Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Metabolic Flux Estimation from Incomplete Labelling Measurements Using the Expectation/Conditional-Maximisation Algorithm

S. Wongsa *, V. Kadirkamanathan *, S.A. Billings * and P.C. Wright †

* Department of Automatic Control and Systems Engineering
University of Sheffield, Sheffield
S1 3JD United Kingdom
Email: {s.wongsa, visakan, s.billings}@shef.ac.uk
† Department of Chemical and Process Engineering
University of Sheffield, Sheffield
S1 3JD United Kingdom
Email: p.c.wright@shef.ac.uk

Abstract— **In this work, the problem of metabolic flux estimation is formulated as a problem of parameter estimation from incomplete labelling data. The expectation/conditional maximisation (ECM) algorithm is used to determined a maximum-likelihood (ML) estimate because of its simplicity and stable convergence. We propose to simplify a nonlinear inverse problem, generally numerically solved by an iterative optimisation algorithm, to a linear regression problem which is arrived at from a linear-in-the-parameter formulation during a partial optimisation process of the ECM algorithm. Three linear least square algorithms, the ordinary least squares (LS), the total least squares (TLS) and the constrained least squares (CLS), have been tested to solve the linear regression in this step. Using simulations, resulting parameter estimates and errors in flux estimation are compared and evaluated. The performance of the algorithms are investigated under two scenarios; when the labelling data are corrupted by a wide range of noise and when the labelling data are incompletely observed. Results suggest that the estimates from the ECM algorithm using CLS produce results superior to other combinations and have potential to be refined to improve its performance in metabolic flux estimation.**

I. Introduction

Metabolic fluxes through metabolic networks are crucial for cell function and a knowledge of these fluxes is essential for understanding and manipulating metabolic phenotypes [8]. Metabolic flux estimation based on data from ^{13}C-tracer experiments has rigorously been studied in the last decades to quantify the intracellular metabolic fluxes since they can rarely be measured directly [7], [12].

The most informative data about the fluxes can be obtained by conducting isotopomer tracer experiments. From the measurements of the ^{13}C enrichments of the intracellular metabolites, the information about the fluxes of the alternative pathways producing a metabolite can be obtained [13].

Based on the isotope balance equations, if the information of the variances of process and measurement noises is provided, the flux estimation can be formulated as a nonlinear least squares problem where the current fluxes are processed to attempt to minimise the (weighted) deviation between measured data and simulated measurements [6], [12]. To estimate the unknown fluxes and the error variances simultaneously, the maximum-likelihood (ML) method is applicable. However, the main difficulty of the ML approach is the high computational complexity caused by optimisation of the likelihood function, especially in an incomplete data situation and a slow convergence speed that makes inference of larger systems infeasible.

To overcome this difficulty, a metabolic flux estimation approach based on the Expectation/Conditional Maximisation (ECM) algorithm [5] has been derived to simplify the ML estimation. The basic idea is as follows: by imputing the missing data from the conditional expectations of the labelling data obtained from the E-step, a linear-in-the-parameter model can be formulated in the conditional maximisation step. The unknown fluxes in the next iteration are then easily solved using a standard linear least square technique. In this paper we examine the feasibility of the developed method by simulating and estimating the intracellular fluxes in the central metabolic pathway of *Corynebacterium glutamicum.*

II. Problem Formulation

A. Metabolic Flux Estimation

The purpose of metabolic flux estimation is to obtain an estimate $\hat{\theta}$, which is as close as possible to the 'true' flux distribution, given the pathway structure of the model. In isotopic steady state, the vector representation of metabolite atom specific activities, combined with atom mapping matrices, enables the equations governing the isotopic steady state to be written in the following form [11], [15]:

$$\mathbf{A}_\theta \mathbf{x} = \mathbf{u} \qquad (1)$$

where $\mathbf{A}_\theta \in \mathbb{R}^{n \times n}$ is the flux matrix which is a function of the metabolic fluxes $\theta \in \mathbb{R}^l$, $\mathbf{x} \in \mathbb{R}^n$ is the vector of positional enrichments of the intracellular metabolites and $\mathbf{u}$ is the vector of measured quantities, e.g. the uptake rates and the extracellular fluxes. The subscript θ will be used for any variable which is dependent on the metabolic flux. If $\mathbf{A}_\theta$ is assumed to be nonsingular, the positional enrichment vector $\mathbf{x}$ can be obtained by inverting $\mathbf{A}_\theta$:

$$\mathbf{x} = \mathbf{A}_\theta{}^{-1}\mathbf{u} \tag{2}$$

Note that in certain pathological situations, the network structure can become disconnected when fluxes vanish, i.e. the forward and backward rates of a particular reaction are equal, and the matrix $\mathbf{A}_\theta$ consequently becomes singular. To prevent the singularity of $\mathbf{A}_\theta$, the net fluxes should be bounded by non-zero lower bounds [11].

In the presence of internal uncertainty, e.g. the simplification of the proposed structure with main fluxes represented to the true network structure, a noise term is added to (2):

$$\begin{aligned} \mathbf{x} &= \mathbf{A}_\theta^{-1}\mathbf{u} + \mathbf{w} \\ &= \mathbf{m}_\theta + \mathbf{w} \end{aligned} \tag{3}$$

where $\mathbf{m}_\theta = \mathbf{A}_\theta^{-\mathbf{1}}\mathbf{u}$ and $\mathbf{w}$ is the modelling error which is assumed to be normally distributed with zero mean and covariance matrix $\mathbf{Q}_w = \sigma_w^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix with corresponding dimension and $\sigma_w > 0$. We now consider two scenarios regarding the availability of the measurements:

1) Complete Data Case: It can be shown that if all positional enrichments of the intracellular metabolites are available (1) can be written as a linear-in-the-flux model [14]:

$$\mathbf{A}_x\theta = \mathbf{u} \tag{4}$$

where $\mathbf{A}_x \in \mathbb{R}^{n\times l}$ is the metabolite matrix which is a function of the positional enrichments (or higher representation level, e.g. the isotopomer distributions) and $\mathbf{u}$ is as defined in (2). Formulating the system equation as (4) enables linear least squares techniques to be applied:

- **Ordinary linear least squares (LS)**
  If we assume all the errors are in the extracellular fluxes $\mathbf{u}$, (4) is replaced by

  $$\mathbf{A}_x\theta = \mathbf{u} + \varepsilon. \tag{5}$$

  The unknown fluxes can be estimated by minimising the sum of the squares of the residuals, i.e.

