Reliable estimation of the key variables and of their rates of change in the alcoholic fermentation

IOAN CRISTIAN TRELEA (✉), ERIC LATRILLE, SOPHIE LANDAUD, GEORGES CORRIEU

Laboratoire de Génie et Microbiologie des Procédés Alimentaires

INA P-G, INRA, 78850 Thiverval-Grignon, France

E-mail: trelea@grignon.inra.fr
Tel: +33(0)130815490
Fax: +33(0)130815490

Abstract

The paper establishes a rigorous probabilistic framework for the reconciliation of apparently conflicting data coming from various physical and chemical measurements, related to the key biological variables of the alcoholic fermentation: the ethanol and the residual sugar concentrations. The analysis is carried out on a database consisting of 15 beer fermentation experiments, for which off-line determinations of ethanol concentration, fermentable sugar concentration, wort density and refractive index are available, as well as on-line records of evolved CO_2. The basic reconciliation method uses mass balance and monotonicity constraints derived from the biological knowledge of the fermentation process. In order to provide interpolated values and rate estimates, smoothness requirements are added. The reconciliation procedure gives more reliable estimates than any given measurement, detects outliers, helps fixing problems in the experimental setting and is also applicable on-line.

Keywords: data reconciliation, maximum likelihood, regularisation

Introduction

In alcoholic fermentation, reliable determination of the key biological variables (ethanol and fermentable sugar concentrations), as well as of their rates of change, is important for monitoring, scheduling, fault detection, control and fundamental study of the process. To be truly useful, such estimations should be performed frequently enough, e.g. every hour. In practice, the accuracy of the various existing measurement techniques is relatively limited, and the measurements themselves are rather infrequent, typically every 24 hours. Experimental data also contains inconsistencies, such as contradictory variations of stoichiometrically related quantities or monotonicity violations of thermodynamically irreversible reactions.

The paper develops a systematic approach for dealing with unavoidable measurement noise present in experimental data. The most probable values of conflicting measurements are determined by the maximization of a suitable likelihood function. It is shown that each measurement enters the likelihood function with a weight inversely proportional to its accuracy, which is intuitively reasonable. The most probable values must also satisfy linearity constraints, derived from the known mass balance of the fermentation process. Furthermore, the biochemical reactions involved in the fermentation process are known to be thermodynamically irreversible. This fact is incorporated in the design as a set of inequality constraints that most probable values must satisfy. The result of this basic reconciliation procedure, labeled in the text as the estimation method M_1, is a set of reliable values free from the mentioned inconsistency problems.

The problem of interpolation between existing data points, and the even more delicate problem of the rate estimation from infrequent and noisy measurements, are dealt with in two distinct ways. The first way is the direct extension of method M_1, augmented with an additional smoothness requirement for the estimation of intermediate values. The resulting methods are called M_2 and M_3, and differ with respect to the imposed terminal rate constraints. Conceptually, they are similar to M_1, but numerically difficult to handle because of the very large number of variables involved. The second way requires the choice of a family of smooth analytical functions which satisfy the mentioned stoichiometric and thermodynamic constraints. This is called method M_4. The main drawback of this method is that the selected family of functions does not necessarily capture all the details of the experimental data, such as biologically meaningful phases observed in the course of the fermentation.

It is demonstrated that all these techniques are also applicable on-line, when only part of the fermentation data is currently available. The computations involved require a basic PC computer and a standard numerical library. The usage of such a data reconciliation and robust rate estimation technique in real-time would be of great help in conducting industrial fermentations, based on at-line measurements only.

Materials and methods

Experimental

Data base

The experimental data base [1] consisted of a set of 15 lager beer fermentation runs, coded R01 to R15. The fermentation was carried out in 15 L stainless steel tanks (LSL Biolafitte, France), under "gentle" agitation at 100 rpm. The outlet gas was passed through a condenser at 0.5°C. The lager wort and the industrial yeast strain*, Saccharomyces cerevisiae uvarum*, were provided by the Institut Français de Brasserie et Malterie (IFBM, France). Starter cultures were carried out at 20°C in 5 L of wort during 3 days. Before inoculation, temperature was decreased to fermentation temperature (1 day) and the starter cultures were centrifuged three times (4000 rpm) in physiological saline. The 15 experimental runs differ with regard to fermentation temperature (10, 13 and 16°C), top pressure (50, 450 and 800 mbar), initial yeast concentration (5, 10, and 20 million cells per mL) and initial wort concentration (50, 70, 100 and 140 g/L of fermentable sugar).

Analytical methods

The ethanol concentration was determined using a Carlo Erba 5300 gas chromatograph equipped with a stainless steel column (200 mm, ∅0.3 mm) coated with Chromosorb 101 (SGE, USA). The concentration of fermentable sugar (the sum of the concentrations of fructose, glucose, maltose ant maltotriose) was determined using a High Performance Liquid Chromatography system (Waters, USA) with an Aminex HPX-87C column (300 mm, ∅7.8 mm, BioRad, USA) at 85°C. The density of the filtered and degasified wort was determined with a 10 mL pycnometer. The refractive index was measured with an ATAGO refractometer. The evolved CO_2 was recorded with a domestic gas meter, delivering a pulse for every liter of gas.

Data reconciliation

Mass balance in alcoholic fermentation

During alcoholic fermentation, the yeast grows and transforms the fermentable sugar and the amino acids into carbon dioxide, ethanol, glycerol, succinic acid and secondary metabolites, such as higher alcohols and esters which give the final product its characteristic flavor. An overall balanced equation of this process reads [2]:

$$C_6H_{11.2}O_{5.6}\ \text{(sugar)} + 0.0224\ C_4H_6O_2N\ \text{(amino acid)} \rightarrow 1.94\ CO_2\ \text{(carbon dioxide)} + 1.92\ C_2H_6O\ \text{(ethanol)} + 0.134\ CH_{1.67}O_{0.5}N_{0.17}\ \text{(biomass)} + 0.04\ C_3H_8O_3\ \text{(glycerol)} + 0.008\ C_4H_6O_4\ \text{(succinic acid)} - 0.389\ H_2O\ \text{(water)} \quad \text{Eq. (1)}$$

The original equation in [2] was corrected for the actual fermentable sugar composition in the considered wort: 20% glucose $C_6H_{12}O_6$ and 80% maltose $C_6H_{11}O_{5.5}$ (by weight). As far as the mass balance is concerned, it can be seen that the pathway leading from sugar to carbon dioxide and ethanol is dominant. The biomass growth come next. The consumption of amino acids and the production of other metabolites is so low that it can be safely neglected in an experimental mass balance. Constant stoichiometric coefficients mean that the consumption and production of the compounds in Equation (1) are related linearly.

