Work in progress – please don't cite without authors' permission

Bayesian Estimation of Input-Output Tables for Russia

Oleg Lugovoy [a,b] Andrey Polbin [b] Vladimir Potashnikov [b]

Conference paper

Abstract

The paper pursues two goals. First, to apply Bayesian statistics for updating IO tables for 1996-2004 period, i.e. within "old" definition of industries. Second, to estimate IOT for 2004-2010, in new definition of activities, based on national accounts and industries-level data. Both goals are experimental since as we know Bayesian statistics is not yet in common use here. However, we believe, that this approach has several advantages over R.A.S. and Maximum entropy methods. First, it is a natural and flexible way to incorporate any kind and amount of information either as a prior distribution or observable data. Second, Bayesian methods provide full density profile on estimated parameters with covariates. And third, from computational perspective minimizing highly dimensional function with hundreds of parameters, like the cross entropy measure, might be much harder than evaluation of posterior distribution using modern sampling algorithms, such as Markov Chain Monte Carlo methods. Comparison of performance of various methods will be provided.

[a] Environmental Defense Fund, USA
[b] The Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia

Introduction

Russian statistical system is under transition for almost two decades from Soviet type Material Product System to SNA. The main transitional break in methodology took place in 2004-2005 when Russian statistical agency "Rosstat" started reporting based on the new definition of economic sectors consistent with NACE[§], and stopped reporting using definition of activities inherited from the Soviet statistical system. This methodological break splits all industry level statistics into two periods with little consistency between each other. As a result, Rosstat stopped updating IOT in 2003, based on the only benchmark survey conducted in 1995. The next survey is scheduled for 2012 with expected publication of results in 2015 or later. Official backward estimation is not expected. Therefore Russian statistics will miss IOT at least from 2004 to 2010. Also quality of officially updated IOT from 1996 to 2003 based on 1995 benchmark is questionable.

…

The paper is organized as follows. First we discuss a conceptual framework for updating IOT using Bayesian statistics, testing the methodology on artificial data in comparison with RAS and Maximum entropy methods, and applying it for updating Russian IOT for 2003. In the following section we discuss a possibility to estimate IOT for 2004-2010 years based on available information from National accounts, sequentially introducing additional information (constrains) to the estimation process.

Updating IOT with Bayesian methods

In this section we discuss a methodology for updating IOT using Bayesian framework and Monte Carlo Markov Chains method as alternative to RAS and maximum entropy methods.

[§] Statistical classification of economic activities in the European community (in French: Nomenclature statistique des activités économiques dans la Communauté européenne).

Conceptual framework

The basic problem of updating an IO matrix or more generally a SAM can be formulating as follows: find an unknown IO matrix with known sums of rows and columns and a known IO matrix for some previous year. Mathematically speaking, we need to find a matrix *A* with following restrictions:

$$Y = AX,$$
$$\sum_i a_{i,j} = \bar{a}_j, \quad a_{i,j} \geq 0 \tag{1}$$

where *Y*, *X* are known vectors and $\bar{a}_j$ are known sums of columns. And there is a known matrix A^0 from previous year. In classical framework the solution of this problem is usually reduced to finding such matrix *A*, which minimize some distance function from known matrix A^0 under system of restrictions (1).

In this paper we propose to follow Bayesian methods in estimation of IO tables. Bayesian approach provides a natural and flexible way to incorporate any kind and amount of information either as a prior distribution or observable data. Bayesian methods also provide full density profile on estimated parameters with covariates.

In Bayesian econometrics it is assumed that a researcher has some prior beliefs about estimated parameter vector θ before observing the data, which could be summarized by prior density function $p(\theta)$. When new data comes the researcher update the beliefs about parameters according Bayes theorem:

$$p(\theta | Y) = \frac{L(Y | \theta) p(\theta)}{\int L(Y | \theta) p(\theta) d\theta} \propto L(Y | \theta) p(\theta) \tag{2}$$

where $p(\theta | Y)$ is the posterior density and $L(Y | \theta)$ is the likelihood.

Bayesian inference is easy since the posterior density contain all the information one may need. The researcher could be interested in point estimate, credible set and correlation of parameters and construct it from posterior distribution. In Bayesian framework point parameter estimate is chosen to minimize expected loss function with expectation taken with respect to the posterior distribution. The most common loss function used for Bayesian estimation is the mean square error and the corresponding point parameter estimate is simply the mean of the posterior distribution.

Despite the attractiveness of this method, in the past, Bayesian inference was not so popular due to numerical integration needed in equation (2). In some cases when the prior on θ is conjugate with posterior on θ the posterior density can be obtained analytically. But in more general setup we know posterior density up to normalizing constant. Recently developed computer-intensive sampling methods such as Monte Carlo Markov Chain (MCMC) methods have revolutionized the application of Bayesian approach. MCMC methods are iterative sampling methods that allow sampling from posterior distribution $p(\theta | Y)$.

Heckelei *et al. (*2008) shortly discuss IOT update with Bayesian method and give an example on artificial data. In this paper authors present a Bayesian alternative to the cross-entropy method for deriving solutions to econometric models represented by undetermined system of equation. In the context of balancing an IO matrix they formulate posterior distribution in the following way:

$$p(z | data) \propto I_{\Psi}(z) p(z) \tag{3}$$

$$z = vec(A) \tag{4}$$

Equation (4) means vectorization of matrix *A*. In equation (3) $p(z)$ is some prior distribution, $p(z | data)$ is the posterior distribution and $I_{\Psi}(z)$ is the indicator function that assigns weights of 1 if *z* satisfies the constraints (1) and 0 otherwise. Authors interpret the indicator function as the likelihood function. As estimate of *z* Heckelei *et al.* (2008) consider mode of posterior distribution which could be found with some optimization routine. And they illustrate proposed method balancing small 4x4 matrix with independent normal prior taking A^0 as prior mean.

But proposed by Heckelei *et al.* (2008) method actually reduced to minimization yet another distance function from known matrix A^0. In this paper we concentrate on finding full density profile of posterior distribution with MCMC techniques and applying it to real data.

For convenience we consider equality and inequality constraints of the system of restriction (1) separately. Inequality constrains could be simply introduced in prior distribution by assigning 0 value of density in inadmissible domain. For example one could specify independent truncated normal distribution between 0 and 1 for each parameter of the matrix *A*. On the other hand if we have certain beliefs about some parameters we could introduce it as additional linear equality constraints. For example it is convenient to assign 0 values for elements of unknown matrix *A* if corresponding elements in the matrix A^0 are zeros.