  $$\min_\theta \|\mathbf{A}_x\theta - \mathbf{u}\|_2. \tag{6}$$

  If rank$(\mathbf{A}_x) = \min(n, l)$, the unique least squares solution θ_{LS} is given by the pseudoinverse of $\mathbf{A}_x$, $\hat{\theta}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{u} = \mathbf{A}^\dagger\mathbf{u}$ [1]. However, if the system is rank-deficient, i.e. rank$(\mathbf{A}_x) < \min(n, l)$, a unique solution which minimises $\|\theta\|_2$ is calculated instead using the singular value decomposition (SVD) [1].
- **Total least squares (TLS)**
  In the classical least squares approach, the matrix $\mathbf{A}_x$is assumed to be free of error, which is not strictly correct for metabolic system equation as the positional enrichments $\mathbf{x}$ are normally corrupted by noise, i.e.

  $$(\mathbf{A}_x + \mathbf{E})\theta = \mathbf{u} + \varepsilon \tag{7}$$

  To take errors in $\mathbf{A}_x$ into account, the problem should be tackled as a total least squares (TLS) problem where perturbations are allowed in $\mathbf{A}_x$. The total least squares solution of (7) is the minimum norm of

  $$\min_{\mathbf{E},\varepsilon}\|(\mathbf{E}\ \ \varepsilon)\|_F \ \text{ subject to } \ (\mathbf{u}+\varepsilon) \ \in \ \text{range}(\mathbf{A}+\mathbf{E}) \tag{8}$$

  Here $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{E}$ and ε are the errors of $\mathbf{A}_x$ and $\tilde{\mathbf{u}}$, respectively.
  Making use of SVD and the reduced-rank approximation theorem, if $\hat{\sigma}_l > \sigma_{l+1}$ where $l = \dim\theta$, $\hat{\sigma}_n$ and σ_n are respectively the nth singular values of the matrix $\mathbf{A}_x$, and the augmented matrix $(\mathbf{A}_x\ \ \mathbf{u})$, a unique minimum norm solution of the TLS exists as follows [9]:

  $$\hat{\theta}_{TLS} = (\mathbf{A}_x^T\mathbf{A}_x - \sigma_{l+1}^2\mathbf{I})^{-1}\mathbf{A}_x^T\mathbf{u} \tag{9}$$

2) Incomplete Data Case: With current technology some components of the positional enrichments of the intermediate metabolites cannot be measured, i.e. some components of $\mathbf{x}$ are missing [13]. While measuring only the output $\mathbf{y} \in \mathbb{R}^m$, the measurement equation can be written as

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v} \tag{10}$$

where $\mathbf{C} \in \mathbb{R}^{m\times n}$ is the output matrix which is a constant matrix of a linear combination of states $\mathbf{x}$, $\mathbf{v}$ represents the measurement error and is assumed to be an additive zero-mean homogeneous Gaussian distributed process with an $m \times m$ covariance matrix $\mathbf{Q}_v = \sigma_v^2\mathbf{I}$ and $\sigma_v > 0$. Therefore, the problem is how to estimate the unobserved fluxes indirectly through the observed variables. The Maximum-Likelihood (ML) criterion serves as a benchmark to choose the parameter values for which the observed data is most likely to occur.

Given the pdf of the measurements $\mathbf{y}$ for a given unknown parameters $\mathbf{\Psi} = (\theta, \sigma_w, \sigma_v)^T$

$$p(\mathbf{y}|\mathbf{\Psi}) = \frac{\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{g}_\theta)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{g}_\theta)\}}{\sqrt{2\pi^m \det|\mathbf{R}|}} \tag{11}$$

with the mean $\mathbf{g}_\theta$, and covariance $\mathbf{R}$ given by

$$\mathbf{g}_\theta = \mathbf{C}\mathbf{A}_\theta^{-1}\mathbf{b} = \mathbf{C}\mathbf{m}_\theta \tag{12}$$

$$\mathbf{R} = \mathbf{C}\mathbf{Q}_w\mathbf{C}^T + \mathbf{Q}_v \tag{13}$$

The unknown parameters $\mathbf{\Psi}$ =$\{\theta, q_w, q_v\}^T$ are estimated by maximising the likelihood function over the parameter space $\mathbf{\Omega}$, i.e.

$$\hat{\theta}_{ML} = \arg\max_{\mathbf{\Psi}\in\mathbf{\Omega}}\{-\log(\det|\mathbf{R}|) - (\mathbf{y}-\mathbf{g}_\theta)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{g}_\theta)\}. \tag{14}$$

If the covariance matrices $\mathbf{Q}_w$ and $\mathbf{Q}_v$ are known, (14) can be cast as a weighted nonlinear least square problem and hence can be solved by a nonlinear least squares algorithm, e.g. the Levenberg-Marquardt and the sequential quadratic programming as applied in [3], [12]. A drawback of the ML estimation (14) is that it specifies a complicated nonlinear optimisation problem in several variables and the estimation often becomes considerably hard with high computational effort. An EM algorithm which will be described in the next section has been developed to simplify the direct ML estimation.

III. EM Algorithm Applied to Metabolic Flux Estimation

Each iteration of the EM algorithm consists of two steps: The E (expectation) step and the M (Maximisation) step. Let the vector $\mathbf{y} = (y_1, ..., y_m)^T$ denote the observed incomplete data and the latent/unobserved data be $\mathbf{x} = (x_1, ..., x_n)^T$. The E step consists of computing the expectation of the complete-data log-likelihood $\log p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi})$ with respect to $p(\mathbf{x}|\mathbf{y}, \hat{\mathbf{\Psi}}^k)$ where $\hat{\mathbf{\Psi}}^k$ denotes the estimated parameters at the kth iteration, i.e.

$$\begin{aligned}\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k) &= E[\log p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi})|\mathbf{y}, \hat{\mathbf{\Psi}}^k] \\ &= \int_{\mathbf{x}} \log p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi}) p(\mathbf{x}|\mathbf{y}, \hat{\mathbf{\Psi}}^k) d\mathbf{x} \qquad (15)\end{aligned}$$

In the M step, $\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k)$ is maximised with respect to $\mathbf{\Psi}$, over the parameter space $\mathbf{\Omega}$. This leads to a new parameter estimate $\hat{\mathbf{\Psi}}^{k+1}$:

$$\hat{\mathbf{\Psi}}^{k+1} = \arg\max_{\mathbf{\Psi}\in\mathbf{\Omega}} \mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k) \qquad (16)$$

The algorithm is iterated until $\|\mathbf{\Psi}^{k+1} - \mathbf{\Psi}^k\|$ or $\|\mathbf{Q}(\mathbf{\Psi}^{k+1}; \Psi^k) - \mathbf{Q}(\mathbf{\Psi}^k; \mathbf{\Psi}^k)\|$ is sufficiently small.

