)$
begin
           1. resolve ($\equiv$ set LHS & RHS) for all variables of each equation
           2. compute$\hat{f}_x$by sampling from the joint density of all RHS
              variables
           3. derive the joint distribution $\hat{f}_z$ for $z$
           4. estimate quantiles $\underline{q_{\max}}$, $\overline{q_{\min}}$
           5. compute the distribution $\hat{f}_{xz}$ on the subspace $x - z = 0$
end
```

4 An Example – The DuPont System

In our example below, 5 out of seven random variables are fully specified by a Gaussian density function coined $N\left(\mu,\sigma^2\right)$:

- •P ∼ $N\left(30,5^2\right)$ •TV ∼ $N\left(100,25^2\right)$ •ROI ∼ $N\left(0.4,0.2^2\right)$
- •CA ∼ $N\left(80,20^2\right)$ •PM ∼ $N\left(0.25,0.1^2\right)$

Costs and capital turnover can be calculated by the equation system (1) – (4). Furthermore, profit, transaction volume and profit margin can be computed by the equation system. The densities derived by *SamPro* are shown in figure 3. Note the increase of accuracy due to reduced variances of estimates relative to the measurement variances.

In figure 4 the densities of profit simulated by SamPro and the observed density are shown. Variance reduction and a shift in the mean are evident.

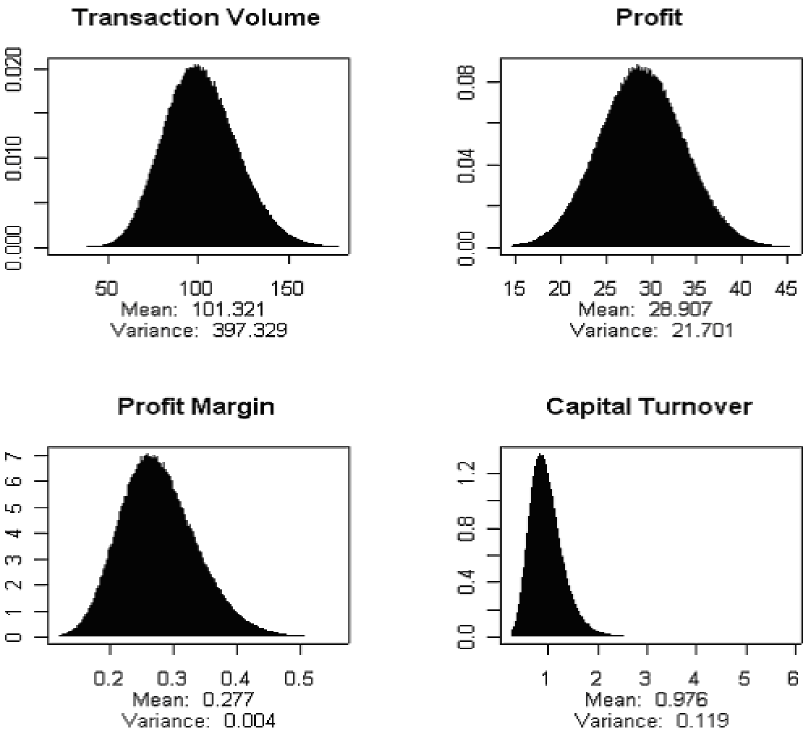

Figure 3. Simulated densities of four variables of the DuPont System

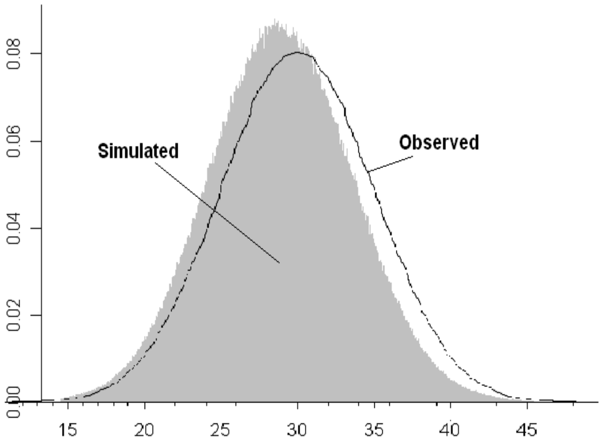

Figure 4. Observed and simulated profit density functions

5 Conclusion

The algorithm SamPro supports an evaluation of stochastic non-linear balance equation systems, which are equally important for business and science. Information fusion leads to significantly improved estimates.

References

1. Gamerman D. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference.* Chapman & Hall, Boca Raton, 2002.

2. Hastings W.K. *Monte Carlo Methods using Markov Chains and their Applications.* Biometrika, 1970, v. 57, p. 97–109.

3. Lenz H.-J., Rödel E. *Statistical quality control of data.* Operations Research '91. Physica-Verlag, Heidelberg, 1992, p. 341–346.

4. Lenz H.-J., Müller R. *On the Solution of Fuzzy Equation Systems.* G. Della Riccia, R. Kruse, and H.-J. Lenz (eds.): Computational Intelligence in Data Mining, Springer, New York, 2000, p. 95-110.