Method M_1: Data reconciliation based on the mass balance

Assumptions

In wine and beer making, it is well known that the ethanol concentration, the amount of produced CO_2, the variations in wort density, in sugar concentration and in the refractive index are roughly proportional to each other [3]. This can be easily explained based on Equation (1): Ethanol, carbon dioxide and sugar enter this equation with constant stoichiometric coefficients. The variation of wort density during fermentation is mainly due to mass loss because of the carbon dioxide evolution. The refractive index of the wort is a measure of the sugar concentration.

The measured variables are affected by noise and do not satisfy the linear relationships exactly. The following standard assumptions about the measurement noise were made:

A_1 All measurements are affected by white (independent), normally distributed, zero mean noise.

A_2 The noise variance is specific to each measured variable (ethanol concentration, fermentable sugar concentration, wort density, refractive index). For each variable, the noise variance is constant for all runs and for all samples in a run.

A_3 The noise variance for each type of measurement (the accuracy of the measurement technique) is known from separate experiments.

These assumptions were verified *a posteriori*, based on the experimental data.

Constraints

The data reconciliation procedure consisted in finding the most probable values of the measured variables, based on the available measurements, and compatible with the consistency constraints derived from the fermentation process knowledge. The constraints considered in this work are:

C_1 In each experimental run, the variations of the ethanol concentration, fermentable sugar concentration, wort density and refractive index are proportional to each other.

C_2 The proportionality coefficients in constraint C_1 (the yield coefficients) are the same for all runs. This assumption is reasonable because the wort volume was shown to remain constant during the fermentation process (Appendix), and hence the density and the refractive index variations are proportional to the mass variations of the ethanol and of the fermentable sugar. In the range of operating conditions specified above, the stoichiometric coefficients in Equation (1) (mass balance) do not depend on temperature, top pressure, initial yeast concentration and initial sugar concentration [4], which varied from one run to another.

C_3 In each experimental run, the initial ethanol concentration is zero.

C_4 In each experimental run, the final fermentable sugar concentration is zero.

C_5 In each experimental run, the ethanol concentration is monotonically increasing, while wort density, sugar concentration and refractive index are monotonically decreasing. This condition, fully confirmed experimentally, comes from the fact that, under anaerobic conditions, the ethanol can not be utilized further, and the reaction described by Equation (1) can not be reversed, due to thermodynamic restrictions.

Mathematical formulation

Mathematically, the problem can be stated as follows. Let n be the number of experimental runs, i (from 1 to n) the index of the current run, m_i the number of samples taken during the run i, j (from 1 to m_i) the index of the current sample, E_{ij}, D_{ij}, S_{ij} and R_{ij} the measured values of the ethanol concentration, wort density, sugar concentration and refractive index respectively, e_{ij}, d_{ij}, s_{ij} and r_{ij} the most probable values of the same variables. Based on assumptions A_1-A_3, the most probable values are those which maximize the associated likelihood function, i.e. the probability of observing the given measured values. It is more convenient from a numerical point of view, but mathematically equivalent, to minimize the minus logarithm of the likelihood function, which takes the form [5]:

$$L_1 = m\log\left(4\pi^2\sigma_E\sigma_D\sigma_S\sigma_R\right) + \frac{1}{2\sigma_E^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(e_{ij}-E_{ij}\right)^2 + \frac{1}{2\sigma_D^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(d_{ij}-D_{ij}\right)^2 + \frac{1}{2\sigma_S^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(s_{ij}-S_{ij}\right)^2 + \frac{1}{2\sigma_R^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ij}-R_{ij}\right)^2 \qquad \text{Eq. (2)}$$

where $m = \sum_{i=1}^{n} m_i$ is the total number of measurements, σ_E, σ_D, σ_S, and σ_R are the known standard deviations of the measurement process (assumption A_2). The first term in Equation (2) is constant, and can be neglected for minimization purposes.

The proportionality requirements C_1 and C_2 were introduced by expressing three of the measured variables in terms of the fourth one, using proportionality (yield) coefficients common to all runs:

$$d_{ij} = d_{i1} - Y_D e_{ij} \qquad \text{Eq. (3)}$$

$$s_{ij} = s_{i1} - Y_S e_{ij} \qquad \text{Eq. (4)}$$

$$r_{ij} = r_{i1} - Y_R e_{ij}, \qquad i = 1...n, \quad j = 2...m_i \qquad \text{Eq. (5)}$$

The constraints C_3 and C_4 are straightforward:

$$e_{i1} = 0 \qquad \text{Eq. (6)}$$

$$s_{i m_i} = 0, \qquad i = 1...n \qquad \text{Eq. (7)}$$

Taking into account constraints C_1 and C_2, the condition C_5 reduces to:

$$e_{ij} - e_{i\,j+1} \le 0, \qquad i = 1...n, \quad j = 1...m_i - 1 \qquad \text{Eq. (8)}$$

Numerical resolution

As stated, the data reconciliation problem requires the minimization of a quadratic function with a mixture of linear, nonlinear, equality and inequality constraints. The nonlinearity comes from the product terms *Ye* in Equations (3-5). Furthermore, the dimension of the optimization problem, i.e. the number of unknowns, equal here to $4m + 3$, is quite large, as indicated in Table 1. The numerical resolution can be improved dramatically by breaking the original problem into:

- one nonlinear unconstrained minimization with 3 unknown yields, Y_D, Y_S, and Y_R.
- n quadratic independent subproblems (for fixed yields) of dimension $4m_i$, with linear equality and inequality constraints. The dimension of the quadratic subproblems can be reduced further to $m_i + 1$ by introducing the equality constraints (Eq. 3-7) into the cost function (Eq. 2), which is straightforward.

This decomposition is particularly useful because effective algorithms for optimizing quadratic functions with linear constraints exist [6]. The top-level nonlinear optimization was solved with a Levenberg-Marquardt algorithm for nonlinear least squares [6].

Missing data and outliers

The fermentation data was accumulated over years, and, for some samples, not all four measurements could be performed, mainly due to equipment failures. Discarding those samples completely would result in unnecessary information loss. Samples with at least two available measurements were retained. The missing data was handled by deleting the appropriate terms in the likelihood function.

After the data reconciliation process was performed, it turned out that some measurements were clearly unrealistic. Measurements which disagreed with the estimated values by more than 3 standard deviations were considered outliers and systematically discarded, as if they were missing data.