A. The E Step of the Algorithm

In metabolic flux estimation based on tracer experiments, not only are some positional enrichments not available but also the available measurements are noisy, which therefore establishes the incomplete data. Here we use the complete positional enrichment $\mathbf{x}$ as the latent data because it is unknown if measurements include noise.

To evaluate (15), we first derive the joint probability density function $p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi})$ from

$$\begin{aligned}p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi}) &= p(\mathbf{y}|\mathbf{x}, \mathbf{\Psi}) \cdot p(\mathbf{x}|\mathbf{\Psi}) \\ &= \frac{1}{\sqrt{2\pi^{(n+m)}\det|\mathbf{Q}_w\mathbf{Q}_v|}} \\ &\quad \cdot \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_\theta)^T \mathbf{Q}_w^{-1}(\mathbf{x} - \mathbf{m}_\theta) \\ &\quad -\frac{1}{2}(\mathbf{y} - \mathbf{C}\mathbf{x})^T \mathbf{Q}_v^{-1}(\mathbf{y} - \mathbf{C}\mathbf{x})\} \qquad (17)\end{aligned}$$

Substituting (17) into (15) and applying the expectation to each term with respect to $\mathbf{x}$ given $\mathbf{y}$ and θ^k yields:

$$\begin{aligned}\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k) =& -\frac{n+m}{2}\log(2\pi) - \frac{1}{2}\log(\det|\mathbf{Q}_w\mathbf{Q}_v|) \\ & -\frac{1}{2}(E[\mathbf{x}] - \mathbf{m}_\theta)^T \mathbf{Q}_w^{-1}(E[\mathbf{x}] - \mathbf{m}_\theta) \\ & -\frac{1}{2}\mathrm{tr}(\mathbf{Q}_w^{-1}(E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E^T[\mathbf{x}])) \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{C}E[\mathbf{x}])^T \mathbf{Q}_v^{-1}(\mathbf{y} - \mathbf{C}E[\mathbf{x}]) \\ & -\frac{1}{2}\mathrm{tr}(\mathbf{Q}_v^{-1}\mathbf{C}(E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E^T[\mathbf{x}])\mathbf{C}^T)\end{aligned} \qquad (18)$$

where $\mathrm{tr}(\mathbf{X}) = \sum_i \mathbf{X}_{ii}$ is the trace operator of a square matrix $\mathbf{X}$ and all expectations are with respect to $\mathbf{x}$, given $\mathbf{y}$ and θ^k.

The conditional autocorrelation matrix $E[\mathbf{x}\mathbf{x}^T|\mathbf{y}, \mathbf{\Psi}^k]$ and the conditional vector $E[\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k]$ can be obtained from the conditional mean and covariance of $p(\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k)$ (22):

$$E[\mathbf{x}\mathbf{x}^T|\mathbf{y}, \mathbf{\Psi}^k] = \mathbf{P}^k + \hat{\mathbf{x}}^k\hat{\mathbf{x}}^{k^T} \qquad (19)$$

$$E[\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k] = \hat{\mathbf{x}}^k \qquad (20)$$

$\hat{\mathbf{x}}^k$ and $\mathbf{P}^k$ denote the conditional mean and covariance of $p(\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k)$.

The conditional pdf $p(\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k)$ is derived by the rule of conditional probability, i.e.

$$p(\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k) = \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi}^k)}{p(\mathbf{y}|\mathbf{\Psi}^k)} \qquad (21)$$

$p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi}^k)$ is already derived in (17). If $p(\mathbf{x}, \mathbf{y}|\mathbf{\Psi}^k)$ is Gaussian, it can then be expressed in the following form:

$$p(\mathbf{x}|\mathbf{y}, \mathbf{\Psi}^k) = \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}}^k)^T \mathbf{P}^{k^{-1}}(\mathbf{x} - \hat{\mathbf{x}}^k)\}}{\sqrt{2\pi^m \det|\mathbf{P}^k|}} \qquad (22)$$

Here $\hat{\mathbf{x}}^k$ and $\mathbf{P}^k$ denote the conditional mean and covariance of $\mathbf{x}$, respectively, and are given by:

$$\hat{\mathbf{x}}^k = \mathbf{P}^k(\mathbf{C}^T\mathbf{Q}_v^{k^{-1}}\mathbf{y} + \mathbf{Q}_w^{k^{-1}}\mathbf{m}_\theta^k) \qquad (23)$$

$$\mathbf{P}^k = (\mathbf{C}^T\mathbf{Q}_v^{k^{-1}}\mathbf{C} + \mathbf{Q}_w^{k^{-1}})^{-1} \qquad (24)$$

Substituting (19) and (20) into (18) and multiplying (18) by 2 yields another form of the conditional expectation of the complete-data log likelihood

$$\begin{aligned}\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k) =& C_0 - \log(\det|\mathbf{Q}_w|) - \log(\det|\mathbf{Q}_v|) \\ & -(\hat{\mathbf{x}}^k - \mathbf{m}_\theta)^T \mathbf{Q}_w^{-1}(\hat{\mathbf{x}}^k - \mathbf{m}_\theta) - \mathrm{tr}(\mathbf{Q}_w^{-1}\mathbf{P}^k) \\ & -(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}^k)^T \mathbf{Q}_v^{-1}(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}^k) \\ & -\mathrm{tr}(\mathbf{Q}_v^{-1}\mathbf{C}\mathbf{P}^k\mathbf{C}^T)\end{aligned} \qquad (25)$$

where C_0 is a constant term.

