Central European Journal of Economic Modelling and Econometrics

A Bayesian Approach to Matrix Balancing: Transformation of Industry-Level Data under NACE Revision

Jakub Boratyński*

Submitted: 6.08.2016, Accepted: 12.11.2016

Abstract

We apply Bayesian inference to estimate transformation matrix that converts vector of industry outputs from NACE Rev. 1.1 to NACE Rev. 2 classification. In formal terms, the studied issue is a representative of the class of matrix balancing (updating, disaggregation) problems, often arising in the field of multi-sector economic modelling. These problems are characterised by availability of only partial, limited data and a strong role for prior assumptions, and are typically solved using bi-proportional balancing or cross-entropy minimisation methods. Building on Bayesian highest posterior density formulation for a similarly structured case, we extend the model with specification of prior information based on Dirichlet distribution, as well as employ MCMC sampling. The model features a specific likelihood, representing accounting restrictions in the form of an underdetermined system of equations. The primary contribution, compared to the alternative, widespread approaches, is in providing a clear account of uncertainty.

Keywords: matrix balancing, Bayesian inference, NACE revision, transformation matrix, multi-sector modelling

JEL Classification: C11, C81, D57, C67, C68

*University of Łódź; e-mail: jakub.boratynski@uni.lodz.pl

1 Introduction

Multi-sector economic models, such as computable general equilibrium (CGE), input-output or input-output-econometric models, rely on data characterised by relatively detailed disaggregation. These models typically include several tens of industries and commodities, and sometimes also further distinguish individual regions, socio-economic groups of households, occupational groups of labour etc. The published data rarely meet the needs of multi-sector modelling immediately. Major difficulties include insufficient disaggregation, inconsistencies of different data sources, significantly delayed publication (by even a few years), limited time scope, and changes in classifications. Consequently, in the field of multi-sector economic modelling, there is substantial literature devoted to the problems of data processing, such as updating, balancing or estimation, based on partial information, with an important role of prior assumptions.

The data issue addressed in this paper arose as a part of a modelling task which required the use of comparable input-output tables for different years, namely 2000, 2005 and 2010. However, the latest available, 2010 table is based on different industry classification (NACE Rev. 2) than the previous ones (NACE Rev. 1.1). In order to get a consistent set of input-output tables, transformation procedure was necessary (applied to the older tables), with elements of both judgement and estimation. As an important step, that procedure involved identifying a matrix of coefficients allowing to transform a vector of industry outputs from NACE Rev. 1.1 to NACE Rev. 2 classification. In this article we focus on that step only, in order to illustrate estimation approach with a view to its possible extensions to other similar data problems.

As we will show in the next section, the problem of estimation of transformation coefficients is formalised as the so called matrix balancing problem. In the area of multi-sector modelling that problem can be summarised as follows: find matrix Z, knowing its row and column totals, along with an initial guess Z_0 (most often formulated using past data). The unknown matrix is typically an input-output table (or its part) or a social accounting matrix (SAM) – the primary data sources for multi-sector models. Balancing may be performed in (a rather standard) situation, where e.g. the input-output table relates to year $t-5$, while some partial (aggregate) data are already available for year t. In the literature one can find numerous studies on the matrix balancing problem. Two broad approaches are dominant, namely bi-proportional balancing (RAS; the name 'RAS' refers to symbols R, A, and S used in the original mathematical notation) methods and entropy-based estimation methods (including Minimum Cross Entropy and Generalised Cross Entropy).

RAS method involves iterative scaling of rows and columns of the initial estimate, Z_0, of the unknown matrix until the match with target row and column totals is reached (Miller and Blair, 2009, p. 313–336, Lahr and De Mesnard, 2004). Different mutations of RAS have been developed, to deal with with negative entries (Junius and Oosterhaven, 2003), sign changes (Lenzen *et al.*, 2014), and information beyond

just row and column totals (Lenzen, Gallego and Wood, 2009, Gilchrist and St Louis, 1999).

The competing technique of matrix balancing uses constrained (numerical) optimisation. It consists in minimizing divergence of Z from Z_0, under constraints of row and column totals being equal to their target values (in practice, the problem is often reformulated such that proportions – e.g. column shares of Z – are estimated rather than Z directly). Although a variety of divergence metrics can be used (e.g. sum of absolute or squared differences, either weighted or unweighted – see Jackson and Murray, 2004), most applications refer to information-theoretic concept of minimum cross entropy (Golan, Judge and Miller, 1996, p. 11–14, 29-31; see also Golan, Judge and Robinson, 1994, Robinson, Cattaneo and El-Said, 2001).

It has been shown that in fact minimum cross entropy and RAS methods lead to equivalent results under certain assumptions (McDougall, 1999). At the same time, generalisation of entropy-based techniques, proposed by Golan, Judge and Miller (1996), extends the scope of their applications. In the context of matrix balancing, it allows for example to solve problems with noise, i.e. with constraints not binding strictly (Golan and Vogel, 2000, who also utilise data for multiple periods), or perform estimation with unknown column totals, using alternative information input (Peters and Hertel, 2016). In general, entropy methods make a convenient framework for the estimation of systems of equations when data are weakly informative, and supplementary prior information is available.

With acknowledgement of advantages of generalised entropy methods, they have been criticised by Heckelei, Mittelhammer and Jansson (2008) for awkward specification and interpretation of prior information. The same authors propose Bayesian highest posterior density (HPD) estimation as an alternative to entropy-based methods. HPD formulation developed by Heckelei, Mittelhammer and Jansson (2008) is the starting point of our analysis. Recognizing the fact that all of the discussed methods lack a comprehensive account of uncertainty, we employ inference based on the full posterior distribution. Our additions to the approach proposed by Heckelei, Mittelhammer and Jansson (2008) include the use of Dirichlet distribution as means of specifying prior knowledge (for both the uninformative and informative cases), and application of MCMC simulation to analyse the posterior distribution. Our goal is to demonstrate how processing of data for multi-sector modelling could possible benefit from the Bayesian approach. The considered problem of NACE Rev. 1.1 to NACE Rev. 2 transformation, although practical, should be treated as illustrative.

2 Problem formulation and numerical example

For ease of exposition, we will first consider a slightly simplified case, along with a 2x2 numerical example. Next we extend the formulation slightly to cover the actual, practical estimation problem.

The transformation of industry output data from NACE Rev. 1.1 to NACE Rev. 2

can be written as follows:

$$\begin{aligned} y_i &= \sum_k \lambda_{ki} \cdot x_k \\ \sum_i \lambda_{ki} &= 1 \\ \lambda_{ki} &\geq 0 \end{aligned} \tag{1}$$

where x_k (for $k = 1, \ldots, K$) is output of NACE Rev. 1.1 industry k, y_i (for $i = 1, \ldots, I$) is output of NACE Rev. 2 industry i. Transformation coefficients λ_{ki} are interpreted as share of NACE Rev. 1.1 industry k output classified to industry i under NACE Rev. 2. The coefficients form a $K \times I$ transformation matrix $\mathbf{\Lambda}$. For the problem to be consistent, it is required that data satisfy $\sum_k x_k = \sum_i y_i$. We shall refer to conditions (1) as 'accounting constraints'.