At the next step let us consider linear equality constraints and rewrite it in the following form:

$$Bz = T \tag{5}$$

where *B* is the known matrix, *T* is the known vector and $z = vec(A)$ is the unknown vector of estimated parameters. System (5) represents undetermined linear system of equations. And from linear algebra it is known that any solution of linear system (5) could be written in the form:

$$z = \tilde{z} + F^{(1)}\xi^{(1)} \tag{6}$$

where $\tilde{z}$ is the particular solution of the system (5) and $F^{(1)}$ is the fundamental matrix of solutions of homogeneous system $Bz = 0$. And any vector $\xi^{(1)}$ solves system (5). The particular solution and the fundamental matrix could be obtained by Gaussian elimination algorithm.

Columns of the fundamental matrix $F^{(1)} = [f_1^{(1)},..,f_k^{(1)}]$ represent basis of the Euclidean subspace. At the next step we could find the basis of the orthogonal complement of this subspace $F^{(2)} = [f_1^{(2)},..,f_{n-k}^{(2)}]$. Let us consider linear transformation of the original space:

$$\begin{bmatrix} \xi^{(1)} \\ \xi^{(2)} \end{bmatrix} = \begin{bmatrix} F^{(1)} & F^{(2)} \end{bmatrix}^{-1} (z - \tilde{z}) \tag{7}$$

In the new system of coordinates prior density has the following form:

$$p_{\xi}(\xi) = \det\left[F^{(1)}\ F^{(2)}\right] p_Z(\tilde{z} + F^{(1)}\xi^{(1)} + F^{(2)}\xi^{(2)}) \tag{8}$$

If we specify posterior distribution in the form (3) than posterior distribution will be the conditional distribution of random vector $\xi^{(1)}$ given the zero value of the random vector $\xi^{(2)}$:

$$p_{\xi}(\xi \mid data) = p_{\xi^{(1)}|\xi^{(2)}}(\xi^{(1)} \mid \xi^{(2)} = 0) \tag{9}$$

If prior distribution is multivariate normal distribution, posterior distribution of vector $\xi^{(1)}$ is also multivariate normal and we could compute posterior mean and covariance matrix analytically. But it doesn't guarantee nonnegative values of estimated matrix *A*. In general setup we use truncated prior distribution and know posterior density up to normalizing constant. To conduct inference about parameters we approximate posterior distribution (9) applying MCMC sampling methods. After generating the sample of vectors $\xi^{(1)}$ we could move to initial space using formula (6) and obtain the sample of vectors *z*, which represents elements of unknown matrix *A*.

To obtain sample from posterior distribution for examples in this paper we perform the Metropolis sampling algorithm, which is a special case of a broader class of Metropolis-Hasting algorithms, and apply a standard single-site updating scheme. As a proposal density for generating candidate parameter values we use normal distribution for each parameter of vector $\xi^{(1)}$. Standard deviations of the proposal density are iteratively selected during adaptive phase to guarantee acceptance rate for each parameter to be between 35 and 45 percent.

Monte Carlo Experiments

To illustrate the proposed Bayesian method for updating IO matrices in this section we perform Monte Carlo experiments and compare results with the RAS and the cross-entropy methods. In Bayesian framework we assume that there exist several additional known matrices $A^{-1},..,A^{-T}$ from all previous years and they could provide additional information for the estimation purpose. The main hypothesis of our set up is that incorporating additional information in Bayesian framework about variation of IO coefficients in time could improve estimate of unknown coefficients. This information could shed light on relative stability of IO coefficients.

To perform Bayesian method we need to specify some prior distribution for parameters and we assume independent truncated normal distributions for each IO coefficient and use coefficients of last known IO table as prior mean. To specify standard deviations in prior distribution we estimate standard deviation for each coefficient from all available matrices $A^0, A^{-1},..,A^{-T}$ with a following formula:

$$\sigma_{i,j} = \sqrt{\frac{1}{T}\sum_{t=-T}^{0}\left(a_{i,j}^{t} - \frac{1}{T+1}\sum_{t=-T}^{0} a_{i,j}^{t}\right)^2} \tag{10}$$

So in Monte Carlo simulations we would assume that we don't know the true data generating process and apply the same procedure for all data sets. And for robustness of results we experiment with different stochastic processes for IO coefficients.

We perform 10 000 Monte Carlo experiments. Monte Carlo are carried out by randomly generating (data generating process would be described later) of six 4×4 matrices $A^{-4},..,A^0,A$. Than we generate randomly vector *X* and compute vector *Y* from equation (1). In the next step we treat IO matrix *A* as unknown and estimate it with known vectors *X*, *Y* and matrices $A^{-4},..,A^0$. Bayesian method is performed by assuming independent normal distribution for each parameter as prior distribution with A^0 as prior mean and estimated standard deviation for each coefficient from matrices $A^{-4},..,A^0$ as prior standard deviation.

To compute posterior distribution of coefficients we apply Markov chain Monte Carlo (MCMC) method with one chain and sampled length of 50 000 simulation. In each experiment we also estimate matrix *A* with the RAS and the cross-entropy methods using A^0 as prior information about unknown coefficients. To compare relative performance of methods we need to specify some measure of closeness between true coefficients of matrix *A* and its estimated values. In contrast to the RAS and the cross-entropy method, Bayesian approach provides full probability profile of estimated parameters and as point estimates of coefficients we choose the mean of posterior distribution. To compare results for each of the methods in each Monte Carlo experiment we compute following statistics:

1. Root mean square error:

$$RMSE = \sqrt{1/16\sum_{i=1}^{4}\sum_{j=1}^{4}\left(a_{ij} - \hat{a}_{ij}\right)^2} \tag{11}$$

2. Mean absolute error:

$$MAE = 1/16\sum_{i=1}^{4}\sum_{j=1}^{4}\left|a_{ij} - \hat{a}_{ij}\right| \tag{12}$$

3. Mean absolute percentage error:

$$MAPE = 1/16\sum_{i=1}^{4}\sum_{j=1}^{4}\left|\frac{a_{ij} - \hat{a}_{ij}}{a_{ij}}\right| \tag{13}$$

And now we describe data generating process. For robustness of results we experiment with stationary and nonstationary processes for IO matrix coefficients. The main assumption in data generating is that there are different variances of error terms for coefficients stochastic processes. Coefficients are generated only for the first three rows of the matrix A^t and the last element is computed as $a_{4,j}^t = 1 - \sum_{i=1}^{3} a_{i,j}^t$. And if one of the coefficients falls out the boundary restrictions we treat the current experiment as unsuccessful and through it out. We experiment with following data generating procedures:

1. Independent IO coefficients. We assume that each coefficient $a_{i,j}^t$ has normal distribution $a_{i,j}^t \sim N(m_{i,j}, \sigma_{i,j}^2)$. At the first step of generating data in one Monte Carlo experiment we generate mean parameters $m_{i,j}$ from uniform distribution $U[0, 0.5]$. Than we compute standard deviation as $\sigma_{i,j} = k_{i,j} m_{i,j}$, where $k_{i,j}$ are random variables from uniform distribution $U[0, 0.05]$. So we assume that standard deviations of simulated coefficients are not greater than 5% of its value. And at the last step we generate $a_{i,j}^t$ from the distribution $N(m_{i,j}, \sigma_{i,j}^2)$.

