Computers and Chemical Engineering 75 (2015) 196–208

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

A general framework for data reconciliation—Part I: Linear constraints

Oliver Cencic[a,*], Rudolf Frühwirth[b]

[a] Institute for Water Quality, Resource and Waste Management, Vienna University of Technology, Karlsplatz 13/226, A-1040 Wien, Austria
[b] Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences, Nikolsdorfer Gasse 18, A-1050 Wien, Austria

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form 3 October 2014
Accepted 2 December 2014
Available online 12 December 2014

Keywords:
Data reconciliation
Nonnormal distributions
Linear equality constraints
Markov chain Monte Carlo (MCMC)
Independence Metropolis–Hastings sampler
Material flow analysis (MFA)

ABSTRACT

This paper presents a new method, based on Bayesian reasoning, for the reconciliation of data from arbitrary probability distributions. The main idea is to restrict the joint prior probability distribution of the involved variables with model constraints to get a joint posterior probability distribution. This paper covers the case of linearly constrained variables, with the focus on equality constraints. The procedure is demonstrated with the help of three simple graphical examples. Because in general the posterior probability density function cannot be calculated analytically, it is sampled with a Markov chain Monte Carlo (MCMC) method. From this sample the density and its moments can be estimated, along with the marginal densities, moments and quantiles. The method is tested on several artificial examples from material flow analysis, using an independence Metropolis–Hastings sampler.

1. Introduction

The goal of material flow analysis (MFA) is to model and quantify all flows and stocks of a system of interest. For this reason as much information about the system as possible is collected which comprises direct measurements when available, but more often data taken from official statistics, reports, publications, expert estimates and similar sources (Laner et al., 2014). Unfortunately these data are often in conflict with known conservation laws such as mass or energy balances, preventing the calculation of unknown quantities or parameters of the model that cannot be measured directly. The basic idea of data reconciliation (DR) is to resolve these contradictions by statistically adjusting the collected data based on the assumption that their uncertainty is described by a probability density function.

DR has been widely used in chemical engineering for more than 50 years to adjust plant measurements. Most solving techniques that have been developed in this period of time are based on a weighted least-squares minimization of the measurement adjustments subject to constraints involving reconciled, unmeasured and fixed variables (Narasimhan and Jordache, 2000; Romagnoli and Sanchez, 2000; Bagajewicz, 2010). The underlying main assumption of this approach is that of normally distributed (Gaussian) measurement errors with zero mean (Johnston and Kramer, 1995). However, in scientific models in general and in MFA models in particular, data is often not normally distributed. If, for instance, a process model is correct, mass flows and concentrations cannot take negative values, and transfer coefficients are restricted to the unit interval.

Another example is provided by expert opinions that frequently have to be relied on in MFA due to scarce or missing data. They are often modeled by uniform, triangular or trapezoidal distributions. The more detailed the expert's knowledge about the quantity under consideration is, the more precisely the distribution can be modeled. If a sufficient number of measurements of the quantity is available, one can either fit a parametric model to the measured data or use a nonparametric model such as the empirical distribution function or the kernel estimate of the probability density function. In the following we will denote a variable as “measured” if there is prior information on the variable of any kind, which is not necessarily a proper measurement.

Although it was demonstrated in Crowe (1996) that the assumption of a normal distribution is acceptable for unknown distributions having relative standard deviations smaller than 30%, it is questionable in the context of macro-scale MFA (e.g. region, country) where relative standard deviations larger than 30% are not uncommon. In addition, the normal distribution is unsuitable to describe uncertainties with strong intrinsic asymmetry.

In the following we propose a numerical DR procedure that is also able to deal with data that cannot be modeled by normal

[*] Corresponding author. Tel.: +43 1 58801 22657; fax: +43 1 58801 9 22657.
E-mail address: oliver.cencic@tuwien.ac.at (O. Cencic).

http://dx.doi.org/10.1016/j.compchemeng.2014.12.004
0098-1354/© 2014 Elsevier Ltd. All rights reserved.

distributions. In this paper we treat the case of linearly constrained variables, with the focus on equality constraints; the cases of nonlinear and inequality constraints will be the subject of a subsequent paper. We start from the following assumptions:

1. There are N measured or unmeasured variables that take values in a subset $D \subseteq \mathbb{R}^N$.
2. The $I \leq N$ measured variables form an I-dimensional random variable with known joint density. The latter is called the prior density. The prior density can be either objective, i.e. the model of a measurement process, or subjective, i.e. the formalization of an expert opinion.
3. The variables are subject to linear equality constraints that define an affine subspace $S \subset \mathbb{R}^N$ of dimension $P < N$.

In Section 2 it is shown that the density of the variables conditional on the constraints is obtained by restricting their prior density to the set $D \cap S$ and normalizing the restricted density to 1. The resulting density is called the posterior density. The prior density plays a key role in the DR mechanism proposed below. No matter how it is obtained, it is good practice to study its influence on the posterior distribution.

In the case of a low-dimensional variable space, the construction of the posterior density can be demonstrated graphically. To show this, we present some simple examples.

Example 1.1. Let us assume that there are two measured variables x_1 and x_2 with the prior density $f(x_1, x_2)$ defined on $D \subseteq \mathbb{R}^2$. The constraint equation $x_1 = x_2$ defines a 1-dimensional subspace S, i.e. a line in $\mathbb{R}^2$. If the prior density is restricted to points on this line and normalized to 1, the posterior density of x_1, x_2 is obtained. By computing the marginal distributions of the posterior we get the posterior densities of x_1 and x_2, which are identical in this case. The values of $f(x_1, x_2)$ along S can be visualized by intersecting the prior density surface with the vertical plane through S.