We assume that output vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, are known, while $\mathbf{\Lambda}$ is estimated. Therefore, the purpose is not transformation of output itself, as data in *both* NACE Rev. 1.1 and NACE Rev. 2 breakdowns are available for certain time window (6-year, in the case of Poland). Rather the recovered $\mathbf{\Lambda}$ matrix may be used to transform data on other categories 'linked' to output (e.g. intermediate inputs) which are not backward-revised ('backcasted') by statistical agencies after classification change. Worth noting, λ_{ki} coefficients could in principle be compiled from source micro data. In fact, behind backward revisions of industry data made by statistical agencies, there must exist – at least implicitly – some transformation matrix. Such information, however, is not published, and thus the need for estimation. Finally, even though access to source data and assumptions underlying data backcasting could ultimately nullify our estimation problems, our considerations equally apply to similarly structured problems where the estimand refers to non-existent data – at least at the time of performing estimation. Consider a simple numerical illustration of problem (1):

			NACE Rev. 1.1 ($\mathbf{x}$)
	$\lambda_{11} \cdot 200$	$\lambda_{12} \cdot 200$	**200**
	$\lambda_{21} \cdot 50$	$\lambda_{22} \cdot 50$	**50**
NACE Rev. 2 ($\mathbf{y}'$)	**150**	**100**	

In the above example there are 3 independent linear equations (accounting constraints) and 4 variables, which implies that the system is underdetermined (indefinite). However, here one can easily identify the lower and upper bounds of coefficients, as shown below:

$$\lambda_{11} \in \langle 0.5, 0.75\rangle \quad \lambda_{12} \in \langle 0.25, 0.5\rangle$$

$$\lambda_{21} \in \langle 0, 1\rangle \quad \lambda_{22} \in \langle 0, 1\rangle$$

It should also be noted that in the above case assigning a value to just one coefficient determines, through the accounting constraints (1), the values of all remaining coefficients.
In the context of the above example, the aim of applying Bayesian procedure can be twofold:

1. in the case of uninformative priors for coefficients, determine their lower and upper bounds, based on data and accounting constraints;

2. in the case of informative priors for coefficients, update them using information from data and accounting constraints.

We will consider both cases in turn. For the stylized example aim 1 has actually been achieved already, but for larger problems, as well as for more general, and perhaps non-linear models, the task will be non-trivial. We should stress that constraints form a system of simultaneous equations (plus the non-negativity constraints), and thus estimation cannot be split into individual equations.

3 Bayesian inference for the stylized example

Let $\boldsymbol{\lambda}_k$ denote row vectors of transformation matrix $\boldsymbol{\Lambda}$ (each vector being of length I). Since elements $\boldsymbol{\lambda}_k$, for each k, are interpreted as shares (proportions) – they are non-negative and sum to one – they can be thought of as realisations from Dirichlet distribution (see Bolshev article from *Encyclopedia of Mathematics*; see also Ferguson, 1973, Darroch and Ratcliff, 1971):

$$\boldsymbol{\lambda}_k \sim \text{Dirichlet}\left(c_k \cdot \boldsymbol{\lambda}_k^{(0)}\right) \tag{2}$$

where $\boldsymbol{\lambda}_k^{(0)}$ are interpreted as mean prior coefficient values (satisfying the condition $\sum_i \lambda_{ki}^{(0)} = 1$), and c_k is concentration parameter (the higher c_k, the more concentrated are random draws from Dirichlet distribution around prior means). When $\lambda_{k1}^{(0)} = \ldots = \lambda_{kI}^{(0)} = 1/I$ and $c_k = I$, the distribution of $\boldsymbol{\lambda}_k$ is (multivariate) uniform across the simplex defined by conditions $\lambda_{ki} \geq 0$ and $\sum_i^I \lambda_{ki}$.
In formulating the likelihood function we follow the concept of Heckelei, Mittelhammer and Jansson (2008, p. 9–10), who attribute the value 1 to all solutions $\boldsymbol{\Lambda}$ that satisfy accounting constraints, and 0 otherwise. Under such an approach, each specific solution of (1) is viewed equally plausible. For our problem the likelihood function

rewrites as:

$$\begin{aligned} L\left(\boldsymbol{\Lambda};\boldsymbol{y},\boldsymbol{x}\right) &\equiv 1 \qquad \text{when } \boldsymbol{y} = \boldsymbol{\Lambda}'\boldsymbol{x} \\ L\left(\boldsymbol{\Lambda};\boldsymbol{y},\boldsymbol{x}\right) &\equiv 0 \qquad \text{when } \boldsymbol{y} \neq \boldsymbol{\Lambda}'\boldsymbol{x} \end{aligned} \tag{3}$$

Note that for specific, fixed $\boldsymbol{\Lambda}$, given $\boldsymbol{x}$, there is only one valid $\boldsymbol{y}$ vector, characterised by probability 1; that is the sampling distribution of $\boldsymbol{y}$ (given $\boldsymbol{x}$) is degenerate, concentrated at one point.

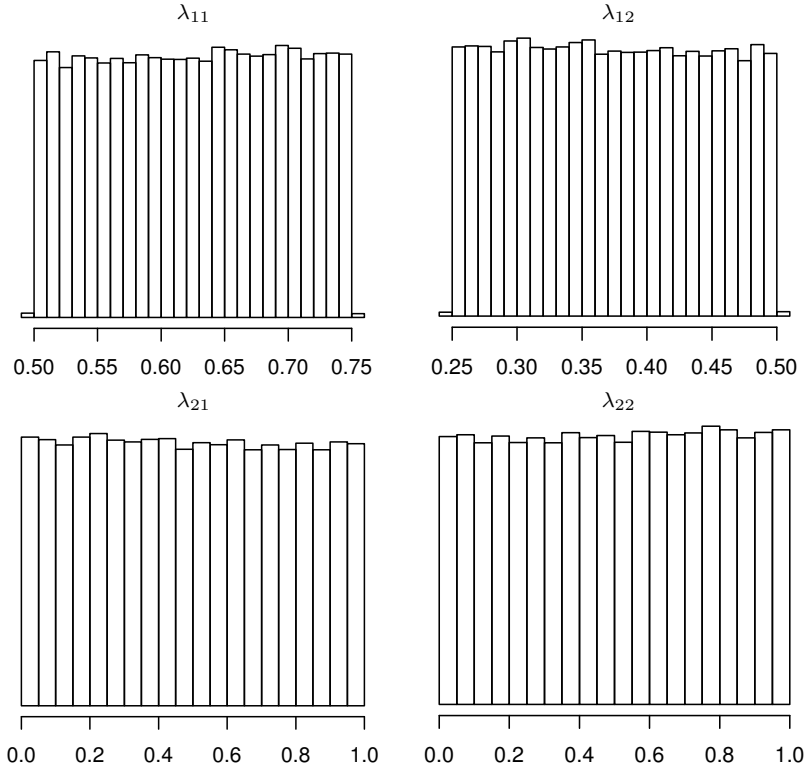
In order to be able to implement the model defined by (1)-(3) in Stan modelling language (Carpenter *et al.*, 2016), a 'noise' component was added, leading to the following model ultimately applied in simulation:

$$\begin{aligned} &\boldsymbol{\lambda}_k \sim \text{Dirichlet}\left(c_k \cdot \boldsymbol{\lambda}_k^{(0)}\right) \\ &\boldsymbol{y} = \boldsymbol{\Lambda}'\boldsymbol{x} + \boldsymbol{e} \\ &\boldsymbol{e} \sim \text{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right) \\ &\boldsymbol{\Sigma} = \text{Diag}\left(\sigma_1^2, \ldots, \sigma_I^2\right) \\ &\sum_i \lambda_{ki} = 1 \\ &\lambda_{ki} \geq 0 \end{aligned} \tag{4}$$

where the standard deviation σ_i is set arbitrarily to a small value (we chose $\sigma_i = 0.1$ for all i). Adding small normal errors e_i should be treated as a workaround to overcome the lack of sampling statements adequate for the peculiar likelihood formulation (3). Figure 1 shows marginal posterior distributions of individual λ_{ki} parameters for the stylized 2×2 example. As a consequence of using uninformative priors, combined with 'flat-surface' likelihood function, the posteriors are approximately uniform (the fact that minor non-zero density appears just outside the bounds is the result of small normal error terms, and does not change the general picture). The simulation correctly identifies parameter bounds – they are equal to the bounds derived analytically. Nevertheless, as mentioned above, Bayesian MCMC sampling may prove useful in cases where analytical solutions are unavailable or difficult.