2. AR(1) process for IO coefficients. We assume that each coefficient $a_{i,j}^t$ are from the following stationary process:

$$a_{i,j}^t = (1-\rho)m_{i,j} + \rho a_{i,j}^{t-1} + \varepsilon_{i,j}^t, \quad \varepsilon_{i,j}^t \sim N(0, \sigma_{i,j}^2) \tag{14}$$

For generating parameters $m_{i,j}$ and $\sigma_{i,j}$ we apply the same procedure as in the previous point. For simplicity we fix parameter $\rho = 0.7$ and the initial conditions for the $a_{i,j}^{-4}$ equal to the unconditional mean $m_{i,j}$. And at the last step we generate $\varepsilon_{i,j}^t$ from the distribution $N(0, \sigma_{i,j}^2)$.

3. Random walk process for IO coefficients. We assume that each coefficient $a_{i,j}^t$ are from the random walk process:

$$a_{i,j}^t = a_{i,j}^{t-1} + \varepsilon_{i,j}^t, \quad \varepsilon_{i,j}^t \sim N(0, \sigma_{i,j}^2) \tag{15}$$

At the first step we generate the initial conditions for the $a_{i,j}^{-4}$ using the same procedure for generating $m_{i,j}$ in previous points and analogically generate values for standard deviations $\sigma_{i,j}$. And finally generate realizations $\varepsilon_{i,j}^t$ from the distribution $N(0, \sigma_{i,j}^2)$.

Table 1 summarizes relative performance of the Bayesian method in comparison with the RAS and the cross-entropy methods. Results indicate that the Bayesian approach outperforms the competing methods according to the introduced measures of fit. In the case of the independent process Bayesian method wins the RAS and the cross-entropy methods in 70-80 percent of simulations and in the case of random walk wins in 60-70 percent of simulations.

Table 1. Relative performance of Bayesian method.

	Independent process		AR(1) process		Random walk	
	Entropy	RAS	Entropy	RAS	Entropy	RAS
RMSE	72.2%	73.2%	67.3%	67.8%	62.0%	63.0%
MAE	76.0%	77.8%	71.2%	73.1%	66.1%	67.7%
MAPE	76.2%	78.1%	72.4%	74.1%	67.1%	69.0%

Figure 3 - Figure 5 in appendix demonstrate scatter plots of the RMSE, MAE and MAPE statistics for the Monte Carlo experiments for each data generating process. On the Y-axis is the corresponding statistic obtained using the Bayesian approach. On the X-axis is the statistic obtained using one of the competing methods. All graphs demonstrate that most of the points lie below the solid line at 45 degrees. The Monte Carlo experiments show that Bayesian method with additional information about variation of IO coefficients could potentially improve estimate of unknown IO matrix.

An example: Updating 2003 IO table for Russia

In this part of the work we illustrate an application of the proposed method to real data. There are available official Russian publications of IO accounts for the period from 1995 to 2003. These accounts used the All-Union Classifier of Economy Branches (OKONH). At different years Rosstat published IO accounts for different number of industries and the longest period of the consistent symmetric IO tables with 22 industries is the period from 1998 to 2003.

In our empirical implication we use symmetric IO tables at basic prices from 1998 to 2002 for the estimation of IO matrix coefficients of 2003. We assume that we know only vectors of total outputs and intermediate demand. Value added we also treat as unknown and estimate corresponding coefficients of IO matrix. We apply the same procedure for estimation as in the Monte Carlo experiments: we assume independent truncated normal distributions for each IO coefficient and use coefficients of 2003 IO table as prior mean. To specify standard deviations in prior distribution we estimate standard deviation for each coefficient on the period from 1998 to 2002. To compute posterior distribution of coefficients we apply Markov chain Monte Carlo (MCMC) method with one chain and sampled length of 300 000 simulation.

Figure 6 in appendix shows scatter plot of the posterior mean of the estimated coefficients in comparison to the true values. All points are concentrated around the solid line at 45 degrees and estimates are close enough to the true values. We also compare the performance of the Bayesian method to the RAS and the cross-entropy methods as in the Monte Carlo experiments. Table 2 shows closeness statistics of the three methods.

Table 2. Results of updating 2003 IO table.

	RMSE	MAE	MAPE	RMSPE
Bayes	0.0074	0.0029	0.1844	**0.4502**
RAS	0.0067	**0.0026**	**0.1728**	0.4604
Entropy	**0.0065**	**0.0026**	0.1797	0.4552

where RMSPE is root mean square percentage error:

$$RMSPE = \sqrt{1/(m*n)\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{a_{ij}-\hat{a}_{ij}}{a_{ij}}\right)^2} \qquad (16)$$

The main idea of computing the additional closeness statistic is that estimated standard deviations are approximately proportional to the coefficients values. And all other things being equal, Bayesian method should outperform the other methods according to this statistic because of the corresponding specification of the prior distribution. Nevertheless Bayesian estimate demonstrates the poorest results according the other measures of fit. But this result is not surprising because only the 1995 IO accounts were constructed on the basis of the detailed survey method. The other IO accounts based on nonsurvey methods.

Bayesian estimation of IOT with limited information

In this section we consider a standard IOT exercise:

$$Y = A * X,$$

where Y is an intermediate demand, X is a gross output, and A is a IO-matrix.

We assume that A is unknown, and are trying to get any inference regarding A based on known Y and X. It should be noted, we are not trying to estimate A, but we are interested in tolerance range for every cell of A, if we don't know any other information except Y and X.

Based on the methodology discussed above, we use MCMC method with uniform priors (0,1) to sample A matrix for Use table in 2006, OKVED (NASE) classification, 15 products by 15 activities. The resulting distribution for every cell of A is presented on the Figure 7 in Appendix. As it follows from the figure, the most distributions are asymmetric and skewed to the zero. The asymmetry is easy to explain since the cells of the table are linearly dependent. Large numbers (closer to the unit) are not possible in all cells of a column at one time, whereas low values (closer to zero) are very likely in most cells of a column in one time.

The Figure 1 below shows distribution of pairwise correlation coefficients between cells (225 combinations). Figure 2 represents scatter plots for 9 selected correlation coefficients.

As it follows from the Figure 1, most of the correlation coefficients values are close to zero (not surprising for a sparse matrix). Some of the coefficients have relatively large absolute values, therefore any additional information imposed on such coefficients as a prior information, will impose constrains on other cells of the matrix. Figure 8 and Figure 9 in the Appendix show changes in the estimated distributions in a case when we provide "narrow" prior information for one of the coefficients of the table. The information about the coefficients (A(D,D) in our case) significantly affects distributions of all other coefficients.