B. The M Step of the Algorithm

The M step determines new parameter values that maximises $\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k)$ with respect to $\mathbf{\Psi}$; that is

$$\mathbf{Q}(\mathbf{\Psi}^{k+1}; \hat{\mathbf{\Psi}}^k) \geq \mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k) \tag{26}$$

with parameters to be identified $\mathbf{\Psi} = (\theta, \sigma_w, \sigma_v)^T$. As can be seen from (25), there is only one flux-dependent term, i.e. $(\hat{\mathbf{x}}^k - \mathbf{m}_\theta)^T \mathbf{Q}_w^{-1} (\hat{\mathbf{x}}^k - \mathbf{m}_\theta)$. The Q-function $\mathbf{Q}(\mathbf{\Psi}; \hat{\mathbf{\Psi}}^k)$ is a quadratic function of the conditional mean $\mathbf{m}_\theta$ and the optimal value of $\mathbf{m}_\theta$ given θ^k is obviously the expected mean $\hat{\mathbf{x}}^k$. Given $\mathbf{m}_\theta^{k+1}$, θ^{k+1} can be obtained by taking the inverse function of the optimal $\mathbf{m}_\theta$, i.e. $\theta = \mathcal{F}^{-1}(\mathbf{m}_\theta)$. However, the inverse function $\mathcal{F}^{-1}(\mathbf{m}_\theta)$ is nonlinear with respect to θ due to the inverse of the matrix $\mathbf{A}_\theta$. Therefore the M-step has no closed form solution and a numerical optimisation procedure is usually required.

Alternatively, in this work we will avoid the use of iterative optimisation for the M-step. Instead of finding $\mathcal{F}^{-1}(\mathbf{m}_\theta)$, we exploit the GEM framework [4] and the relationship of equations (1) and (4). Rather than solving for the optimal θ, we find an appropriate value and associated parameters σ_v and σ_w such that

$$\mathbf{Q}(\mathbf{\Psi}^{k+1}; \hat{\mathbf{\Psi}}^k) \geq \mathbf{Q}(\mathbf{\Psi}^{k+(S-1)/S}; \hat{\mathbf{\Psi}}^k) \geq \cdots \geq \mathbf{Q}(\mathbf{\Psi}^k; \hat{\mathbf{\Psi}}^k) \tag{27}$$

The value of $\mathbf{\Psi}$ on the sth CM-step of the $(k+1)$th iteration is denoted by $\mathbf{\Psi}^{k+s/S}$ where S is the number of CM steps of the ECM algorithm. It can be shown that the ECM algorithm possesses the convergence properties of the GEM algorithm [4].

The appropriate framework is motivated by the fact that $\hat{\mathbf{x}} = \mathbf{m}_\theta$, i.e. $\mathbf{Q}$ is optimum, implying that $\mathbf{A}_{\hat{x}}\theta = \mathbf{u}$ and therefore, the appropriate solution can be obtained using the various least squares solutions. Once θ^{k+1} is obtained the noise variances can be calculated by the optimisation of the Q-function over σ_w and σ_v.

The concept of partial optimisation of the expected complete-data log-likelihood over each parameter with all other parameters held fixed is introduced by [5] and is known as the Expectation/Conditional Maximisation (ECM) algorithm which is a class of the Generalised Expectation Maximisation (GEM) algorithm for which the M-step requires $\mathbf{\Psi}^{k+1}$ to be chosen such that the Q-function increases rather than maximise it over all $\mathbf{\Psi} \in \mathbf{\Omega}$ [4].

The pseudo-code for the ECM algorithm applied to flux estimation is as follows:

1) Generate an initial estimate $\{\theta^k, \sigma_v^k, \sigma_w^k\}$
2) **CM1** - Identification of θ^{k+1}:
   Given current θ^k, calculate $\mathbf{A}_\theta^k$ and $\hat{\mathbf{x}}^k$ from the isotopic balance equations and (23), respectively. Formulate the vector $\tilde{\mathbf{u}}$ and the matrix $\mathbf{A}_x^k$ of (5) by treating the conditional estimate $\hat{\mathbf{x}}^k$ as a set of *complete* positional enrichments. θ^{k+1} is hence obtained using a linear least squares technique.
3) **CM2** - Identification of σ_v^{k+1}:
   The measurement noise variance σ_v^2 at the $(k+1)$th iteration has a closed form:

$$\begin{aligned}\sigma_v^{2^{k+1}} &= \frac{1}{m}(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}^{k+1/3})^T(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}^{k+1/3}) \\ &\quad + \mathrm{tr}(\mathbf{C}\mathbf{P}^{k+1/3}\mathbf{C}^T)\end{aligned} \tag{28}$$

   where $\hat{\mathbf{x}}^{k+1/3}$ and $\mathbf{P}^{k+1/3}$ (using (24)) are calculated from $\mathbf{\Psi}^{k+1/3} = \{\theta^{k+1}, \sigma_v^k, \sigma_w^k\}$ and m is the number of measured positional enrichments.
4) **CM3** - Identification of $\sigma_w^{2^{k+1}}$:
   The process noise variance σ_w^2 at the $(k+1)$th iteration also has a closed form:

$$\begin{aligned}\sigma_w^{2^{k+1}} &= \frac{1}{n}(\hat{\mathbf{x}}^{k+2/3} - \mathbf{A}_\theta^{k+1}\mathbf{b})^T(\hat{\mathbf{x}}^{k+2/3} - \mathbf{A}_\theta^{k+1}\mathbf{b}) \\ &\quad + \mathrm{tr}(\mathbf{P}^{k+2/3})\end{aligned} \tag{29}$$

   where $\hat{\mathbf{x}}^{k+2/3}$ and $\mathbf{P}^{k+2/3}$ are calculated from $\mathbf{\Psi}^{k+2/3} = \{\theta^{k+1}, \sigma_v^{k+1}, \sigma_w^k\}$ step and n is the number of positional enrichments of all intermediate metabolites in the pathway.
5) If $\frac{\|\mathbf{\Psi}^{k+1} - \mathbf{\Psi}^k\|}{\|\mathbf{\Psi}^k\|} < \tau_1$ or $\frac{\|\mathbf{Q}(\mathbf{\Psi}^{k+1}; \Psi^k) - \mathbf{Q}(\mathbf{\Psi}^k; \mathbf{\Psi}^k)\|}{\|\mathbf{Q}(\mathbf{\Psi}^k; \mathbf{\Psi}^k)\|} < \tau_2$, where τ_1 and τ_2 are preselected thresholds, then accept $\mathbf{\Psi}^{k+1}$ as the estimated parameter $\hat{\mathbf{\Psi}}$ and exit, otherwise return to step 2.

C. Reduction of solution space using the free fluxes

It is known that the EM algorithm sometimes is very slow to converge and the rate of convergence is linear and depends on the proportion of information in the observed data [4]. To improve the convergence speed, the concept of the *free fluxes* [12] is applied to reduce the solution space. The assumptions on the model stoichiometry, the flux directions and flux measurements provide additional linear equality constraints of a form $\mathbf{A}_{eq}\theta = \mathbf{b}_{eq}$ to the flux estimation problem in **CM1** step, resulting in a reduction of the search space.