Consider now the case of informative priors. It might be (and in fact often is, as we shall see in the next section) the case that one can identify 'natural' counterparts of certain NACE Rev. 1.1 industries in the NACE Rev. 2 classification. Then it is reasonable to presume a priori that the majority of output of a NACE Rev. 1.1 industry will be classified to the corresponding NACE Rev. 2 industry. A convenient way of operationalizing such assumption is to formulate the mean of prior share (transformation coefficient), along with its standard deviation. Note that expressing prior information in terms of *both* mean and standard deviation is only available for *one* element of each vector $\boldsymbol{\lambda}_k$, while for the remaining elements only the means can be specified. This feature can be considered a limitation of the Dirichlet distribution, and points towards possibility of using more elastic distributions, appropriate for

Figure 1: Posterior distribution of λ_{ki} for 2×2 example with uninformative priors



compositional data. A major alternative is the additive logistic normal (ALN) class (see Aitchison and Shen, 1980; Aitchison, 1982). For example, the ALN distrubution was used by Osiewalski (2001, p. 146–165) in the Bayesian estimation of complete demand system, being a model that explains the *structure* of household consumption expenditures.

In order to translate prior assumptions about the mean and standard deviation of a distinguished element of a $\boldsymbol{\lambda}_k$ vector, λ_{kd}, consider the fact that the marginal of Dirichlet distribution (2) is the beta distribution (Ferguson, 1973, p. 211):

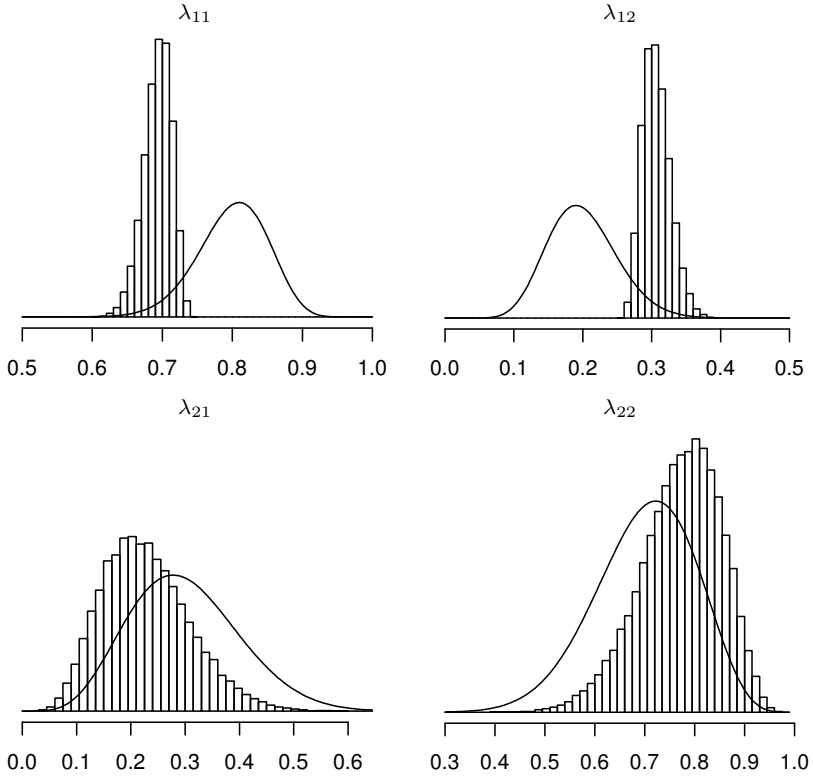
$$\lambda_{kd} \sim \text{Beta}\left(c_k \cdot \lambda_{kd}^{(0)},\, c_k - c_k \cdot \lambda_{kd}^{(0)}\right) \tag{5}$$

Denoting by s_{kd} the standard deviation of λ_{kd}, the concentration parameter c_k can be determined as follows:

$$c_k = \left(\frac{\lambda_{kd}^{(0)} \cdot \left(1 - \lambda_{kd}^{(0)}\right)}{s_{kd}^2} - 1 \right) \tag{6}$$

For the stylized 2×2 example we assumed prior $\lambda_{11}^{(0)} = 0.8$, and the related standard deviation $s_{11} = 0.05$; we also assumed $\lambda_{22}^{(0)} = 0.7$ and $s_{22} = 0.1$; prior assumptions concerning $\lambda_{12}^{(0)}$ and $\lambda_{21}^{(0)}$ are then implicit. Results – in the form of posterior and prior marginal distributions (illustrated by histograms and densities, respectively) – are shown in figure 2.

Figure 2: Posterior distribution of λ_{ki} for 2×2 example with informative priors



4 Estimating 44×63 transformation matrix based on actual data

In the actual, practical estimation problem we have industry-level nominal output data for Poland for the years 2000-2005 (available from Eurostat) under both NACE Rev. 1.1 and NACE Rev. 2. Here we present estimation based on single year (2000)

data only. NACE Rev. 1.1 data cover 44 industries, while NACE Rev. 2 – 63 industries, both mostly at the two-digit classification level, with some industries being aggregated into larger groups. In such a case, the coefficient estimated matrix $\mathbf{\Lambda}$ is of 44×63 size. There are two notable differences compared to the stylized case considered above.

1. *Matrix* $\mathbf{\Lambda}$ *is sparse.* At the discussed disagrregation levels, output of a particular NACE Rev. 1.1 industry can usually be reclassified into one of only few NACE Rev. 2 industries, according to the so called correspondence tables between the two classifications. In the considered case, $\mathbf{\Lambda}$ contains 198 non-zero elements (i.e. elements subject to estimation), which implies that it is approximately 7% dense. Correspondence tables also facilitate specification of prior assumptions, by restricting the number of potential links between NACE Rev. 1.1 and NACE Rev. 2 industries, and also by providing additional sector-specific information on the relationships (see for example correspondence tables available from Eurostat at `http://ec.europa.eu/eurostat/web/nace-rev2/correspondence_tables`). Worth noting, matrix sparsity also differentiates the considered problem from the typical setting of matrix balancing task, referring to dense matrices such as input-output tables.

2. *Data are not fully consistent.* Inconsistencies become apparent when our data are confronted with the correspondence tables. At the analysed disaggregation level, according to correspondence tables, there are four cases of one-to-one mapping between NACE Rev. 1.1 and NACE Rev. 2 industries. In these cases output of the related industries should be identical under both classifications, but actually there are differences (notably for the water collection, treatment and supply sector the difference is as much as 40%). At the same time, data satisfy the condition $\sum_i y_{it} = \sum_k x_{kt}$, implying that errors spread over to other industries (and there may be as well other inconsistencies, not being that evident).