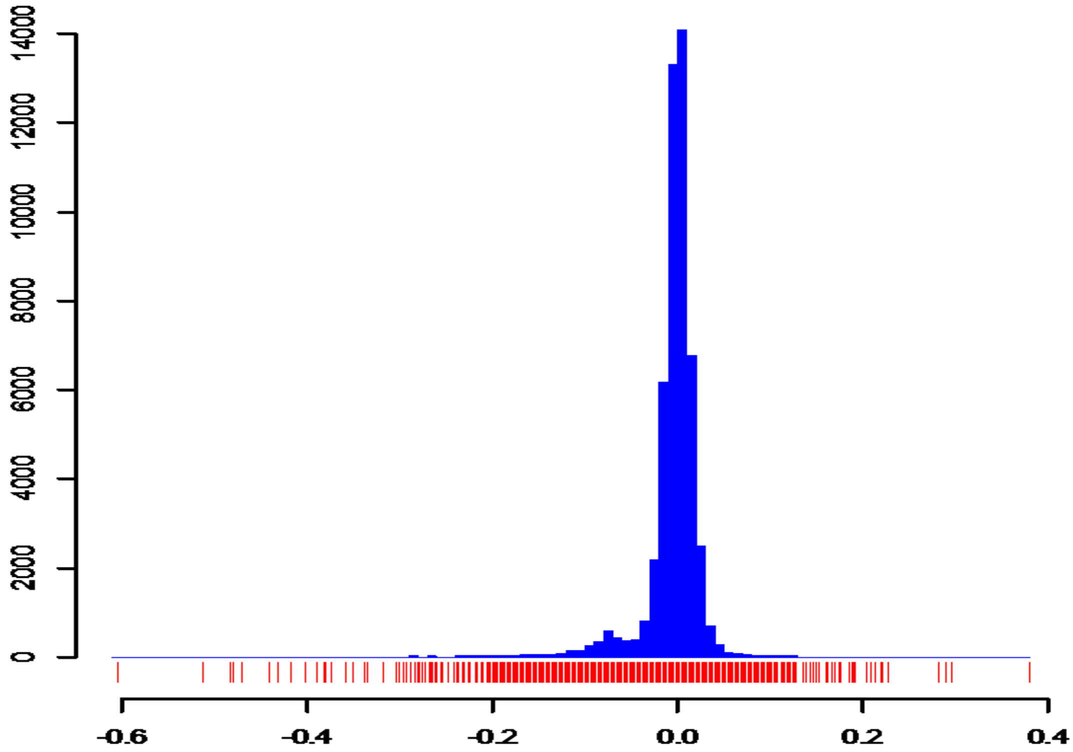

Figure 1. Distribution of pairwise correlation coefficients between estimated cells of USE-2006 matrix.

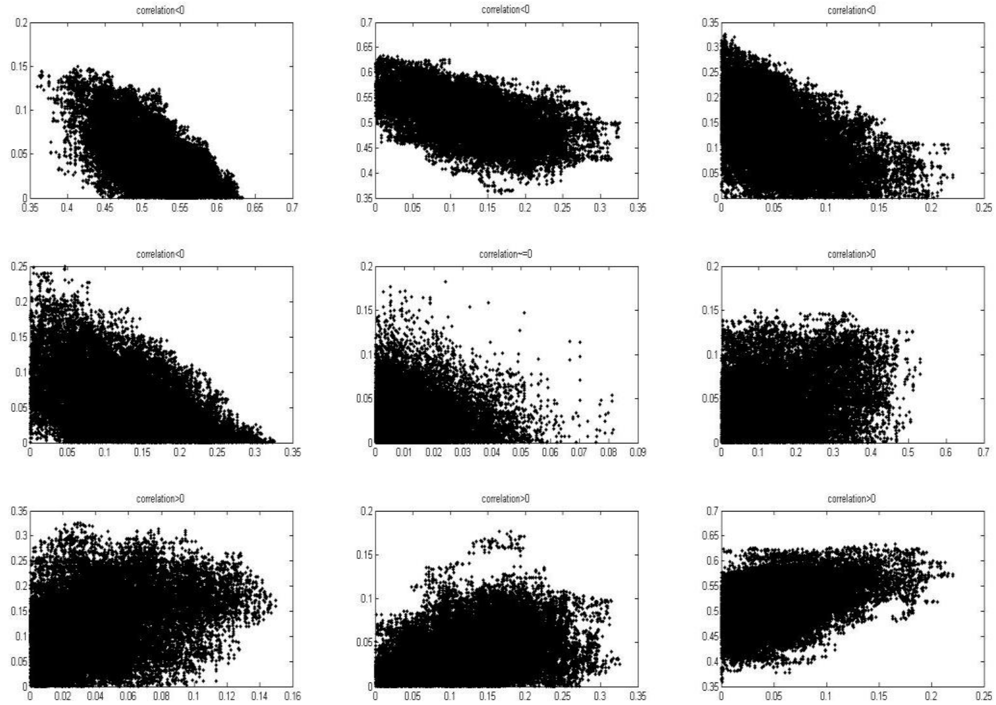

Figure 2. Correlation scatterplots for selected cells of the estimated Use table.

Preliminary findings

- The discussed Bayesian methodology is a feasible alternative to RAS, Maximum entropy, and other methods applied for updating, balancing and updating input-output tables.
- The Bayesian techniques have several advantages over other methods:
    - Flexibility in an experiment design: MCMC method allows working with very wide range of functional forms, distributions, and constrains.
    - Accommodation of various types of information into estimation process: through priors.
    - Lower sensitive to data availability and data quality: uncertainty in data can be naturally introduced into estimation process.
    - Provide full density profile on estimated parameters with covariates: in contrary to standard point estimate methods, MCMC output is a sample of values satisfying the considered model, data and priors; the

sample can be used for estimation of shape of the parameters distribution, confident intervals, correlations and covariates.

- As demonstrated by Monte-Carlo experiment, the proposed Bayesian methodology outperforms RAS and Maximum entropy methods in average in 70% cases (based on 10 thousands experiments).
- Some inference can be drawn regarding IOT coefficients based on information on columns and rows sums of the matrix. Any additional information about IOT coefficient(s), former matrices, variance of the coefficients, might significantly improve the estimate.

From our point of view, Bayesian techniques have a great potential as application to IOT.

References

Eurostat (2008). European Manual of Supply, Use and Input-Output Tables. Methodologies and Working Papers. Luxembourg: Office for Official Publications of the European Communities.

Golan, A., G. Judge and S. Robinson (1994). 'Recovering information from incomplete or partial multisectoral economic data'. The Review of Economics and Statistics. 76(3), 541-549.

Golan, A., G. Judge and D. Miller (1996). Maximum Entropy Econometrics. Chichester UK: Wiley.