A special case of missing data appeared in the runs R10 and R13, which, for technical reasons, were stopped before sugar exhaustion. The constraint C_4 was eliminated in these cases.

Method M_2: Interpolation using a free-form model

Motivation

The data reconciliation method M_1 gives reliable estimations of the measured variables at the same sampling moments at which measurements were taken. In our case, the sampling interval was not constant, and roughly equal to 24 hours. However, in many applications, such as process monitoring and control, it is important to have much more frequent estimates of the key variables, e.g. every hour, as well as rate estimates. The data reconciliation method M_1 was extended to provide intermediate values, giving method M_2. The main difficulty comes from the fact that "target" measured values are not available for the intermediate moments. Intermediate values have to be determined from an additional smoothness or *regularisation* condition. The set of all unknown values, including intermediate ones, plus the regularisation condition, form the free-form model.

Various regularisation functions have been proposed in the literature, such as maximum entropy, minimum average slope, minimum average curvature etc. [7]. The adequacy of a particular regularisation function depends on the application. In this work, the minimum average curvature criterion was selected, which is also used, for example, in interpolation by spline polynomials.

Mathematical formulation

Let $\tau = 1$ h be the time interval at which interpolated values are desired. For each experimental run i there are p_i such values, and generally $p_i >> m_i$. Let k (from 1 to p_i) be the index of the interpolated value and k_j, with j from 1 to m_i, those values of k for which at least two measurements are available. The Equation (2) was modified by introducing the new regularization term and dropping out the constant term:

$$L_2 = \frac{1}{2\sigma_G^2}\sum_{i=1}^{n}\sum_{k=2}^{p_i-1}\left(\frac{e_{k-1}-2e_k+e_{k+1}}{\tau^2}\right)^2 + \frac{1}{2\sigma_E^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(e_{ik_j}-E_{ij}\right)^2 + \frac{1}{2\sigma_D^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(d_{ik_j}-D_{ij}\right)^2 + \frac{1}{2\sigma_S^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(s_{ik_j}-S_{ij}\right)^2 + \frac{1}{2\sigma_R^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ik_j}-R_{ij}\right)^2 \quad \text{Eq. (9)}$$

The coefficient $\sigma_G = 0.02$ $gL^{-1}h^2$ is the weight of the regularization term in the likelihood function. It expresses the tradeoff between the smoothness requirement and the fit to experimental data. Clear theoretical guidelines for selecting its value are not available, and the

choice is subjective to some extent, as well as the choice of the regularization function itself. It was found, however, that the results are relatively insensitive to reasonable modifications of σ_G, such as dividing or multiplying the given value by a factor of 2.

The constraints C_1-C_5 are expressed similarly to method M_1:

$$d_{ik} = d_{i1} - Y_D e_{ik} \qquad \text{Eq. (10)}$$

$$s_{ik} = s_{i1} - Y_S e_{ik} \qquad \text{Eq. (11)}$$

$$r_{ik} = r_{i1} - Y_R e_{ik}, \qquad i = 1...n, \;\; k = 2...p_i \qquad \text{Eq. (12)}$$

$$e_{i1} = 0 \qquad \text{Eq. (13)}$$

$$s_{i\,p_i} = 0, \qquad i = 1...n \qquad \text{Eq. (14)}$$

$$e_{ik} - e_{i\,k+1} \le 0, \qquad i = 1...n, \;\; k = 1...p_i - 1 \qquad \text{Eq. (15)}$$

Missing values and outliers were handled in the same way as in the method M_1. For the numerical resolution, the same decomposition in a nonlinear unconstrained optimisation with quadratic linearly constrained subproblems was used.

Method M_3: Interpolation using a free-form model with initial and final rate constraints

In the considered fermentation experiments, it is reasonable to assume that the initial fermentation rate is close to zero, since, after inoculation, the yeast needs at least a few tens of minutes to adapt to the new medium. The final fermentation rate is also zero, because the experiments were conducted until no CO_2 evolution was observed. The data reconciliation method M_3 is the same as method M_2, except that two additional constraints were added:

C_6 The initial ethanol production rate is zero in all experimental runs.

C_7 The final ethanol production rate is zero in all experimental runs.

Mathematically, this is expressed as:

$$e_{i1} - e_{i2} = 0, \qquad i = 1...n \qquad \text{Eq. (16)}$$

$$e_{i\,p_i-1} - e_{i\,p_i} = 0, \qquad i = 1...n \qquad \text{Eq. (17)}$$

Method M_4: Interpolation using a parametric model

An alternative to interpolation by free-form models and regularisation functions is the use of parametric models. A parametric model is a family of functions depending on one or a few parameters which determine the actual curve shape. The main difficulty is the choice of the mathematical form of the functions, which should be compatible with the constraints, while still preserving enough flexibility to accommodate the experimental data. For the present application, the so-called incomplete beta functions were found useful [5]:

$$\beta(q,\lambda,\mu)=\frac{\int_0^q x^{\lambda-1}(1-x)^{\mu-1}dx}{\int_0^1 x^{\lambda-1}(1-x)^{\mu-1}dx}, \qquad q\in[0 \quad 1] \qquad \text{Eq. (18)}$$

In this formula, $\lambda > 0$ and $\mu > 0$ are shape parameters. Their effect is illustrated in Figure 1. In order to accommodate the experimental data, two scale parameters were added. Let t_{im_i} be the (known) time when the last sample was taken in experiment *i*, and e_{im_i} be the (unknown) final ethanol concentration in that experiment. The ethanol concentration at any moment *t* between 0 and t_{im_i} was expressed as:

$$e_i(t)=e_{im_i}\beta\left(\frac{t}{t_{im_i}},\lambda_i,\mu_i\right) \qquad \text{Eq. (19)}$$

The constraints C_1 and C_2 are expressed as before:

$$d_i(t)=d_{i1}-Y_D e_i(t) \qquad \text{Eq. (20)}$$

$$s_i(t)=s_{i1}-Y_S e_i(t) \qquad \text{Eq. (21)}$$

$$r_i(t)=r_{i1}-Y_R e_i(t), \qquad i=1...n, \;\; t\in\lfloor 0 \quad t_{im_i}\rfloor \qquad \text{Eq. (22)}$$