Taking into account the sparsity of $\mathbf{\Lambda}$ requires adjustments to the previous model formulation, (4). This stems from the fact that Dirichlet-distributed random vector cannot contain zero elements. Therefore we define as $\tilde{\boldsymbol{\lambda}}_k$ vectors consisting of just the non-zero elements of rows of $\mathbf{\Lambda}$ matrix, $\boldsymbol{\lambda}_k$. More precisely, $\tilde{\boldsymbol{\lambda}}_k$ consists of those components of $\boldsymbol{\lambda}_k$ that represent *possible* – according to correspondence tables – transitions ('flows') from NACE Rev. 1.1 to NACE Rev. 2 industries. Here the Dirichlet distribution is in fact overly restrictive, as some of the possible connections between classifications may appear void in reality and, consequently, $\tilde{\boldsymbol{\lambda}}_k$ components should be allowed to take zero values too. However, we view this limitation as merely theoretical, since elements of $\tilde{\boldsymbol{\lambda}}_k$ can take values arbitrarily close to zero which is sufficient for practical applications.
The second feature – data inconsistencies – necessarily entails the use of error terms, in this case playing the role beyond that of a 'workaround' needed to formulate sampling

statements in the program implementation. One can add to this a more fundamental reason, i.e. the fact that data supplied by statistical agencies are themselves subject to uncertainties, and, in particular, backward revisions of data due to classification changes involve problems related to inadequacy of sampling schemes used in older business surveys to the new data requirements, thus leading to application of various estimation techniques (van den Brakel, 2010, describes the related issues in the context of the recent NACE revision).
Taking into account the above comments, the ultimate formulation of our model is as follows:

$$\begin{aligned}
&\tilde{\boldsymbol{\lambda}}_k \sim \text{Dirichlet}\left(c_k \cdot \tilde{\boldsymbol{\lambda}}_k^{(0)}\right)\\
&\sum_{j=1}^{J_k} \tilde{\lambda}_{kj} = 1\\
&\tilde{\lambda}_{kj} > 0\\
&\lambda_{ki} = f_k(\tilde{\boldsymbol{\lambda}}_k) \quad \text{for } (k,i) \in \Omega\\
&\lambda_{ki} = 0 \quad \text{for } (k,i) \notin \Omega\\
&\boldsymbol{y} = \boldsymbol{\Lambda}'\boldsymbol{x} + \boldsymbol{e}\\
&\boldsymbol{e} \sim \text{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)\\
&\boldsymbol{\Sigma} = \text{Diag}\left(\sigma_1^2, \ldots, \sigma_I^2\right)
\end{aligned} \tag{7}$$

where $f_k()$ is a *mapping* of $\tilde{\boldsymbol{\lambda}}_k$ vector to non zero elements of the full-length $\boldsymbol{\lambda}_k$ vector. Note that for each k – i.e. for each NACE Rev. 1.1 industry – the vector $\tilde{\boldsymbol{\lambda}}_k$ (and, accordingly, $\tilde{\boldsymbol{\lambda}}_k^{(0)}$) will in general have different length, J_k, equal to the number of corresponding NACE Rev. 2 industries. Information from correspondence tables is represented by the Ω set, comprising of pairs (k,i), indicating the possible paths of output reclassification. The distribution of error terms is specified explicitly as multivariate normal with a diagonal covariance matrix, $\boldsymbol{\Sigma}$.
It should be underlined that, regarding error terms, their standard deviations are assumed fixed, rather than estimated. Intuitively, single-year data are insufficient for joint estimation of transformation coefficients and the scale of resulting errors. Estimation of $\boldsymbol{\Lambda}$ seems meaningful only upon belief that errors are in some way restricted – otherwise the data, together with accounting restrictions, would not be informative enough from the point of view of recovering unknown $\boldsymbol{\Lambda}$. We chose to set standard deviations σ_i to 1% of observed output of individual NACE Rev. 2 industries, y_i. This reflects prior belief that overall the inconsistencies in official data are in fact relatively small. According to an alternative interpretation (had the previous one seemed unjustified), the assumed scale of error terms is just enough to accommodate the apparent inconsistencies (discussed above) alone. Worth to add, in the four cases of one-to-one relationship between NACE Rev. 1.1 and NACE Rev. 2 industries, the corresponding coefficients were exogenously set to one, effectively eliminating output

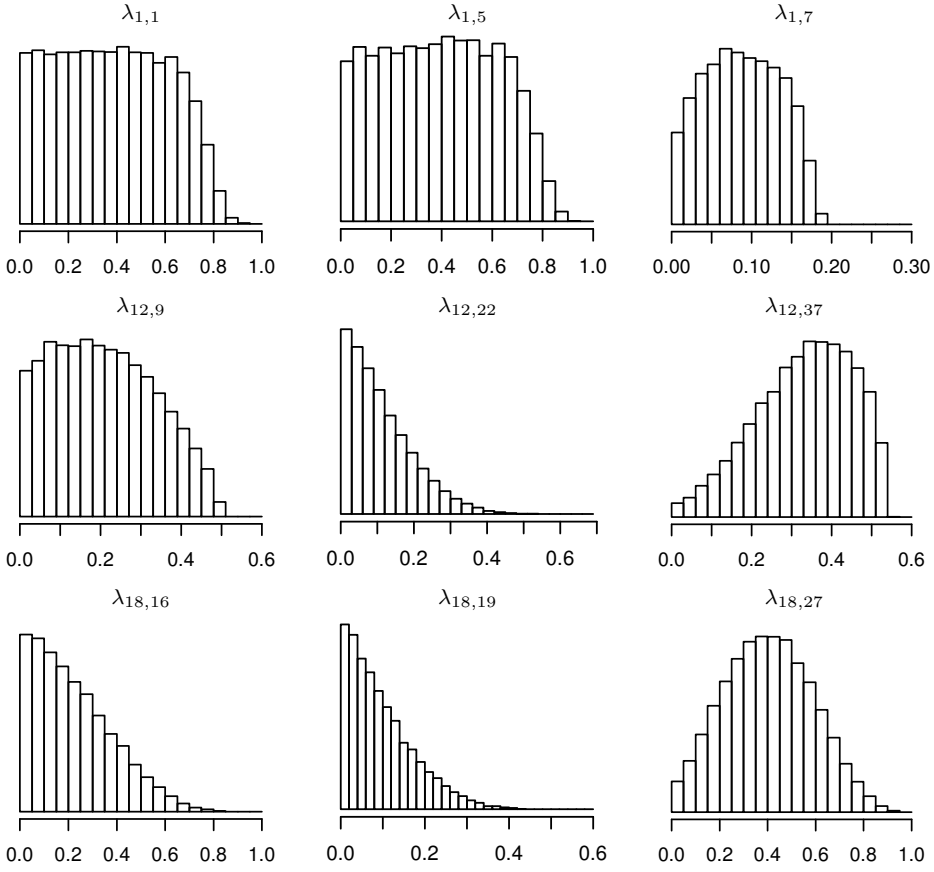
data for those industries from the estimation procedure.

A more standard approach to the modelling of error terms would be to assign a prior distribution to the covariance matrix, $\mathbf{\Sigma}$, rather than treat that matrix as fixed. However, with such an approach we have encountered serious computational difficulties in the posterior sampling for our model – in fact we managed to find satisfactory solutions only for special cases, using a small, stylized case. Therefore we treat the issue of error specification in matrix balancing problems as a topic for further research. The formulation involving fixed error variances has proven practical from the point of view of computational efficiency. In order to illustrate how estimation results are affected by the choice of error scale, the study is supplemented with sensitivity analysis – along with the central case of σ_i's equal to 1% of observed output, we also consider alternative values: 0.2%, 5%, and 25%. The results of sensitivity analysis are discussed in the concluding part of the current section.