Heckelei T., Mittelhammer R., Jansson T. (2008). 'A Bayesian alternative to generalized cross entropy solutions for undetermined econometric models'. Institute for Food and Resource Economics Discussion Paper 2008: 2

Hoff, P. (2009). A First Course in Bayesian Statistical Methods. Springer

Kalantari, B., I. Lari, F. Ricca, and B. Simeona (2008). 'On the complexity of general matrix scaling and entropy minimization via the RAS algorithm'. Mathematical Programming, Series A 112, 371–401.

Leon Y., L. Peeters, M. Quinqu and Y. Surry (1999). 'The use of maximum entropy to estimate input-output coefficients from regional farm accounting data'. Journal of Agricultural Economics 50, 425-439.

Miller, R. E. and P. D. Blair (2009). Input-Output Analysis: Foundations and Extensions. Cambridge: Cambridge University Press, 2nd edition.

Ntzoufras, I. (2008). Bayesian Modeling Using WinBUGS. Wiley

Robinson, S., A. Cattanbo And M. el-Said (2000). 'Updating and estimating a social accounting matrix using cross entropy methods'. Economic Systems Research 13, 47-67.

Stone, Richard. (1961). Input-Output and National Accounts, OECD, Paris.

Stone, R., Bates, J. and Bacharach M. (1963). A program for growth 3, input-output relation ships 1954-1966. Cambridge.

Temurshoev, U., N. Yamano, and C. Webb (2010). 'Projection of supply and use tables: methods and their empirical assessment'. WIOD Working Paper Nr. 2.

Temurshoev, U. and Timmer, M. P. (2010). 'Joint estimation of supply and use tables'. WIOD Working Paper Nr. 3.