By applying standard techniques of linear algebra to the linear constraint equation, we can find a set of fluxes, known as *free fluxes* , whose values are sufficient to fix the whole flux distribution [2], [11]. The dimension of the free fluxes ϕ is $r =$ dim $(\theta)-$ rank $(\mathbf{A}_{eq})$. Given the free fluxes ϕ, the solution space of the equality constraints is parameterised by

$$\theta = \mathbf{\Gamma} \cdot \phi + \theta_0 \tag{30}$$

where θ_0 is the vector of fixed solution derived from the flux measurements, $\mathbf{\Gamma}$ is the vector space spanned by the null space of $\mathbf{A}_{eq}$.

In practice, the complete flux distributions can be simply computed by using a matrix inversion. By adding the constraint of free fluxes $\mathbf{N}_f \cdot \theta = \mathbf{n}_f$ to the linear equality constraint of the system stoichiometry and measured fluxes, the complete flux distributions is given in a form:

$$\theta = \tilde{\mathbf{A}}_{eq}^{-1} \cdot \tilde{\mathbf{b}}_{eq} \tag{31}$$

where $\tilde{\mathbf{A}}_{eq} = \binom{\mathbf{A}_{eq}}{\mathbf{N}_f}$ and $\tilde{\mathbf{b}} = \binom{\mathbf{b}_{eq}}{\mathbf{n}_f}$.

To apply the free fluxes to the ECM algorithm, the parameter $\mathbf{\Psi}$ becomes $\{\phi, \sigma_v, \sigma_w\}$ and the linear parametrisation (30) is added to each of the CM steps to map from the searched parameter ϕ to the actually required parameter θ.

IV. Computational Experiments

We tested the developed ECM algorithm to estimate the intracellular fluxes of the model of central carbon metabolism of lysine-producing *Corynebacterium glutamicum* on glucose containing glycolysis, pentose phosphate pathway and citric acid cycle. The reaction equations and carbon mappings presented in [12] were used.

A. Flux model

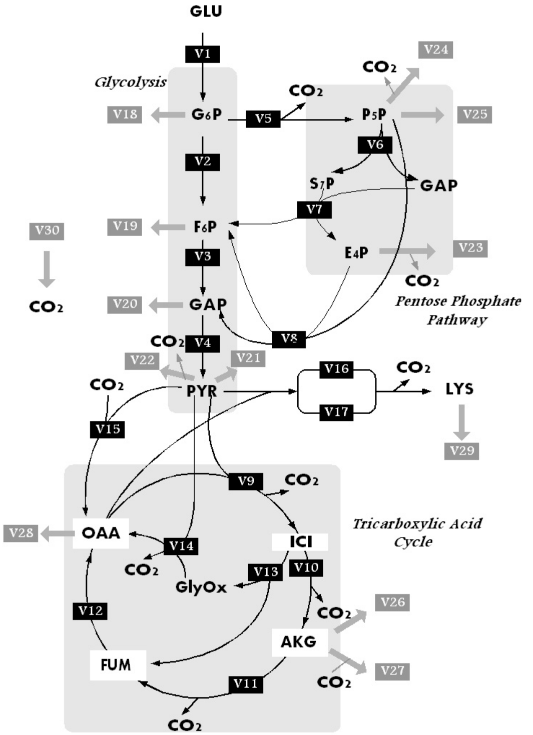


Fig. 1. Metabolic model for the central metabolism of L-lysine-producing *Corynebacterium glutamicum* (adapted from [12]). Thin arrows represent the reactions in the pathway. Shaded arrows indicate the withdrawal of precursors for biosynthesis. V12 and V16 are scrambling reactions. Abbreviations: GLU , glucose; LYS, lysine; G6P, glucose 6-phosphate; F6P, fructose 6-phosphate; GAP, glyceraldehyde 3-phosphate; PYR, pyruvate; P5P, pentose 5-phosphate; E4P, erythrose 4-phosphate; S7P, sedoheptulose; OAA, oxaloacetate; ICI, isocitrate; AKG, α-ketaglutarate;FUM, fumarate; GlyOx, glyoxalase; CO2, CO_2.

The metabolic network for lysine-producing *C. glutamicum* on glucose is shown in Fig. 1. The network consisted of 15 metabolites (excluding 10 biomass metabolites) and 30 reactions of which 7 were bidirectional and of which 14 (1 substrate uptake (V1) + 11 biomass effluxes (V18-V28) + 2 product formulation of lysine (V29-V30)) were assumed to be measurable. Table I summarises the stoichiometric reactions ,the corresponding carbon atom transitions and assumptions on the reaction directionality which were used in this metabolic model. The number of net fluxes for modelling purposes was 32 of which 2 reactions came from the scrambling reactions (V12 and V16). The scrambling reaction Vi consists of two parallel reactions which are referred to as Vi^1 and Vi^2.

TABLE I

The central metabolic network in *C. glutamicum* with the stoichiometric reactions and the corresponding carbon atom transitions. Bidirectional reactions are indicated by double headed arrows. Complete scrambling was assumed for the fumarase reactions $V12^1$ and $V12^2$ and lycine productoin via the succinyl-diaminopimelate dehydrogenase pathway ($V16^1$ and $V16^2$).

Glycolysis	
V1 :GLU + GAP → G6P + PYR	ABCDEF +abc → ABCDEF + abc
V2 :G6P ↔ F6P	ABCDEF ↔ ABCDEF
V3 :F6P ↔ GAP + GAP	ABCDEF ↔ CBA + DEF
V4 :GAP ↔ PYR	ABC ↔ ABC
Pentose phosphate pathway	
V5 :G6P → CO2 + P5P	ABCDEF ↔ A + BCDEF
V6 :P5P + P5P ↔ S7P + GAP	ABCDE + abcde ↔ ABabcde + CDE
V7 :GAP + S7P ↔ E4P + F6P	ABC + abcdefg ↔ defg + abcABC
V8 :P5P + E4P → GAP + F6P	ABCDE + abcd → CDE + ABabcd
Tricarboxylic acid cycle and glyoxylate cycle	
V9 :PYR + OAA → ICI + CO2	abc + ABCD → DCBAcb + a
V10 :ICI → AKG + CO2	ABCDEF → ABCEF + D
V11 :AKG → FUM + CO2	ABCDE ↔ BCDE + A
$V12^1$:FUM ↔ OAA	ABCD ↔ ABCD
$V12^2$:FUM ↔ OAA	ABCD ↔ DCBA
V13 :ICI → GlyOx + FUM	ABCDEF → AB + CDEF
V14 :GlyOx + PYR → OAA + CO2	AB + abc → ABba + c
Anaplerotic pathway	
V15 :PYR + CO2 ↔ OAA	ABC + a ↔ ABCa
Lycine production	
$V16^1$:OAA + PYR → LYS + CO2	ABCD + abc → ABCDcb + a
$V16^2$:OAA + PYR → LYS + CO2	ABCD + abc → abcDCB + A
V17 :OAA + PYR → LYS + CO2	ABCD + abc → ABCDcb + a