As with the stylized 2×2 example, we begin with the uninformative priors case, i.e. we employ the assumption that the prior distribution of each individual $\tilde{\boldsymbol{\lambda}}_k$ is multivariate uniform. The marginal posterior distributions are illustrated in figure 3, using as an example 9 out of 198 non-zero transformation coefficients. Compared to the 2×2 case, the resulting posterior distributions now usually take trapezoid or triangular shapes (not strictly so, partly due to existence of small normal error terms in the model), rather than strictly rectangular (uniform), which is the effect of more complex relationships between different coefficients in the current higher-dimension case. Still, however, with such shapes of histograms, it is justified to report the results in terms of ranges – the maximum minus the minimum MCMC sample value.

Absolute values of those ranges are graphically presented in figure 4. In this figure, as in the remaining similar figures in this paper, table rows and columns represent NACE Rev. 1.1 and NACE Rev. 2 industries, respectively. Both rows and columns are labelled by standard two-digit NACE codes, from respective revisions. The framed cells show correspondence between NACE Rev. 1.1 and NACE Rev. 2 industries, and hence indicate which cells of the $\mathbf{\Lambda}$ coefficient matrix were subject to estimation (the remaining coefficients being zero by definition). In figure 4, the dark gray indicates absolute range equal to 1 (an extreme case for coefficient values ranging from 0 to 1) whereas white indicates range equal to 0 (the no-uncertainty case). Therefore, the darker the hue, the greater the uncertainty about a respective coefficient. As can be seen from figure 4, results are mixed, with areas of small uncertainty – mostly regarding services – and areas of medium to large uncertainty – mostly regarding the manufacturing industries. One can see from the placement of the non-zero cells that lower and upper bounds of at least some coefficients could easily be deduced, rather than simulated. Perhaps this even applies to all unknown coefficients, but it was not our purpose to explore that possibility. Rather, we wanted to show an application of simulation-based approach that also easily extends to more complex formulations. Still it is also clear from the picture that results are overall rather unusable, in terms of direct application to the transformation of input-output tables,

Figure 3: Posterior distribution of selected λ_{ki} coefficients under uninformative priors



unless heavy aggregation is performed. However, they point to areas (industries) on which efforts to acquire additional information should be focused.

As a next step, we introduced informative priors, in line with the approach proposed for the stylized example. For each NACE Rev. 1.1 industry we attempted to identify an explicit match among NACE Rev. 2 industries, in terms of primary activity profile (which usually entails identical or similar industry name). Whenever such a match was found, we assumed that 90% (as a mean value) of output of a given NACE Rev. 1.1 industry is 'transferred' to its NACE Rev. 2 counterpart, with standard deviation being equal to 5 percentage points. In all, 26 such cases were identified (out of the total 44). In additional 7 cases, NACE Rev. 1.1 industry has only one corresponding NACE Rev. 2 industry, leading to coefficient being 1 by identity. In the other 11

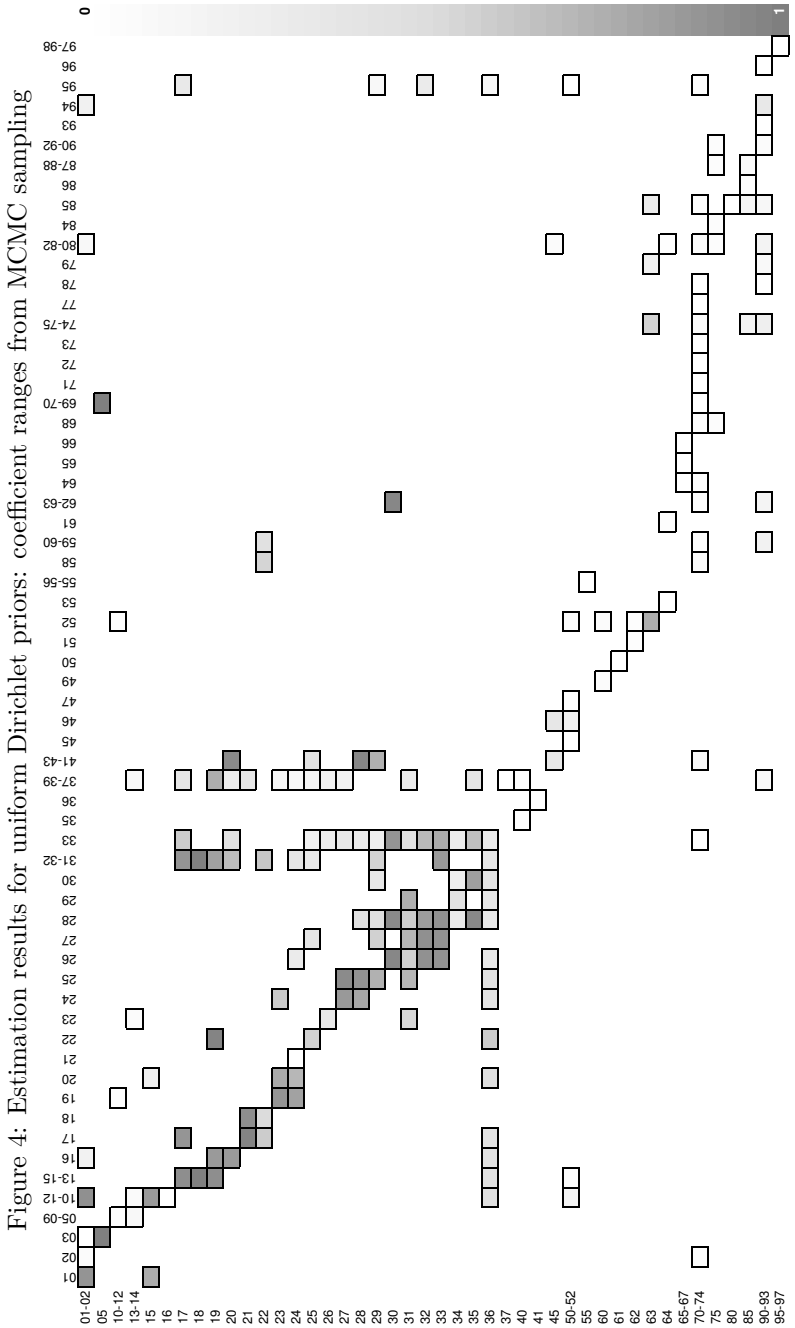


Figure 4: Estimation results for uniform Dirichlet priors: coefficient ranges from MCMC sampling

industries, no single explicit match could be identified, so we remained with uniform Dirichlet distribution for the respective $\tilde{\boldsymbol{\lambda}}_k$ vector. In figures 6-7, the explicit matches, as well as the 7 cases of full certainty, are indicated by bold cell frames.