Appendix

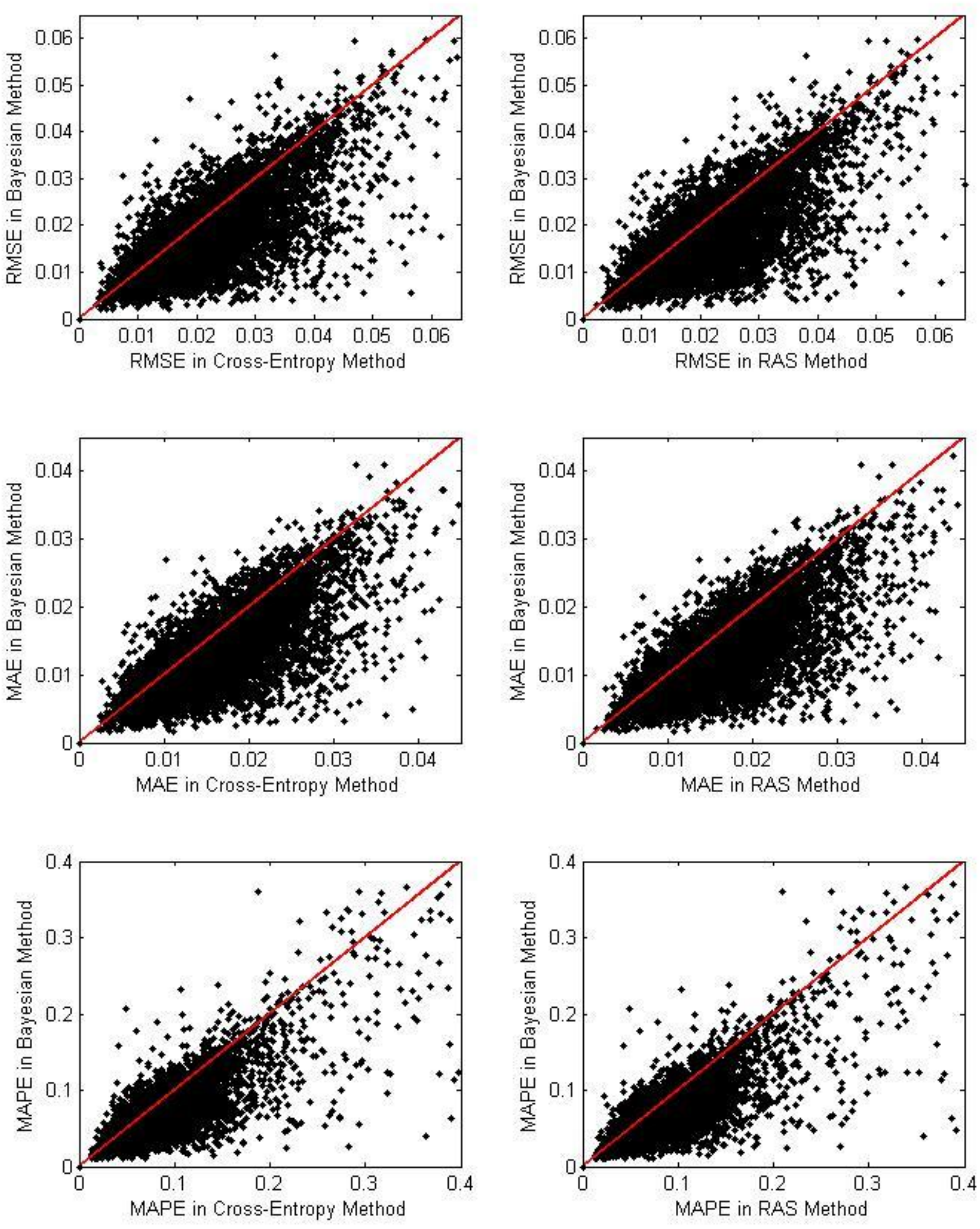

Figure 3. Independent IO coefficients, plots of performance statistics

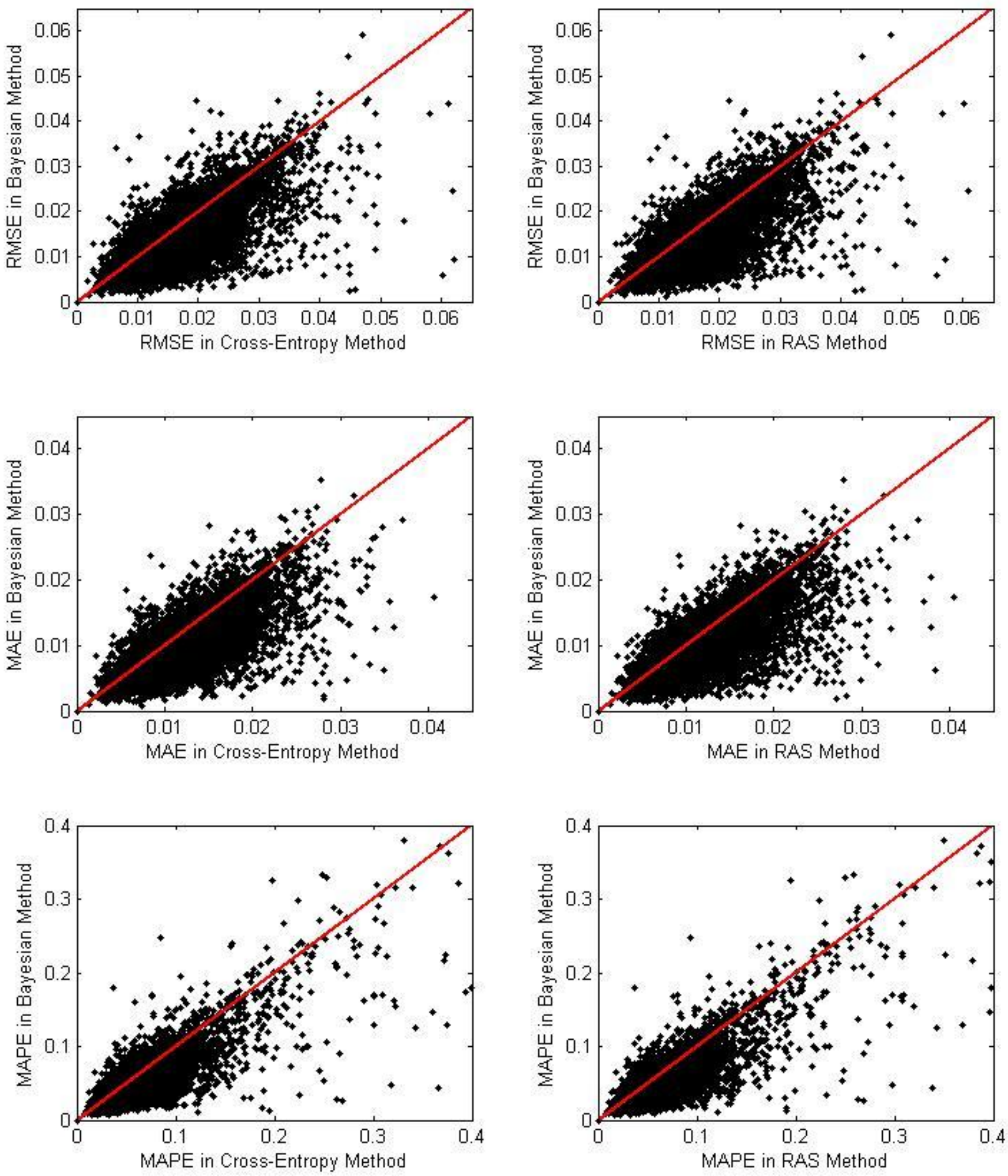

Figure 4. AR(1) process for IO coefficients, plots of performance statistics

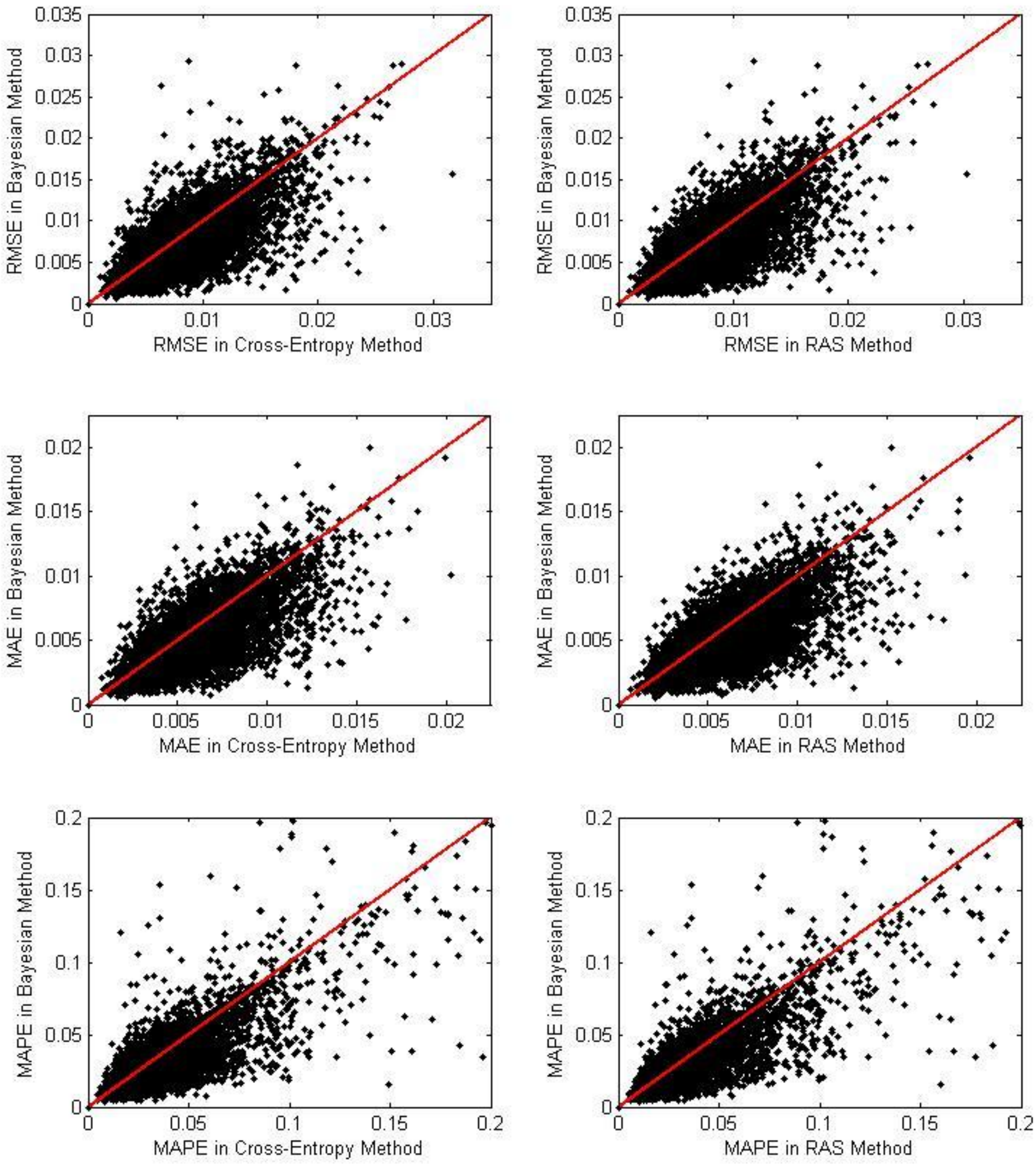

Figure 5. Random Walk process for IO coefficients, plots of performance statistics.