The forward and reverse reactions were considered separately, resulting in 64 fluxes from 32 reactions. The equality constraints due to the measured extracellular fluxes, the metabolite balances, assumptions on directions of fluxes and scrambling reactions resulted in 10 degrees of freedom for the natural fluxes. We have chosen: $\{\overrightarrow{v_5}, \overrightarrow{v_{13}}, \overrightarrow{v_{17}}, \overleftarrow{v_8}, \overleftarrow{v_2}, \overleftarrow{v_4}, \overleftarrow{v_6}, \overleftarrow{v_7}, \overleftarrow{v_{12}}^2, \overleftarrow{v_{15}}\}$ as the free fluxes. The right and left arrows represents the forward and reverse reactions, respectively.

A set of 'true' measured positional enrichment data was generated using the network model in Fig. 1 and a set of free fluxes of $\{65.3, 1.2, 4.7, 11, 313.2, 14.6, 84.2, 5.7, 1.6, 30.4\}$. The only carbon source Glu was assumed to be labelled at the first carbon atom. The measured biomass fluxes and true fluxes were set such that their values are deemed representative of those found in the literature [12].

Two scenarios are considered to evaluate and compare the performance of the ECM algorithm using different linear least squares algorithms.

1) Noisy complete data: In this experiment we evaluate and compare the performance of the ECM algorithm using different LS algorithms over a wide range of noise environments. The labelling data were perturbed by uniformly distributed noise with variances of 5%, 10%, 15% and 20% of the true labelling data. we assume that the positional enrichment was available for every metabolite in the network.

In the CM1 step three linear least squares algorithms, i.e. the ordinary least squares (LS), the total least squares (TLS) and the constrained least squares (CLS) were tested. The solutions of LS and TLS were calculated by the pseudoinverse based on SVD and (9), respectively. If the LS or TLS resulted in a negative flux which is biologically infeasible, it would be replaced by a uniformly random number ranging between 0 and 0.001, provided the Q function was not decreased by the new flux value. If the Q function was decreased, another random number was re-selected. For the CLS algorithm, rather than randomly imputing the infeasible solutions by a positive value, the solution space was constrained to non-negative solutions. The CLS algorithm was handled by the Matlab function Lsqnonneg from the Optimisation Toolbox.

The algorithm was terminated if the thresholds (τ_1,τ_2) = 10^{-6} were reached or the maximum number of iteration of 30 was arrived. Each experiment was run through 50 Monte Carlo trials. The flux estimates are usually given in terms of net and exchange fluxes to enable a rapid interpretation of degree of reversibility by the users. The relations between the forward and backward fluxes and the net and exchanges fluxes are:

$$v_i^{net} = \overrightarrow{v}_i - \overleftarrow{v}_i \tag{32}$$

$$v_i^{xch} = \min(\overrightarrow{v}_i, \overleftarrow{v}_i) \tag{33}$$

To measure the closeness of the estimated fluxes to the true fluxes, the Normalised Mean-Square Error (NMSE) was adopted and defined by

$$NMSE = \frac{1}{l}\sum_{i=1}^{l}(\frac{\theta_i - \hat{\theta}_i}{\theta_i})^2 \tag{34}$$

where l is the number of identified fluxes, θ_i is the true value of flux i and $\hat{\theta}_i$ is the average estimated flux i in l simulations.

Fig. 2 shows that all the NMSE values of the ECM algorithm using LS and CLS were smaller than that of the ECM algorithm using TLS. For low noise level at 5%, the ECM using CLS produced results that were the closest to the true values. The results of Fig. 2 also show the performance difference between the ECM algorithm using LS and CLS decrease with increasing noise level.

2) Incomplete data: This experiment considers errors when labelling data are not fully observed. Generally, the labelling patterns of molecules in central metabolism are not usually measured directly, but rather they are measured from the corresponding amino acids. The precursors in Fig. 1 which usually can be derived from the amino acids are: Gap, Pyr, Co2, P5p, E4p, Oaa, Akg and Lys [3]. G6p and F6p can also be derived from the NMR spectra of glucan, glycogen, trehalose and chitin [10]. Therefore, we assumed that labelling data corresponding to these metabolites were available, resulting in 43 available labelling measurements. The proportion of missing data is around 30%.

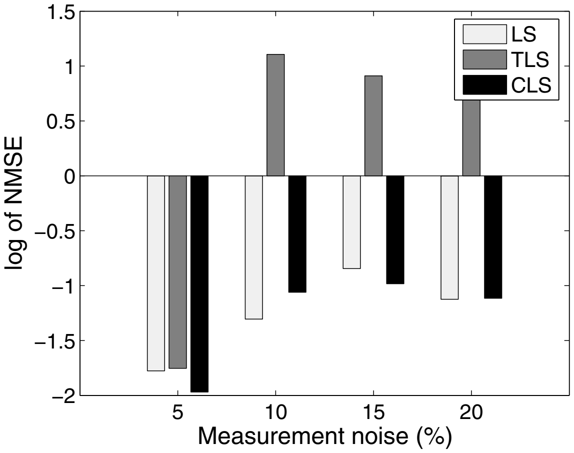


Fig. 2. NMSE performance comparison of the ECM algorithm using LS, TLS and CLS with measurement error (%) added to the true labelling data. The NMSE performance was calculated from the estimates of individual forward and backward fluxes.

The ECM algorithm combining with LS, TLS and CLS were applied using the generated measurements shown in Table III. By starting from 50 different initial values of the free fluxes, the estimate which gave the closest simulated enrichments to the measurements was selected as the best estimated flux distribution. The best estimates obtained from the tested algorithms are quantitatively compared in Table II. The estimates obtained for product formation and biomass effluxes are omitted from Table II for the sake of brevity. As can be seen in Table II, all three ECM algorithms provided good estimation results.