Figure 5: Posterior distribution of selected λ_{ki} coefficients under uninformative priors

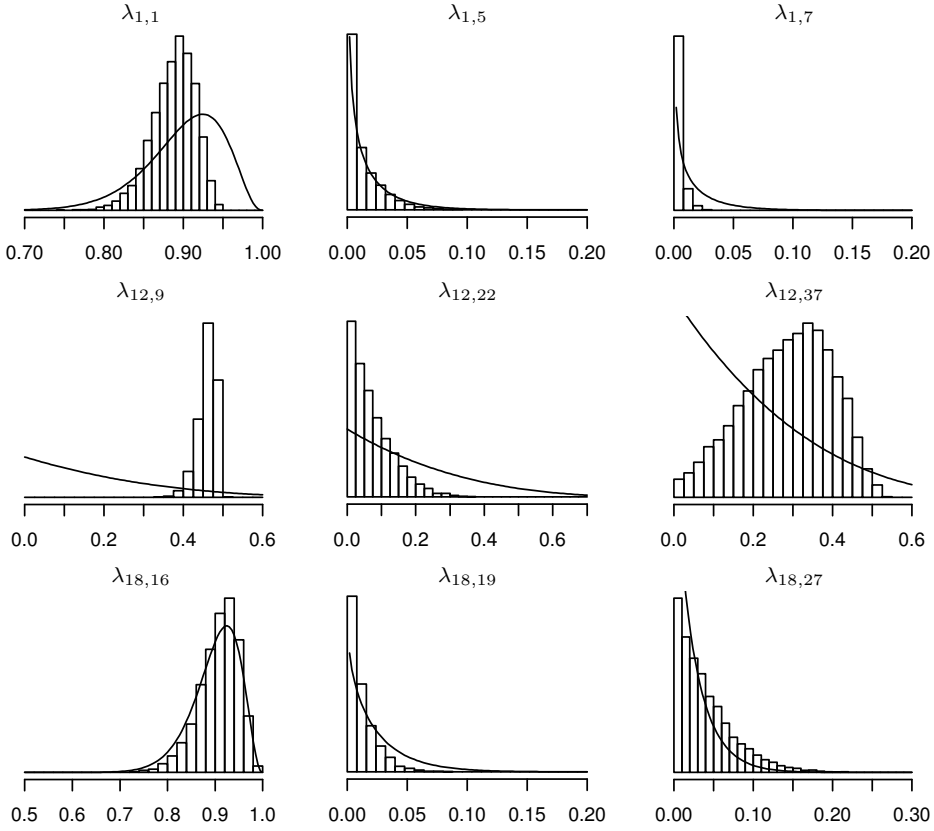


Illustration of how prior distributions are updated to posterior distributions are shown in figure 5, with prior densities being represented by solid lines. For industry No. 1 (row 1 in figure 5), we can notice that uncertainty is reduced somewhat, at least for two coefficients, and also the mean of the major one changes slightly. Estimation hardly adds any information on coefficients for industry No. 18 (row 3). We can observe, however, how the *uninformative* priors for industry No. 12 (row 2) get significantly modified due to information supplied elsewhere in the system.
Figure 6 shows absolute changes of posterior versus prior means of all transformation coefficients. The largest observed change was approximately 0.5 and is represented

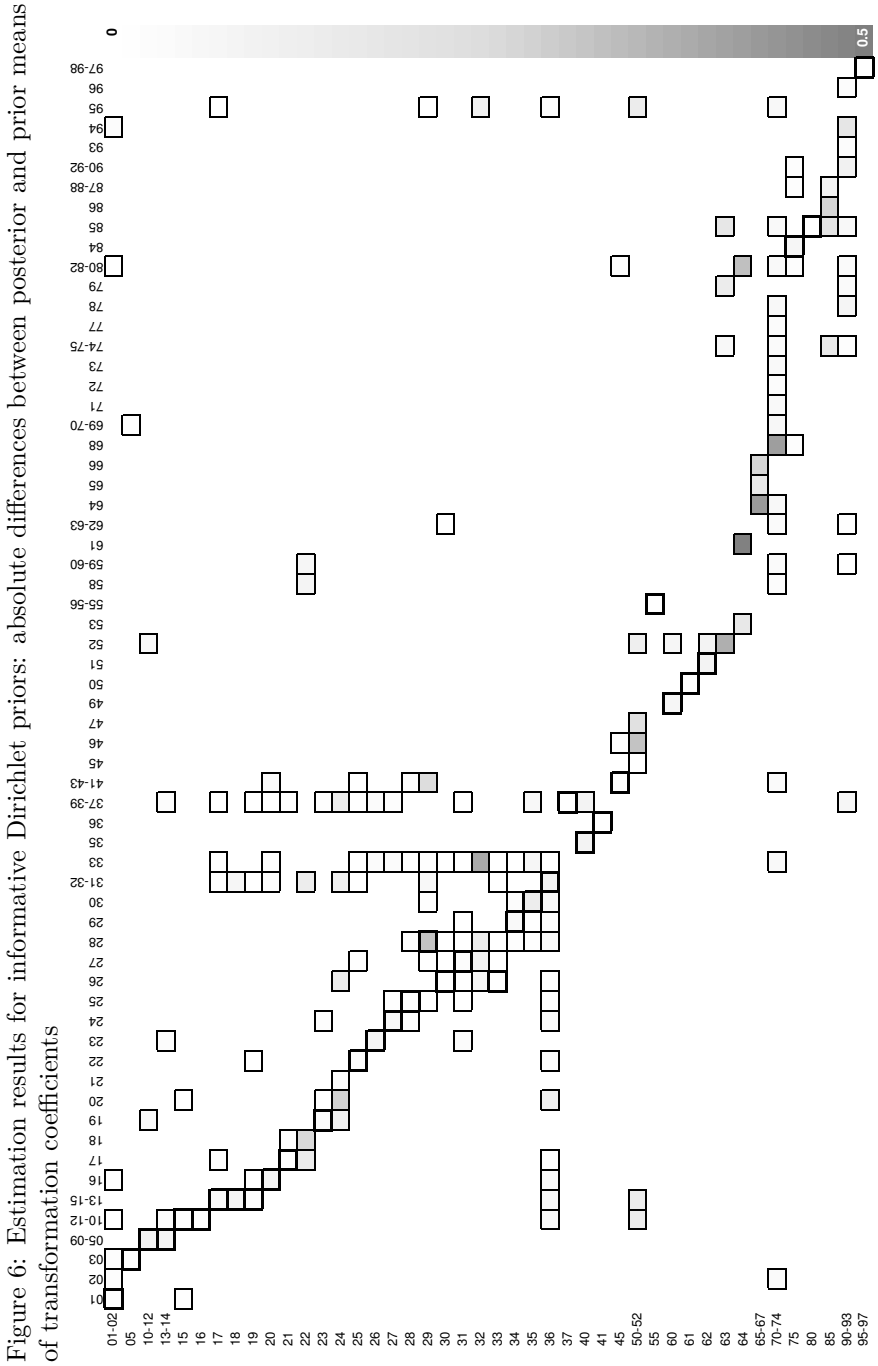


Figure 6: Estimation results for informative Dirichlet priors: absolute differences between posterior and prior means of transformation coefficients