The constraint C_3 is always satisfied, since $\beta(0,\lambda,\mu)=0$ for all $\lambda > 0$ and $\mu > 0$. The constraint C_4 is equivalent to:

$$s_{i1}=Y_S e_{im_i} \qquad \text{Eq. (23)}$$

The constraint C_5 is also satisfied, since the beta functions are monotonically increasing. The theory of the beta functions also implies that the rate constraints C_6 and C_7 are equivalent to:

$$\lambda > 1, \quad \mu > 1 \qquad \text{Eq. (24)}$$

The negative logarithm of the likelihood function, after omission of the constant term, is:

$$L_4 = \frac{1}{2\sigma_E^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(e_i(t_{ij}) - E_{ij}\right)^2 + \frac{1}{2\sigma_D^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(d_i(t_{ij}) - D_{ij}\right)^2 + \frac{1}{2\sigma_S^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(s_i(t_{ij}) - S_{ij}\right)^2 + \frac{1}{2\sigma_R^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_i(t_{ij}) - R_{ij}\right)^2 \qquad \text{Eq. (25)}$$

In the case of the parametric model, the number of unknowns is $5n + 3$, that is considerably lower than in the previous cases (Table 1). However, breaking the original problem into a top-level optimization of the three yields and n independent calculations of five parameters for each experimental run, still saves some computation time and improves accuracy, even if the subproblems require nonlinear optimization.

Results

Data reconciliation at work

An example of the data reconciliation procedure, using method M_1, is shown in Figure 2. This particular experiment was selected because it illustrates, on a single run, most of the benefits expected from data reconciliation.

Firstly, an incompatibility is detected between the total ethanol production of 26.2 g/L and the total sugar consumption of 66.3 g/L. The yield coefficient $Y_S = 1.833$ is determined reliably, based on the whole pool of 15 experiments, as discussed below, so the incompatibility must come from either ethanol concentration or sugar concentration measurement errors, or both. The density and refractive index measurements help resolving the conflict, by indicating that the truth is probably somewhere in between: the most probable ethanol concentrations are higher than the measured ones, and sugar concentrations lower.

Secondly, the reconciliation procedure indicates that the last-but-one ethanol measurement is an outlier, and should not be included in calculations. If it was included, (i) the estimated density, sugar *and* refractive index at 116 h would be significantly higher than measured *and* (ii) at 93 h the ethanol concentration would be lower and the other values higher than measured, because of the constraint C_5. This is less probable than excluding a single ethanol measurement. Visual inspection of the curves suggests the same thing. The first refractive

index measurement is also an outlier, because it goes completely against the evidence provided by the other three variables.

Thirdly, it can be seen that the monotonicity constraint C_5 is active (limiting) between the last two samples, at 93 h and 116 h. Without this constraint, the last estimated value would be lower for ethanol and higher for density, sugar and refractive index. This is in contradiction with the biological reality: sugar concentration can not increase, and ethanol concentration can non decrease. The reconciliation process finds a compromise that assumes small measurement errors while satisfying the biological constraint.

Yield coefficient determination

The effectiveness of the reconciliation process heavily relies on the assumption that the yield coefficients are the same for all runs, as imposed by the constraint C_2. The estimated values of the yield coefficients are reported in Table 2, together with their 95% confidence limits and the theoretical values, estimated from the mass balance equation. It can be noted that all four reconciliation methods provide consistent estimations, since the confidence intervals overlap. The theoretical value of the yield coefficient of the density versus ethanol (Y_D) was estimated from the mass balance (Equation 1) with the assumption that the volume of the fermentation medium remains constant (Appendix), and the density variation is due to mass variation caused by CO_2 release. So the value reported in Table 2 is actually the theoretical yield of the carbon dioxide versus ethanol. This theoretical value agrees with the experimental estimations provided by methods M_1 and M_4. On the contrary, the theoretical value of the fermentable sugar versus ethanol yield (Y_S) is far from agreeing with the experimental ones. The difficulty of measuring the concentration of the fermentable sugar in the wort is well known by the brewers, who always prefer using the wort density instead. In the present case, the systematic underestimation of the fermentable sugar concentration might come from the difficulty to distinguish between the fermentable sugar maltotriose and non-fermentable sugars with higher number of glucose units. With the High Performance Liquid Chromatograph used, the corresponding peaks overlapped significantly.

Validity of the statistical assumptions

The data reconciliation methods were worked out based on standard statistical assumptions about the measurement errors, A_1-A_3. The computation of the confidence intervals for the estimated quantities is also based on these assumptions. As an example, the probability plot of

the scaled residuals (difference between the most probable and the experimental values, divided by the assumed standard deviation of the measurement noise) is shown in Figure 3, for the reconciliation method M_2. The plot indicates that the hypothesis of a zero mean, normally distributed measurement noise looks reasonable. The probability plots for the other methods are very similar. The assumed (*a priori*) and the estimated (*a posteriori*) standard deviations for the considered measurements (ethanol concentration, density, sugar concentration and refractive index) are reported in Table 3. It can be seen that the assumed standard deviations, obtained from previous repetition experiments, were slightly overestimated. However, for any given method, the standard deviations were overestimated by roughly the same factor for all measurement types. This does not hamper the conclusions, since multiplying any of the L_1-L_4 expressions by a constant factor does not change the optimization results. This claim is also supported by the probability plot in Figure 3, where the scaled residuals from all measurement types were mixed, but there is no evidence of data coming from distinct probability distributions.

Interpolation and rate estimation

In many applications it is desired to have frequent (e.g. every hour) estimates of the variables that are measured only rarely (e.g. every day), as well as estimates of the corresponding rates. The results of applying methods M_2, M_3 and M_4 for interpolation and rate estimation are illustrated in Figure 4, on the experimental run R07. All three methods perform a smooth interpolation between existing measurements. The main differences arise at the beginning of the fermentation, when the experimental data is scarce. Method M_2 minimizes the average curvature, and hence favors a constant production rate, unless the experimental evidence suggests otherwise. The resulting initial ethanol production rate is very high. The constraint C_6 embedded in methods M_3 (explicitly) and M_4 (implicitly) forces a zero initial production rate. The results produced by methods M_3 and M_4 are similar, with however a 10 % discrepancy for the maximum production rate.

A different situation is depicted in Figure 5, for the experimental run R03. Here experimental data is available at the beginning of the curve, and all three methods produce similar results before 100 h. Differences in the estimated production rate arise between 100 and 200 h. Methods M_2 and M_3 make no *a priori* assumption about the curve shape (hence the name of free-form interpolation) and detect two distinct phases, before and after 170 h, which might have biological significance. This is opposed to method M_4, whose curve shape is imposed by

the selected parametric model, namely the incomplete beta function. Method M_4 can not detect such fine structure in the data.
As a general rule, the estimated rate is much more sensitive to the interpolation method than the variable itself. In applications where rate estimation is important, the interpolation method and the underlying hypotheses must be considered very carefully.