To compare the overall estimation performance among all tested algorithms, the NMSE values of the estimated natural fluxes (i.e. the individual forward and backward fluxes) were calculated as presented in Fig. 3 (a). In overall, CLS produced the smallest NMSE value of 0.0307, which is only slightly smaller than the resulting NMSE of LS (0.0340) but greatly is smaller than that of TLS (0.1190). The high value of NMSE obtained from the TLS method was caused by a poor estimation performance on the exchange fluxes as evident in Fig. 3 (b). From Fig. 3 (b), we also see that the best estimated net fluxes was achieved from the LS method. The inferior performance of the CLS method in net flux estimation is mainly due to its poor estimation on the net fluxes v_{13}^{net} and v_{14}^{net} which yielded approximately 50 % estimation error (see Table II).

The simulated labelling data from the ECM algorithm using CLS were calculated and are shown with their corresponding measured data in Table III. It can be seen that the estimated labelling data are in good agreement with the measurements.

Furthermore, we evaluate each algorithm performance in terms of convergence speed. In Fig. 4 we plot the time-

TABLE II

ESTIMATED VALUES OF THE NET AND EXCHANGE FLUXES WHEN LS, TLS AND CLS ARE USED IN THE CM1 STEP.

Flux	True net flux	LS	TLS	CLS
V1	100.00	100.00	100.00	100.00
V2	32.50	33.15	33.33	32.72
V3	71.01	71.22	71.28	71.08
V4	159.72	159.93	159.99	159.79
V5	65.30	64.66	64.47	65.09
V6	20.40	20.19	20.12	20.33
V7	20.40	20.19	20.13	20.33
V8	18.60	18.39	18.33	18.53
V9	62.33	62.54	62.61	62.40
V10	61.13	61.43	61.81	61.76
V11	52.93	53.23	53.61	53.56
V12	54.13	54.35	54.41	54.20
V13	1.20	1.12	0.80	0.64
V14	1.20	1.12	0.80	0.64
V15	43.69	44.40	45.10	44.67
V16	6.80	7.42	7.81	7.22
V17	4.70	3.46	2.68	3.86

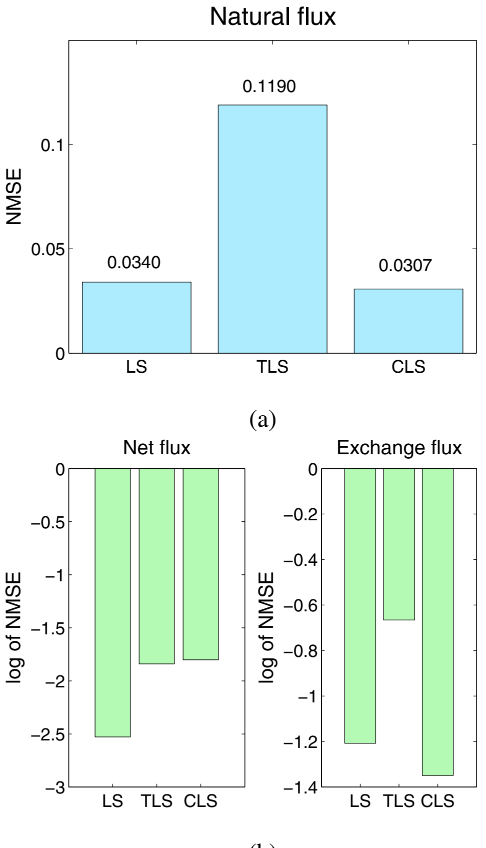
Flux	True exch. flux	LS	TLS	CLS
V2	313.20	315.06	317.66	314.55
V4	14.60	12.97	21.16	13.74
V6	84.20	80.33	76.53	82.79
V7	5.70	12.73	16.65	9.34
V8	11.00	6.76	4.77	8.72
V12	3.20	4.54	7.99	6.22
V15	30.40	30.85	32.81	30.83

dependent Q function of the three algorithms against the iteration number. In terms of number of iterations, CLS converged at iteration 11, followed by LS and TLS at iterations 12 and 13, respectively. The average calculation time per experiment of LS, TLS and CLS were 95, 143 and 96 seconds, respectively (data not shown). In comparison to the nonlinear least squares algorithm using the MATLAB function LSQNONLIN with the maximum iteration of 30 and the termination tolerances on the function value and parameters of 10^{-6}, the nonlinear least squares took around 600 seconds per simulation. Therefore, the ECM could lead to an over six fold speed up in the flux estimation.

V. CONCLUSION

In this study we propose an ECM algorithm to estimate metabolic fluxes from incomplete labelling data. The main focus of the report is to provide a general understanding of the algorithm and to show how this algorithm can be applied to metabolic flux estimation. The main advantage of the proposed method is that the ML estimation problem is simplified by the formulation of a linear-in-the-parameter model in the CM step, resulting in efficiency with the computational effort and speed-up in the estimation process.

For the network studied here, the results from computational experiments show that the ECM algorithms using LS and CLS perform well for both complete and incomplete data cases. In the incomplete data case, the ECM algorithm using CLS produced results superior to the combination of ECM and LS and TLS algorithms in terms of overall estimation accuracy.



(a)

(b)

Fig. 3. Comparison of NMSE performance of the ECM algorithm using LS, TLS and CLS when 30 % of labelling data are missing. (a) NMSE calculated from the estimated natural fluxes. (b) NMSE calculated from the estimated net and exchange fluxes.

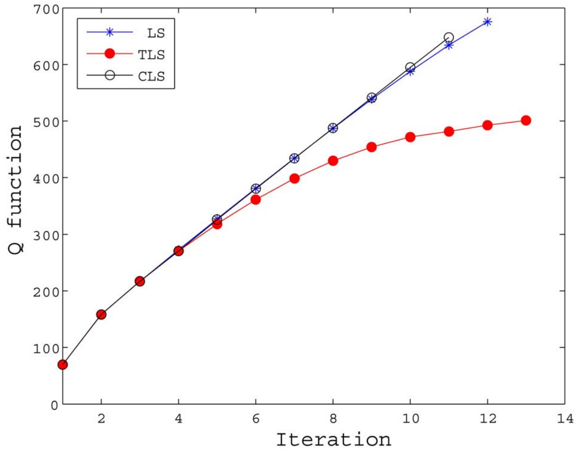


Fig. 4. Convergence of the ECM algorithms combining with LS, TLS and CLS.