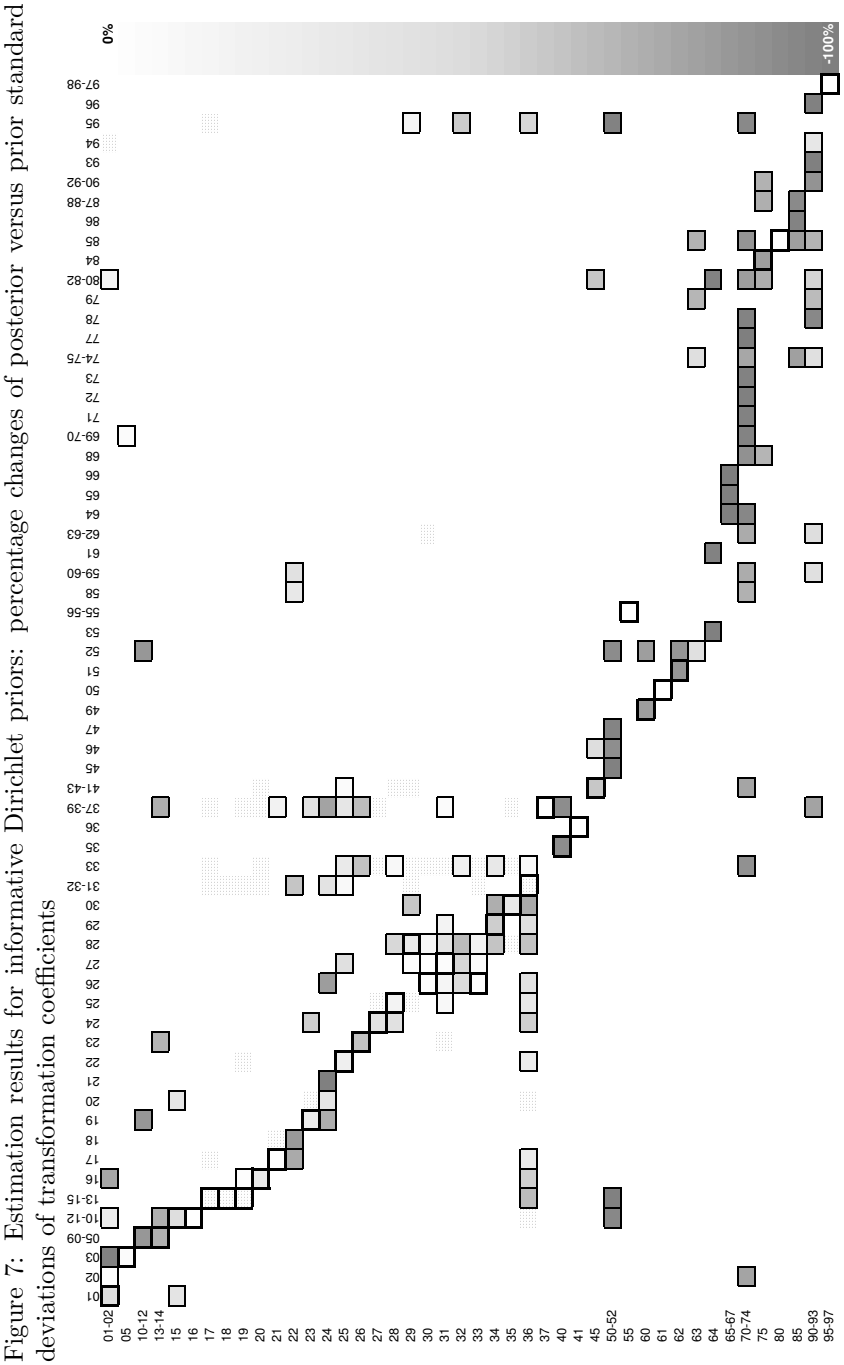


Figure 7: Estimation results for informative Dirichlet priors: percentage changes of posterior versus prior standard deviations of transformation coefficients

by dark gray, whereas white indicates no change. Again, the pattern of results is mixed, but for nearly 20% of non-zero coefficients the absolute change is larger than 0.1, pointing to a significant insight coming from the model.

Finally, figure 7 shows how prior *uncertainty* about coefficient values is changed through Bayesian inference. The contribution is measured as *percentage* reduction in standard deviation of the marginal posterior distribution versus the prior – dark gray indicating 100% reduction, and white – no change. Uncertainty is dramatically reduced for service industries (consistent with the results for uninformative priors, and a consequence of relatively straightforward links between service activities under both NACE versions; in fact, one could treat formulation of priors for each industry in isolation naive, and so initial uncertainty could be viewed naively high). For manufacturing industries the outcomes are more diverse, although a general reduction of coefficient variances is also quite evident. That general picture does not change much even having taken into account that for approximately 20% of coefficients uncertainty has actually increased (indicated by cells with dotted pattern). The latter result is a consequence of dependencies between individual coefficients, imposed by the likelihood function. Consider for example a 7-element vector of shares, having Dirichlet distribution, with mean 0.9 and standard deviation 0.05 for the first element, and equal means for the remaining components. Revision of the first component's mean to 0.8, with standard deviation unchanged, leads to an approximately 5% increase of standard deviations for all the remaining shares. A similar effect might arise e.g. when mean prior share is updated in the estimation process, and there is little information in the data to directly revise the remaining shares.

Results of sensitivity analysis, involving different choices of error standard deviations (0.2%, 5%, and 25% of observed output, y_i, respectively), are presented in figures 8-10 (the plots summarise results for 198 estimated non-zero λ_{ki} coefficients). Under uninformative prior for $\mathbf{\Lambda}$, ranges between 1st and 99th percentile of posterior distribution for each λ_{ki} are first computed. Next, differences (changes) are calculated between the ranges obtained in sensitivity runs, and their base-case estimates (i.e. estimates obtained with error standard deviations set to 1% of observed industry output) – the distributions of those differences are plotted in figure 8. It can be seen that choosing smaller standard deviations (0.2%) changes the results only slightly – the majority (83%) of ranges for λ_{ki} are revised by no more than -0.01 to 0.01; also, in most cases (63%) the ranges become slightly narrower. It should be added that it was hardly possible to compute the solution for 0.2% standard deviations (the sampling process was significantly prolonged), indicating that such an error scale was perhaps to small to accommodate inconsistencies in the data. For the 5% case, results for most coefficients do not change much either – the posterior ranges for 75% of them are revised by no more than -0.02 to 0.02. However, there is also a significant proportion of cases in which the ranges become considerably wider. In turn, the choice of 25% standard deviations implies a substantial increase in posterior uncertainty about λ_{ki} coefficients, as the third plot in figure 8 shows.

Similar conclusions also apply to the case of informative priors for $\mathbf{\Lambda}$. Here it is more appropriate to carry out the analysis in terms of sensitivity of posterior means and standard deviations rather than ranges of λ_{ki} coefficients, since their distributions are now 'less uniform'. Distributions of changes in these posterior means and standard deviations, caused by alternative choices of error scale, are shown in figures 9 and 10, respectively. In the 0.2% case, the results are largely unaffected compared to the base, 1% case (nearly 100% of posterior λ_{ki} means and standard deviations are revised by no more than -0.01 to 0.01). Increase in error standard deviations leads to a rise in uncertainty concerning λ_{ki} and changes in their mean values, although even for the largest error scale (25%) there remains a significant proportion of coefficients that are hardly affected. A general observation from sensitivity analysis (including the runs not reported here) is that ever increasing the assumed error scale eventually leads to posterior distribution of $\mathbf{\Lambda}$ being indistinguishable from the prior – the data become uninformative.