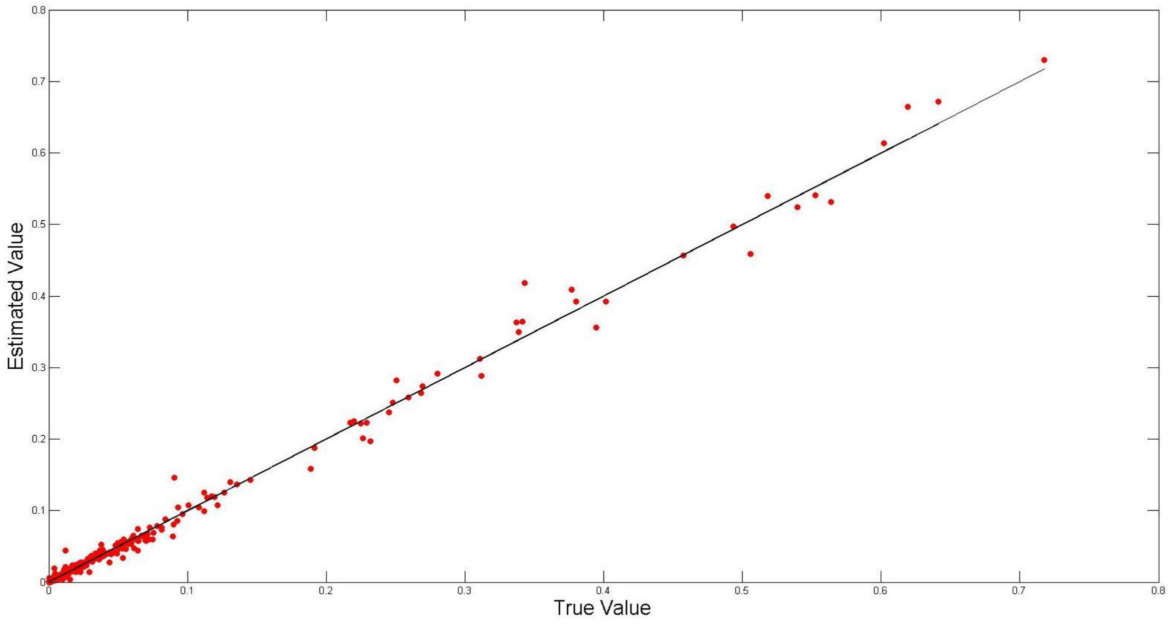

Figure 6. Comparison of the Bayesian estimate of the IO matrix coefficients for 2003 to the true values.

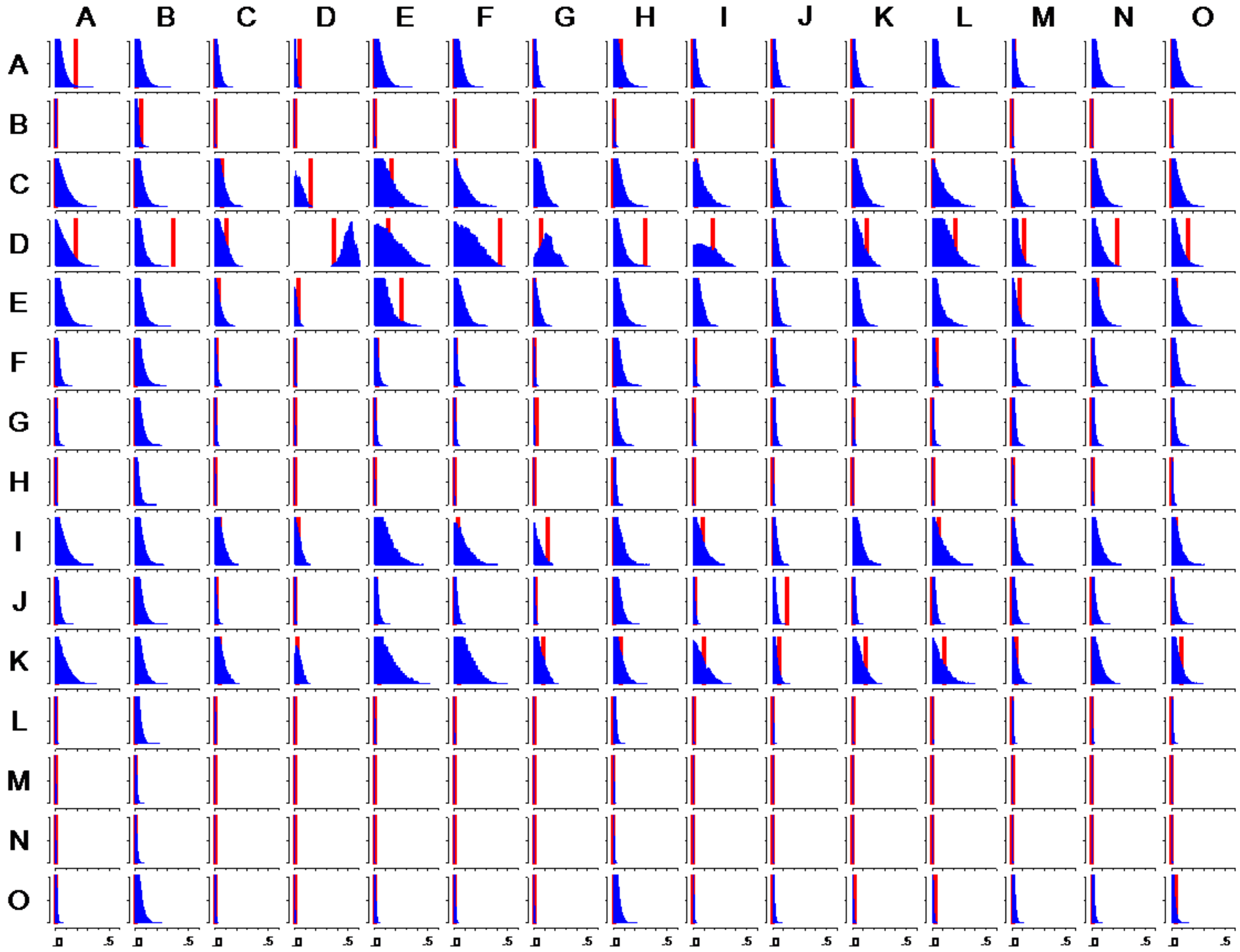

Figure 7. Estimated distributions (blue) of USE table coefficients in a case when intermediate consumption and intermediate demand by industries are known, and actual values (red).