On-line data processing

For process supervision and control purposes, experimental data must be processed on-line, as they arrive. Yield coefficients can not be estimated on-line, because available data is too limited. Rather, fixed values previously estimated on the whole data base should be used (Table 2). In Figure 6, on-line processing, using 2/3 of the data available in experiment R13, is compared with off-line processing, on the whole data set. As far as the ethanol concentration is concerned, all four methods produce similar results, even when applied to limited experimental data. This is also true for density, sugar concentration and refractive index, because they are linearly related to the ethanol concentration. The rate estimations are also good for methods M_2 and M_3, but not for method M_4 (method M_1 does not provide rate estimations). The mathematical form of the equations used in method M_4 was selected for its ability to describe sigmoid shapes, which correspond to finished fermentations, and appears to be less appropriate for running experiments.
Tests performed on the entire database, for various fractions of the available data, suggest that the most robust method for on-line processing is M_1. If rate estimations are needed, M_3 should be preferred.

Detection of CO_2 leak

The measurement of CO_2 evolution is a convenient way for monitoring alcoholic fermentation using inexpensive on-line sensors [8,9,10]. According to Equation (1), the *produced* CO_2 is linearly related to the ethanol concentration. After the wort is saturated in CO_2 and the top pressure in the tank is established, the *released* CO_2 (which is actually measured) equals the produced one. This was verified for the considered database by computing, by ordinary linear regression, the experimental yield of the CO_2 versus ethanol, and the associated confidence limits. The results are presented in Figure 7, together with the theoretical yield obtained from the Equation (1). The agreement is good for most experiments, as the confidence interval includes the theoretical value. In run R11 the CO_2 was not recorded. In run R06 the CO_2

measurement was affected by a known equipment failure. However, runs R04, R12 and R13 appeared to be affected by a subtle CO_2 leak, which was not at all obvious before reliable ethanol concentrations were determined by data reconciliation (here, method M_1 was used). CO_2 leak can also be detected in real time, as soon as enough measurements become available to make the confidence interval of the computed yield small enough.

Accuracy improvement

For all considered variables, data reconciliation provides more reliable estimates than any individual measurement. In order to illustrate the accuracy improvement, the standard errors associated with the estimates of the final ethanol concentration, the initial density, the initial sugar concentration and the initial refractive index are reported in Table 4. The accuracy improvement is given in the 5^{th} column. This is to be compared with an accuracy improvement by a factor of 2 which would result if 4 repetitions of a single measurement were performed. The equivalent number of repetitions of a single measurement is reported in the last column. For example, in order to get a similar accuracy for the initial sugar concentration, 10 measurements should be performed instead of 4 (since the standard error of a mean value is proportional to the inverse of the square root of the number of repetitions). Except for the ethanol, data reconciliation based on 4 measurements provides more reliable estimates than simple repetitions of any given measurement 4 times. This is due to the biological insight introduced into the problem formulation via the consistency constraints $C_1 - C_7$.

Summary

Direct measurement of key biological variables in alcoholic fermentation (fermentable sugar and ethanol concentration in the medium) is often impractical, costly and unreliable. However, related measurements are more readily available and are more accurate (density, refractive index, evolved CO_2). The paper illustrates how the biological insight (through mass balance, monotonicity constraints and smoothness requirements) can help improve the accuracy of either measurement, supply any missing one, provide interpolated values and rate estimates.

Apparently conflicting data are reconciled in a rigorous probabilistic framework, which also helps identifying outliers. Comparison of experimental yield coefficients with theoretical ones points out problems with the experimental setting, such as CO_2 leak and separation of

fermentable and non fermentable sugars by HPLC. Accuracy improvement resulting from "intelligent" data reconciliation is shown to be higher than from simple repetitions of the measurements.

If interpolated values and rates of change of the variables are required, then either explicit smoothness requirements are added, or a family of smooth analytical functions is used to model the data. Analytical functions require much less parameters than free-form models, but can not capture all specific features of the data. All the benefits of the basic data reconciliation are retained in either case.

Data reconciliation, interpolation and rate estimation can be also applied on-line, when only part of the data is currently available. The presented techniques are basically unchanged, but yield coefficients have to be known beforehand and fixed. More care is needed when selecting a technique for rate estimation.

The techniques presented in the paper have been tailored for the existing database, but the same principles can be easily adapted to other applications. For example, data reconciliation could be carried out if only density plus CO_2 or refractive index measurements were performed. Density measurements are already commonplace in industry. Refractive index or carbon dioxide measurements could be added with little extra cost [11]. Ethanol and fermentable sugar concentrations, as well as their rates of change, would be simply estimated from the stated linear relationships. Of course, the accuracy diminishes if the amount of performed measurements decreases.

Nomenclature

Symbol	Units	Significance
A_1-A_3		Statistical assumptions about the measurement noise
C	g L^{-1}	Amount of evolved carbon dioxide
C_1-C_7		Consistency constraints satisfied by the estimated variables
D_{ij}	g L^{-1}	Measured wort density in sample *j* of run *i*
d_{ij}	g L^{-1}	Most probable wort density in sample *j* of run *i*
E_{ij}	g L^{-1}	Measured ethanol concentration in sample *j* of run *i*
e_{ij}	g L^{-1}	Most probable ethanol concentration in sample *j* of run *i*
i		Index of the experimental run
j		Index of the sample in an experimental run
k		Index of the interpolated value
k_j		Index of the interpolated value for which measurement *j* is available
L_1-L_4		Negative logarithms of the likelihood functions associated to methods M_1-M_4

M_1		Basic data reconciliation method based on mass balance and reaction thermodynamics
M_2		Data reconciliation and free-form interpolation method. In addition to M_1, includes smoothness requirements
M_3		Data reconciliation and free-form interpolation method. In addition to M_2, includes initial and final rate constraints
M_4		Data reconciliation and parametric interpolation method. Based on incomplete beta functions
m		Total number of available samples in the data base
m_i		Total number of samples available for the run i
n		Total number of experimental runs in the data base
p		Total number of interpolated values in the data base
p_i		Total number of interpolated values in run i
q		Dimensionless time
R01-R15		Experimental runs
R_{ij}		Measured refractive index in sample j of run i
r_{ij}		Most probable refractive index in sample j of run i
S_{ij}	g L^{-1}	Measured fermentable sugar concentration in sample j of run i
s_{ij}	g L^{-1}	Most probable fermentable sugar concentration in sample j of run i
t	h	Time
t_{ij}	h	Time at which sample j of run i was taken
Y_C	g g^{-1}	Yield coefficient of the carbon dioxide versus ethanol
Y_D	g g^{-1}	Yield coefficient of the wort density versus ethanol
Y_R	g^{-1} L	Yield coefficient of the refractive index versus ethanol
Y_S	g g^{-1}	Yield coefficient of the fermentable sugar versus ethanol
β		Incomplete beta function
λ		First shape parameter of the beta function
μ		Second shape parameter of the beta function
σ_D	g L^{-1}	Standard measurement error for wort density
σ_E	g L^{-1}	Standard measurement error for ethanol concentration
σ_G	g L^{-1} h^2	Weight of the smoothing term in the likelihood function
σ_R		Standard measurement error for refractive index
σ_S	g L^{-1}	Standard measurement error for fermentable sugar concentration
τ	h	Time interval for estimation of interpolated values