Figure 8: Sensitivity analysis for uninformative prior case: distribution of changes in posterior ranges (1st to 99th percentile) of λ_{ki} coefficients under alternative error scale choices

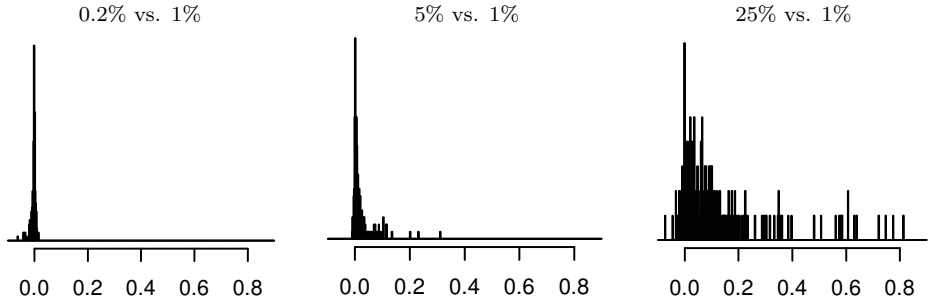


Figure 9: Sensitivity analysis for informative prior case: distribution of changes in posterior means of λ_{ki} coefficients under alternative error scale choices

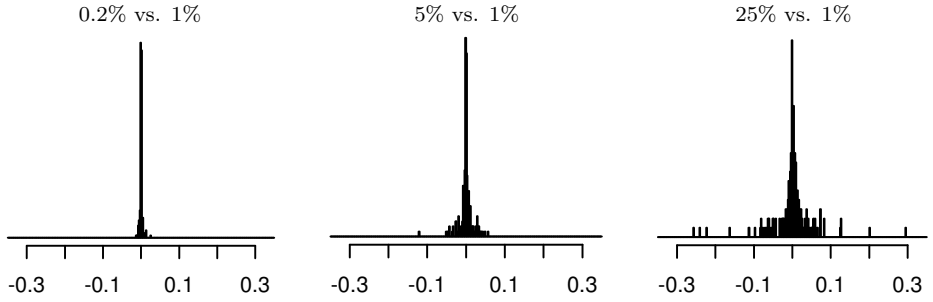
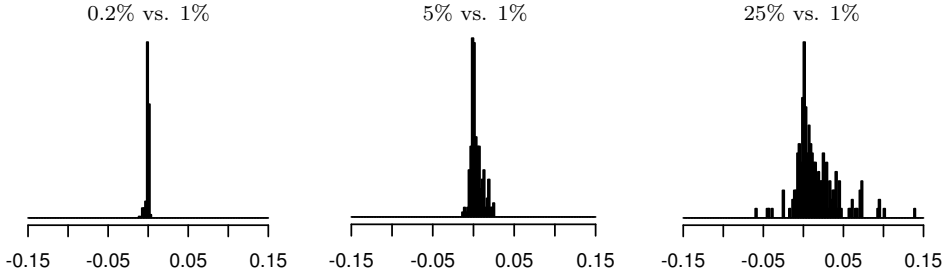


Figure 10: Sensitivity analysis for informative prior case: distribution of changes in posterior standard deviations of λ_{ki} coefficients under alternative error scale choices



5 Conclusions

In the paper we have shown how Bayesian inference, with MCMC sampling, can address practical data problems arising in the field of multi-sector economic modelling. The specific problem being solved consisted in estimating transformation matrix, allowing to convert vectors of industry data between different classifications – the previous NACE Rev. 1.1 and the current NACE Rev. 2. The proposed formulations directly apply to similarly structured problems of input-output table updating, disaggregation etc., very commonly appearing in the literature, and usually solved using bi-proportional or entropy-based techniques. We believe that Bayesian approach, as demonstrated in this paper, adds the important dimension of comprehensive uncertainty analysis, allowing to better understand the relative importance of prior knowledge versus data and accounting restrictions, indicate areas where acquiring additional information is particularly needed etc. Such an approach fits multi-sector modelling problems, characterised inevitably by limited data and important role of prior assumptions.

At this point we may attempt to sketch agenda for further developments, including:

1. model extension that would allow for joint use of data for several periods in the estimation process, supported by assumptions of invariability or a certain pattern of evolution of the estimated coefficients; in similar line, multiple data categories (beyond output) could be jointly used in estimation.

2. a possible switch to additive logistic normal distributions (replacing Dirichlet distribution), allowing for more flexibility in modelling of compositional data;

3. formulating assumptions concerning errors in terms of an informative prior;

4. applying the proposed approach to other multi-sector data problems;

5. explicitly incorporating uncertainty resulting from data estimation into simulation exercises using derived multi-sector models.

Acknowledgements

The author wishes to thank Jacek Osiewalski for insightful comments and discussion, an anonymous referee for valuable suggestions, as well as the participants of seminars at the Chair of Theory and Analyses of Economic Systems and the Chair of Econometric Models and Forecasts, University of Łódź, for critical review of earlier versions of this work.

References

[1] Aitchison, J. (1982), The statistical analysis of compositional data, *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2), 139–177.

[2] Aitchison, J. and Shen, S. (1980), Logistic-normal distributions: Some properties and uses, *Biometrika*, 67(2), 261–272.

[3] Bolshev, L., Dirichlet distribution, [in:] Rehmann, U., [ed.], *Encyclopedia of Mathematics*.

[4] van den Brakel, J. (2010), Sampling and estimation techniques for the implementation of new classification systems: the change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys, [in:] *Survey Research Methods*, volume 4, pages 103–119.

[5] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2016), Stan: A probabilistic programming language, *Journal of Statistical Software*, in press.

[6] Darroch, J. and Ratcliff, D. (1971), A characterization of the Dirichlet distribution, *Journal of the American Statistical Association*, 66(335), 641–643.

[7] Ferguson, T. (1973), A Bayesian analysis of some nonparametric problems, *Annals of Statistics*, 1(2), 209–230.

[8] Gilchrist, D. and St Louis, L. (1999), Completing input–output tables using partial information, with an application to Canadian data, *Economic Systems Research*, 11(2), 185–194.

[9] Golan, A., Judge, G. and Miller, D. (1996), *Maximum entropy econometrics: Robust estimation with limited data*, Wiley & Sons.

[10] Golan, A., Judge, G. and Robinson, S. (1994), Recovering information from incomplete or partial multisectoral economic data, *Review of Economics and Statistics*, 76(3), 541–549.

[11] Golan, A. and Vogel, S. J. (2000), Estimation of non-stationary social accounting matrix coefficients with supply-side information, *Economic Systems Research*, 12(4), 447–471.

[12] Heckelei, T., Mittelhammer, R. and Jansson, T. (2008), A Bayesian alternative to generalized cross entropy solutions for underdetermined econometric models, University of Bonn, *Institute for Food and Resource Economics Discussion Paper* No. 2.

[13] Jackson, R. and Murray, A. (2004), Alternative input-output matrix updating formulations, *Economic Systems Research*, 16(2), 135–148.

[14] Junius, T. and Oosterhaven, J. (2003), The solution of updating or regionalizing a matrix with both positive and negative entries, *Economic Systems Research*, 15(1), 87–96.

[15] Lahr, M. and De Mesnard, L. (2004), Biproportional techniques in input-output analysis: table updating and structural analysis, *Economic Systems Research*, 16(2), 115–134.

[16] Lenzen, M., Gallego, B. and Wood, R. (2009), Matrix balancing under conflicting information, *Economic Systems Research*, 21(1), 23–44.

[17] Lenzen, M., Moran, D., Geschke, A. and Kanemoto, K. (2014), A non-sign-preserving ras variant, *Economic Systems Research*, 26(2), 197–208.

[18] McDougall, R. (1999), Entropy theory and ras are friends, *GTAP Working Papers*, (6).

[19] Miller, R. and Blair, P. (2009), *Input-output analysis: foundations and extensions*, Cambridge University Press.

[20] Osiewalski, J. (2001), *Ekonometria bayesowska w zastosowaniach*, Wydawnictwo Akademii Ekonomicznej w Krakowie.

[21] Peters, J. and Hertel, T. (2016), Matrix balancing with unknown total costs: preserving economic relationships in the electric power sector, *Economic Systems Research*, 28(1), 1–20.

[22] Robinson, S., Cattaneo, A. and El-Said, M. (2001), Updating and estimating a social accounting matrix using cross entropy methods, *Economic Systems Research*, 13(1), 47–64.