TABLE III
MEASURED AND ESTIMATED POSITIONAL ENRICHMENTS RESULTING FORM THE BEST FIT DISTRIBUTION SHOWN IN TABLE II BY THE ECM ALGORITHM USING CLS.

Carbon atom	Measured(%)	Simulated (%)	Error
G6p-1	73.841	73.783	0.057
G6p-2	0.531	0.569	-0.038
G6p-3	2.015	2.156	-0.141
G6p-4	0.282	0.318	-0.036
G6p-5	0.097	0.107	-0.010
G6p-6	6.514	6.905	-0.392
F6p-1	65.489	65.449	0.040
F6p-2	0.701	0.750	-0.049
F6p-3	2.658	2.842	-0.183
F6p-4	0.372	0.419	-0.047
F6p-5	0.128	0.141	-0.013
F6p-6	8.593	9.101	-0.507
Gap-1	1.207	1.283	-0.076
Gap-2	0.416	0.433	-0.017
Gap-3	27.888	27.876	0.012
Pyr-1	2.034	2.131	-0.097
Pyr-2	1.752	1.801	-0.049
Pyr-3	27.130	27.104	0.026
Co2-1	21.973	21.929	0.044
P5p-1	12.371	12.304	0.067
P5p-2	1.779	1.904	-0.124
P5p-3	0.827	0.877	-0.050
P5p-4	0.285	0.296	-0.011
P5p-5	19.104	19.054	0.050
S7p-1	-	16.715	-
S7p-2	-	1.808	-
S7p-3	-	11.518	-
S7p-4	-	1.923	-
S7p-5	-	0.868	-
S7p-6	-	0.293	-
S7p-7	-	18.853	-
E4p-1	2.050	2.131	-0.082
E4p-2	0.687	0.766	-0.079
E4p-3	0.237	0.258	-0.022
E4p-4	15.870	16.637	-0.767
Oaa-1	6.772	6.904	-0.132
Oaa-2	9.411	9.502	-0.091
Oaa-3	22.782	22.756	0.026
Oaa-4	17.290	17.281	0.009
Ici-1	-	17.281	-
Ici-2	-	22.756	-
Ici-3	-	9.502	-
Ici-4	-	6.904	-
Ici-5	-	27.103	-
Ici-6	-	1.801	-
Akg-1	17.290	17.280	0.009
Akg-2	22.781	22.755	0.027
Akg-3	9.411	9.502	-0.091
Akg-4	27.129	27.102	0.026
Akg-5	1.752	1.801	-0.049
Fum-1	-	21.518	-
Fum-2	-	10.156	-
Fum-3	-	25.973	-
Fum-4	-	2.860	-
GlyOx-1	-	17.275	-
GlyOx-2	-	22.748	-
Lys-1	5.011	5.020	-0.009
Lys-2	6.565	6.462	0.102
Lys-3	24.396	24.470	-0.074
Lys-4	17.289	17.280	0.009
Lys-5	25.513	25.386	0.127
Lys-6	4.598	4.840	-0.242

Moreover, the ECM based approach is indeed resulting in a fast convergence.

Despite its overall excellent estimation performance and speed, it is evident that the algorithm produce an inferior performance to some fluxes. As a consequence, the method could probably be refined to further improve its performance and statistical analyses in terms of, for example the sensitivity to a particular measurement.

REFERENCES

[1] A. Bjorck., *Numerical methods for least squares problems*. Philadelphia, PA:Society for Industrial and Applied Mathematics, 1996.

[2] N. Isermann and W. Wiechert., "Metabolic isotopomer labeling systems. PartII: structural flux identifiability analysis," *Mathematical Bisciences*., vol. 183, pp. 175-214, 2003.

[3] A. Marx, A.A. de Graff, W. Wiechert, L. Eggeling, H. Sahm., "Determination of the fluxes in the central metabolism of Corynebacterium glutamicum by nuclear magnetic resonance spectroscopy combined with metabolite balancing," *Biotechnology and Bioengineering*., vol. 49, pp. 111-129, 1996.

[4] G.J. McLachlan and T. Krishnan., *The EM Algorithm and Extensions*. New York:John Wiley and Sons, 1997.

[5] X.L. Meng and D.B. Rubin., "Maximum likelihood estimation via the ECM algorithm: a general framework," *Biometrika*., vol. 80, pp. 267-278, 1993.

[6] K. Schmidt, M. Carlsen, J. Nielsen, and J. Villadsen., "Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices," *Biotechnology and Bioengineering*., vol. 55, No. 6, pp. 831-840, 1997.

[7] K. Schmidt, J. Nielsen, and J. Villadsen., "Quantitative analysis of metabolic fluxes in Escherichia coli, using two-dimensional NMR spectroscopy and complete isotopomer models," *Journal of biotechnology*., vol. 71, pp. 175-190, 1999.

[8] G.N. Stephanopoulos, A.A. Aristidou, and J.Nielsen., *Metabolic engineering - Principles and methodologies*." Academic Press, 1998.

[9] Van Huffel, S. and Vanderwalle, J., *The Total Least Squares Problem: Computational Aspects and Analysis*." Philadelphia:Society for Industrial and Applied Mathematics, 1991.

[10] van Winden, W., Verheijen, P. and Heijnen, S., " Possible Pitfalls of Flux Calculations Based on 13C-Labeling," *Metabolic Engineering*., vol. 3, pp. 151-162, 2001.

[11] W. Wiechert and A.A. de Graaf., "Bidirectional reaction steps in metabolic networks part I: Modeling and simulation of carbon isotope labeling experiments," *Biotechnology and Bioengineering*., vol. 55, pp. 101-117, 1997.

[12] W. Wiechert, C. Siefke, A.A. de Graaf, A. Marx., "Bidirectional reaction steps in metabolic networks part II: Flux estimation and statistical analysis," *Biotechnology and Bioengineering*., vol. 55, pp. 118-135, 1997.

[13] W. Wiechert., "13C Metabolic flux analysis," *Metabolic Engineering*., vol. 3, pp. 195-206, 2001.

[14] J. Yang, S. Wongsa, V. Kadirkamanathan, S.A.Billings and P.C. Wright., "Self adaptive evolutionary algorithm based methods for quantification in metabolic systems," *Computational Intelligence in Bioinformatics and Computational Biology*., San Diego USA, pp. 260-267, 2004.

[15] C. Zupke and G. Stephanopoulos., "Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices," *Genome Inform Ser Workshop Genome Inform*., vol. 10, pp. 489-498, 1994.