Appendix: Experimental verification of the constant volume hypothesis

In the considered experimental setting, an accurate determination of the wort volume variation was not feasible. Instead, it was verified in a separate experiment that the mass deficit due to CO_2 evolution, predicted by Equation (1), explained the density variation almost exactly. The estimation of the volume modification induced by yeast growth, given below, shows a negligible contribution.

The density D_1 of a 80 g/L aqueous solution of fermentable sugar (64 g/L maltose + 16 g/L glucose) was found to be:

$D_1 = 1030.2$ g/L

According to Equation (1), 80 g of fermentable sugar give 40.9 g of ethanol, which remains in the solution, and $m_C = 39.5$ g of carbon dioxide, which is released. The fermeter used in the experiments was equipped with a condenser at 0.5°C, which ensured that neither water nor ethanol were released along with carbon dioxide. The measured density D_2 of a 40.9 g/L aqueous ethanol solution was:

$D_2 = 991.7$ g/L

Thus, a volume $V_1 = 1.000$ L of sugar solution weights:

$m_1 = D_1 \cdot V_1 = 1030.2$ g

After fermentation and CO_2 release, the remaining mass is:

$m_2 = m_1 - m_C = 990.7$ g

and it occupies a volume of:

$V_2 = m_2 / D_2 = 0.999$ L.

Hence, within 0.1% accuracy, the production of the two major compounds of the alcoholic fermentation (carbon dioxide and ethanol) does not change the fermentation volume.

The biomass growth does not change the mass balance, since all compounds are taken from the wort and remain in the wort. In a typical fermentation experiment, $60 \cdot 10^9$ yeast cells are produced in 1 L of wort, representing a volume less than 0.007 L (for a typical cell diameter of 6 μm). A living cell contains at least 95% of water, taken from the original solution. So, possible volume variations due to biomass growth are less than 0.00035 L/L, which is far below the overall experimental accuracy.

Other metabolites are produced in too small concentrations to change the wort volume by more than 0.1%.

References

1. Titica, M., Landaud, S., Trelea, I.C., Latrille, E., Corrieu, G., Cheruy, A.: Modelling of higher alcohol and ester production kinetics based on CO_2 emission, with a view to beer flavor control by temperature and top pressure. Journal of the American Society of Brewing Chemists 58 (2000), 167-174.
2. Williams, L.A.: Heat release in alcoholic fermentation: a critical reappraisal. Am. J. Enol. Vitic., 33 (1982) 149-153.
3. El Haloui, N., Picque, D., Corrieu, G.: Alcoholic fermentation in winemaking: on-line measurement of density and carbon dioxide evolution. Journal of Food Engineering, 8 (1988), 17-30.
4. Gee, D.A., Ramirez, W.F.: A flavour model for beer fermentation. Journal of the Institute of Brewing, 100 (1994), 321-329.
5. Bury, K.: Statistical distributions in engineering. pp 27-48 & 238-266. Cambridge University Press, 1999.
6. Coleman, T., Branch, M.A., Grace, A.: Optimization toolbox for Matlab – User's guide, pp. 2.4-2.26 The MathWorks, Inc, 1999.
7. Sivia, D.S.: Data analysis: a bayesian tutorial, pp. 130-158. Oxford University Press, 1998.
8. Stassi, P., Rice, J.F., Munroe, J.H., Chicoye, E.: Use of the CO_2 evolution rate for the study and control of the fermentation. MBAA Technical Quarterly, 24 (1987), 44-50.
9. Daoud, S.I., Searle, B.A.: On-line monitoring of brewery fermentation by measurement of CO_2 evolution rate. Journal of the Institute of Brewing, 96 (1990), 297-302.
10. Corrieu, G., Trelea, I.C.,Perret, B.: On-line estimation and prediction of density and ethanol evolution in the brewery. MBAA Technical Quarterly 37 (2000), 173-181.
11. Stassi, P., Fehring, J.F., Bale, C.B., Goetzke, G.P., Ryder, D. S.: Optimization of fermenter operations using a fermenter instrumentation sytem. MBAA Technical Quarterly 32 (1995), 57-65.

List of figures

Figure 1. Representation of incomplete beta functions for various combinations of the shape parameters λ and μ.

Figure 2. Data reconciliation using method M_1, applied to experimental run R02. Measurements (o), most probable values (*) and outliers (x).

Figure 3. Probability plot of the 701 scaled residuals for the data reconciliation method M_2. The probability scale was linearized based on the assumption of normally distributed measurement noise, with zero mean and standard deviation indicated in Table 3. The absence of significant departure from the straight line indicates that the assumption is reasonable.

Figure 4. Data reconciliation, smooth interpolation and rate estimation, applied to experimental run R07. Method M_2 (—) uses a free-form model and minimises the average curvature. Method M_3 (– –) is similar to M_2 but additionally imposes zero initial and final production rates. Method M_4 (▬) uses a parametric model based on incomplete beta functions.

Figure 5. Data reconciliation, smooth interpolation and rate estimation, applied to experimental run R03. Methods M_2 (—) and M_3 (– –) detect fine structure in the data, which method M_4 (▬) can not account for.

Figure 6. On-line data processing, applied to experimental run R13. Experimental data (o), off-line estimation using all available data (—), on-line estimation using only data available up to the current moment (▬).

Figure 7. Evolved CO_2 yield versus ethanol (method M_1), computed by ordinary linear regression. Measurements between 10% and 90% of the final ethanol concentration were taken into account, in order to avoid difficulties with the dissolved CO_2 and tank pressurisation. Experimental yields (o), 95% confidence limits (|) and theoretical yield (—).