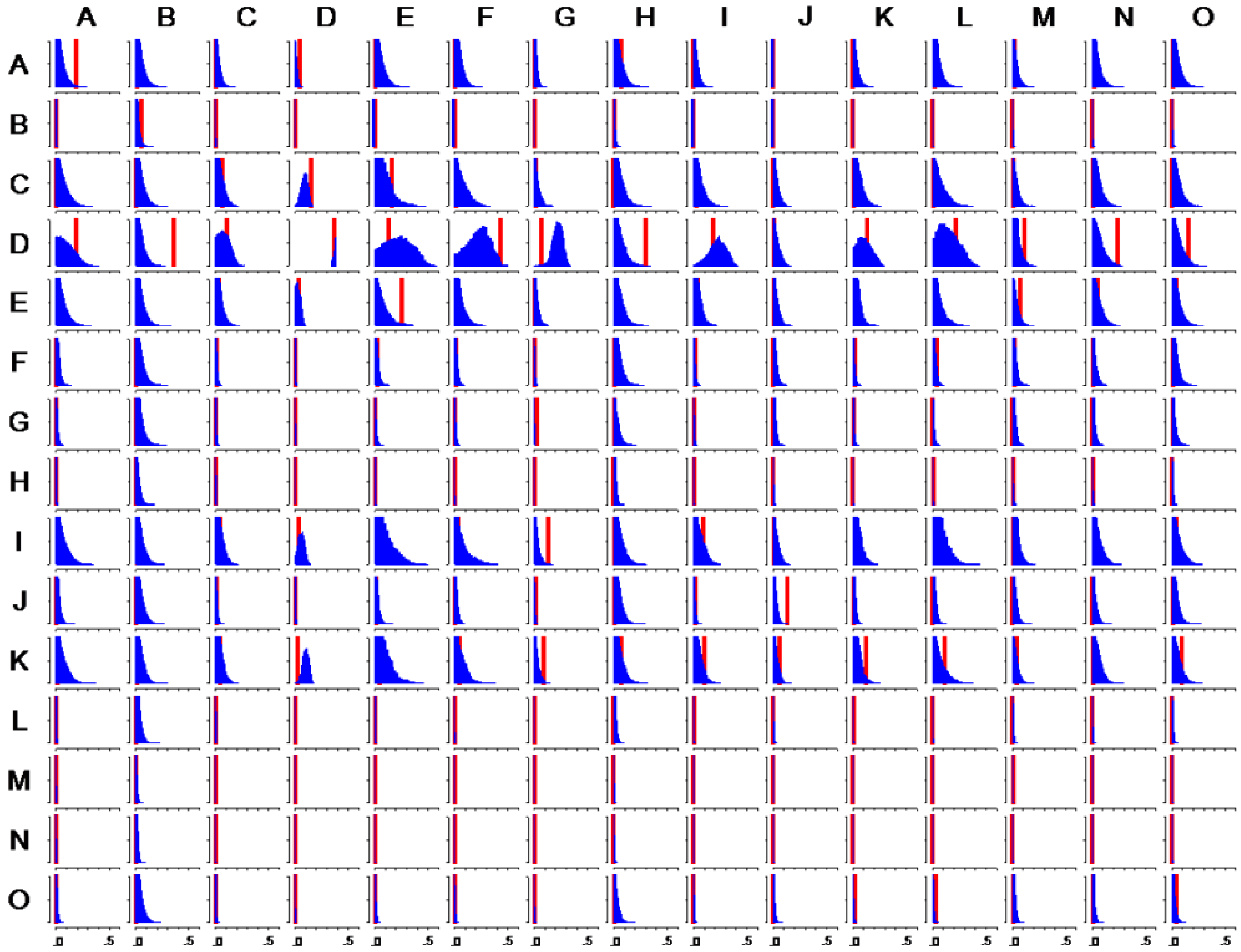

Figure 8. Estimated distributions (blue) of USE table coefficients with added information for Use(D,D) coefficient.

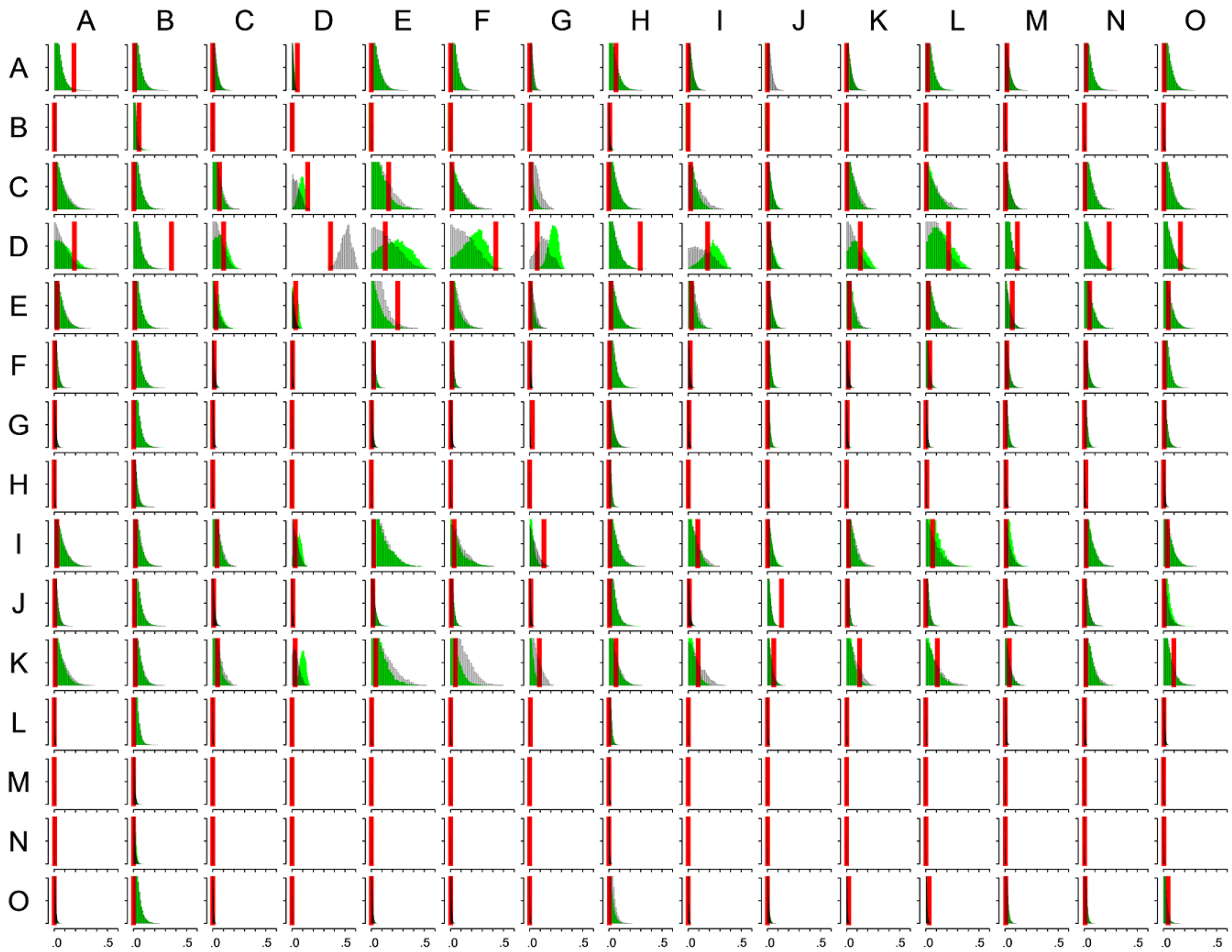

Figure 9. Comparison of estimated coefficients with (green) and without (grey) added information for Use(D,D) coefficient, and actual values (red).