Fig. 1 shows an instance of this problem, with x_1, x_2 independent, $f_1(x_1) = \gamma(x_1\,; 2, 2)$ and $f_2(x_2) = \gamma(x_2\,; 3, 1.5)$, where $\gamma(x\,; a, b)$ is the density of the Gamma distribution with parameters a and b:

$$\gamma(x; a, b) = \frac{x^{a-1}\,\mathrm{e}^{-x/b}}{b^a\,\Gamma(a)}.$$

Example 1.2. Let us assume that there are three measured variables x_1, x_2 and x_3 with the prior density $f(x_1, x_2, x_3)$ defined on $D \subseteq \mathbb{R}^3$. The constraint equation $x_3 = x_1 + x_2$ defines a 2-dimensional subspace S, i.e. a plane in $\mathbb{R}^3$. If the prior density is restricted to points in this plane and normalized to 1, the posterior density of x_1, x_2, x_3 is obtained. By computing the marginal distributions of the posterior we get the posterior densities of x_1, x_2 and x_3, respectively.

Fig. 2 shows an instance of this problem, with x_1, x_2, x_3 independent, $f_1(x_1) = \gamma(x_1\,; 2, 2)$, $f_2(x_2) = \gamma(x_2\,; 3, 1.5)$ and $f_3(x_3) = \gamma(x_3\,; 6, 1.7)$. The values of $f(x_1, x_2, x_3)$ are shown color-coded.

Example 1.3. Let us assume that there are two measured variables x_1, x_2 and one unmeasured variable x_3. The prior density of x_1, x_2, x_3 is defined on $D \subseteq \mathbb{R}^3$, but can be written as $f(x_1, x_2)$, as it does not depend on x_3. The rest of the procedure is the same as in Example 1.2. Due to the lack of an actual constraint the 2-dimensional prior density is not restricted by the 2-dimensional subspace S, the posterior densities of x_1 and x_2 are equal to the priors, and the posterior of x_3 is their convolution.

Fig. 3 shows an instance of this problem, with x_1, x_2 independent, $f_1(x_1) = \gamma(x_1\,; 2, 2)$, $f_2(x_2) = \gamma(x_2\,; 3, 1.5)$ and x_3 not measured. The values of $f(x_1, x_2)$ are shown color-coded. This example demonstrates that the method can also be used to calculate unknown variables and that it even works when the measured variables cannot be reconciled.

In the case of a nonnormal prior density, the normalization constant of the restricted density cannot in general be computed analytically. In the simple examples just discussed, it can be computed numerically by a single or a double integral. For larger dimensions of S, however, numerical integration becomes cumbersome and time-consuming. We therefore propose to avoid explicit calculation of the posterior density altogether by generating a random sample from the unnormalized restricted density. This can be achieved by applying a tool that is frequently used in Bayesian statistics (O'Hagan, 1994), namely Markov chain Monte Carlo (MCMC) (Robert and Casella, 2004; Liu, 2004; Brooks et al., 2011). The method and its implementation in the context of DR is explained in Section 3. Section 4 presents the application of MCMC to four examples in MFA. Finally, Section 5 contains our conclusions and the outlook on further work.

2. Mathematical foundation

Let $\boldsymbol{v}$ be a column vector of N measured or unmeasured variables. Following the notation in Madron (1992), we assume that $\boldsymbol{v}$ is arranged such that $\boldsymbol{v} = (\boldsymbol{y}; \boldsymbol{x})$, where $\boldsymbol{y}$ contains the J unmeasured variables and $\boldsymbol{x}$ contains the $I = N - J$ measured variables.[1] We also may have a vector $\boldsymbol{z}$ of M fixed (nonrandom) variables. DR means that $\boldsymbol{v}$ is modified in such a way that it satisfies a system of constraint equations. If all K equations are linear, the constrained system can be written in the following form:

$$\boldsymbol{By} + \boldsymbol{Ax} + \boldsymbol{Cz} = \boldsymbol{0} \quad \text{or} \quad \boldsymbol{By} + \boldsymbol{Ax} + \boldsymbol{c} = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ are known matrices of dimension $K \times I$, $K \times J$, $K \times M$, respectively, and $\boldsymbol{c}$ is a column vector of dimension $K \times 1$. We assume that

A1. $rank(\boldsymbol{B}, \boldsymbol{A}) = rank(\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{c})$, meaning the system is solvable;
A2. $rank(\boldsymbol{B}, \boldsymbol{A}) = K$, meaning the model equations are linearly independent;
A3. $rank(\boldsymbol{B}) = J$, meaning all unmeasured quantities are observable (they can be calculated).

If any of these assumptions is violated the underlying problems have to be resolved before being able to proceed. One way to achieve this goal is to apply the Gauss-Jordan elimination to matrix $(\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{c})$. The result, known as the reduced row echelon form (or canonical form), serves to detect contradictions (A1), to eliminate dependent equations automatically (A2) and to classify variables, in particular to identify and eliminate unobservable unmeasured variables (A3). For detailed instructions how to proceed see Madron (1992, p. 125). There exist alternative equation-oriented approaches for variable classification (Romagnoli and Sanchez, 2000, p. 33), but in our opinion the Gauss–Jordan elimination is the easiest to understand.

We make further use of the reduced row echelon form in order to identify dependent and free variables of the system. The column numbers of the pivot elements (leading 1 in each row) denote the dependent variables, which can be unmeasured or measured ones. All other variables, which have to be measured ones, are designated as free. The outcome of this classification process depends on the initial order of the variables. Although the posterior density itself is unique, the choice of the free variables can affect its computation, so the initial order of the variables should be chosen carefully (see Section 3.2 and Example 4.4).

[1] The semicolon (comma) denotes vertical (horizontal) concatenation of matrices.