Table 1
Number of unknown values in the reconciliation procedures

	Method M_1		Method M_2		Method M_3		Method M_4	
Basic problem formulation	e_{ij}, d_{ij}, s_{ij}, r_{ij}	$4m$	e_{ik}, d_{ik}, s_{ik}, r_{ik}	$4p$	e_{ik}, d_{ik}, s_{ik}, r_{ik}	$4p$		
	Y_D, Y_S, Y_R	3	Y_D, Y_S, Y_R	3	Y_D, Y_S, Y_R	3		
Total:	$4m+3$ =	**783***	$4p+3$ =	**11403**	$4p+3$ =	**11403**		
After taking into account equality constraints	e_{ij}	$m-2n$	e_{ik}	$p-2n$	e_{ik}	$p-4n$	λ_i, μ_i	$2n$
	d_{i1}, s_{i1}, r_{i1}	$3n$	d_{i1}, s_{i1}, r_{i1}	$3n$	d_{i1}, s_{i1}, r_{i1}	$3n$	d_{i1}, s_{i1}, r_{i1}	$3n$
	Y_D, Y_S, Y_R	3	Y_D, Y_S, Y_R	3	Y_D, Y_S, Y_R	3	Y_D, Y_S, Y_R	3
Total:	$m+n+3$ =	**213***	$p+n+3$ =	**2865**	$p-n+3$ =	**2838**	$5n+3$ =	**78**

*Missing data not taken into account

Table 2
Yield coefficients: theoretical values and estimated values from experimental data

Yield		Mass balance	Method M_1	Method M_2	Method M_3	Method M_4
Y_D	Min*	0.966	0.959	0.983	0.969	0.958
	Most probable		0.973	0.991	0.978	0.973
	Max*		0.987	0.999	0.987	0.989
Y_S	Min*	1.957	1.791	1.828	1.808	1.781
	Most probable		1.833	1.852	1.833	1.826
	Max*		1.874	1.876	1.859	1.871
Y_R $\times 10^{-4}$	Min*	—	2.383	2.442	2.410	2.375
	Most probable		2.426	2.467	2.436	2.423
	Max*		2.469	2.492	2.463	2.471

*95% confidence limits

Table 3
Standard deviations of the measurement noise

Measurement		Assumed	Estimated (*a posteriori*)			
		(*a priori*)	M_1	M_2	M_3	M_4
Ethanol [g/L]	σ_E	1.5	1.3	1.5	1.4	1.4
Density [g/L]	σ_D	1.5	1.2	1.3	1.3	1.5
Sugar [g/L]	σ_S	6.0	5.2	5.3	5.3	5.5
Refractive index [$\times 10^{-4}$]	σ_R	5.0	3.9	4.3	4.2	4.4

Table 4
Accuracy improvement due to data reconciliation

Measurement	Average estimated standard error (M_1)	Assumed standard measurement error	Ratio	Equivalent number of direct measurements
Final ethanol [g/L]	0.93	1.5	1.61	3
Initial density [g/L]	0.53	1.5	2.84	8
Initial sugar [g/L]	1.89	6.0	3.17	10
Initial refractive index	1.59×10^{-4}	5.0×10^{-4}	3.13	10

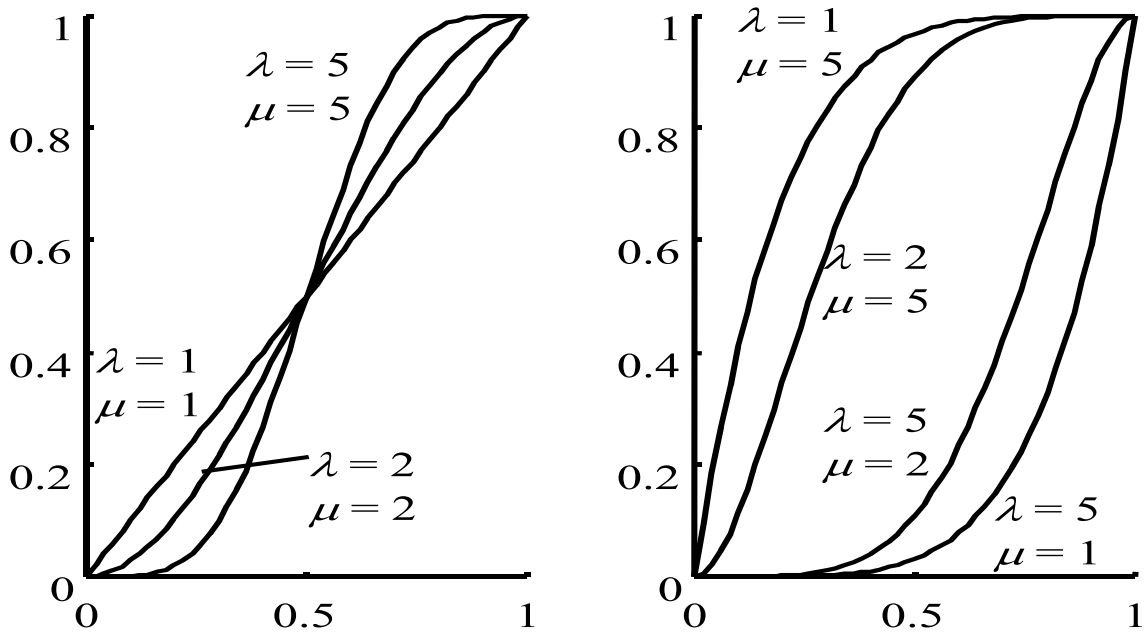

Figure 1

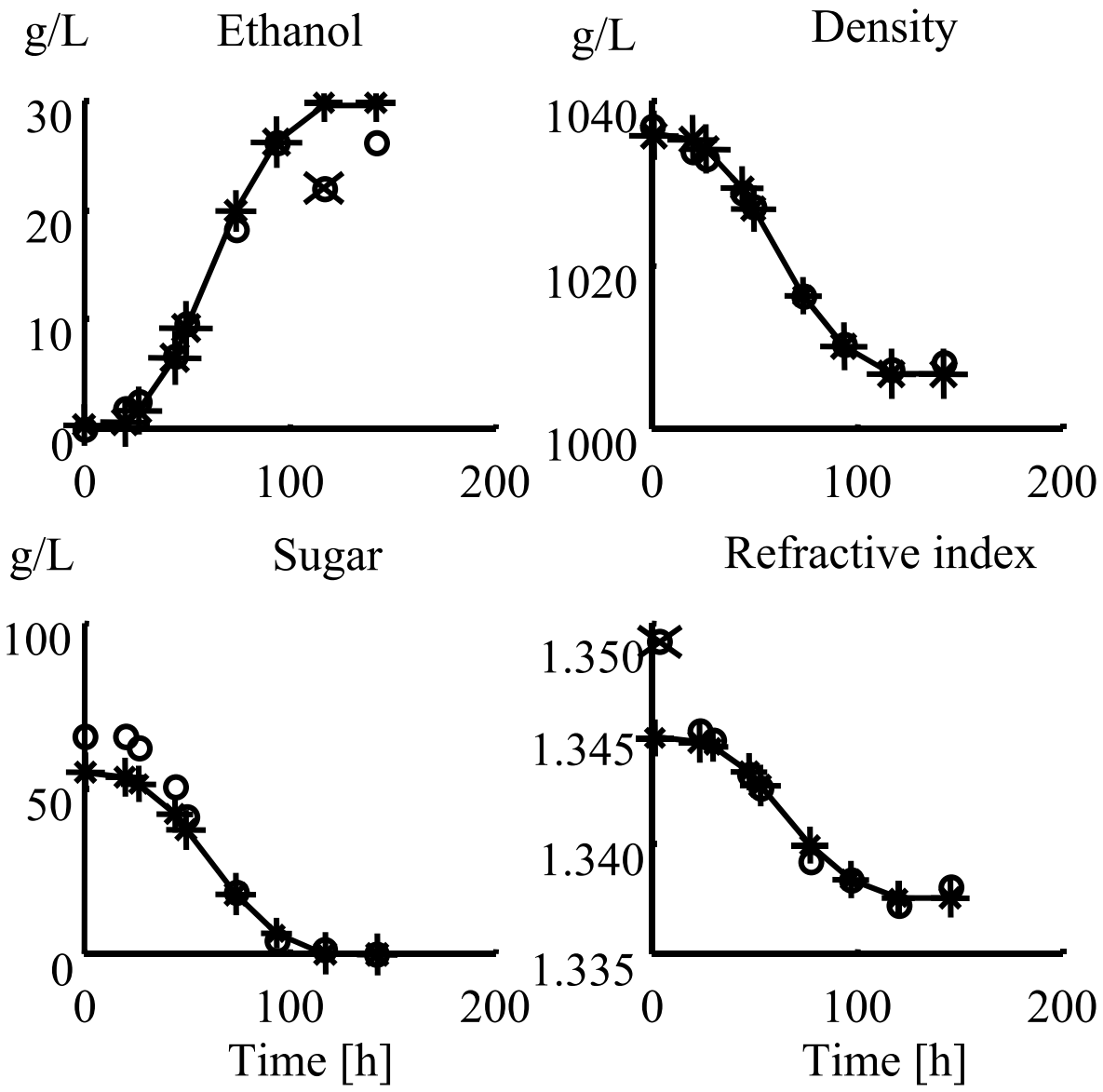
g/L
Ethanol
30
20
10
0
0
100
200
g/L
Density
1040
1020
1000
0
100
200
g/L
Sugar
100
50
0
0
100
200
Time [h]
Refractive index
1.350
1.345
1.340
1.335
0
100
200
Time [h]

Figure 2

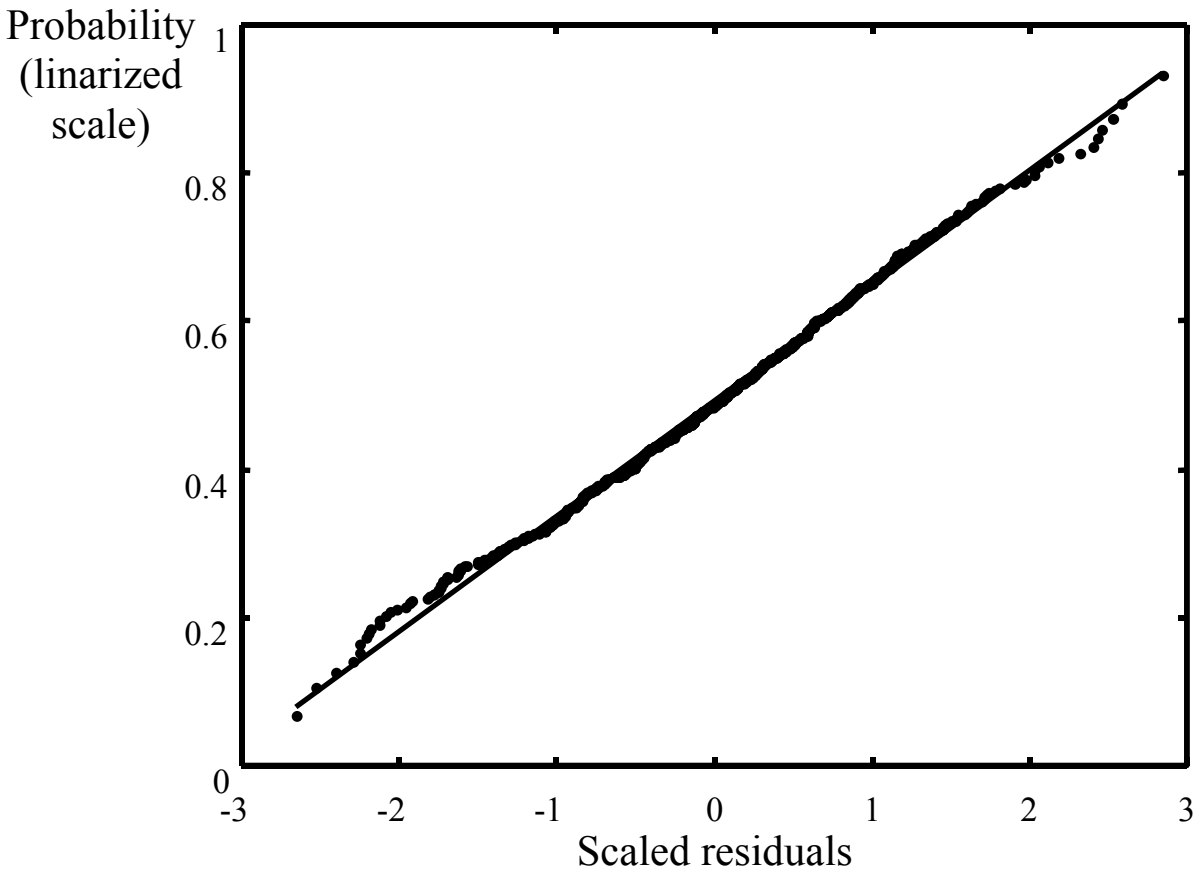
Probability
(linarized
scale)
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Scaled residuals

Figure 3

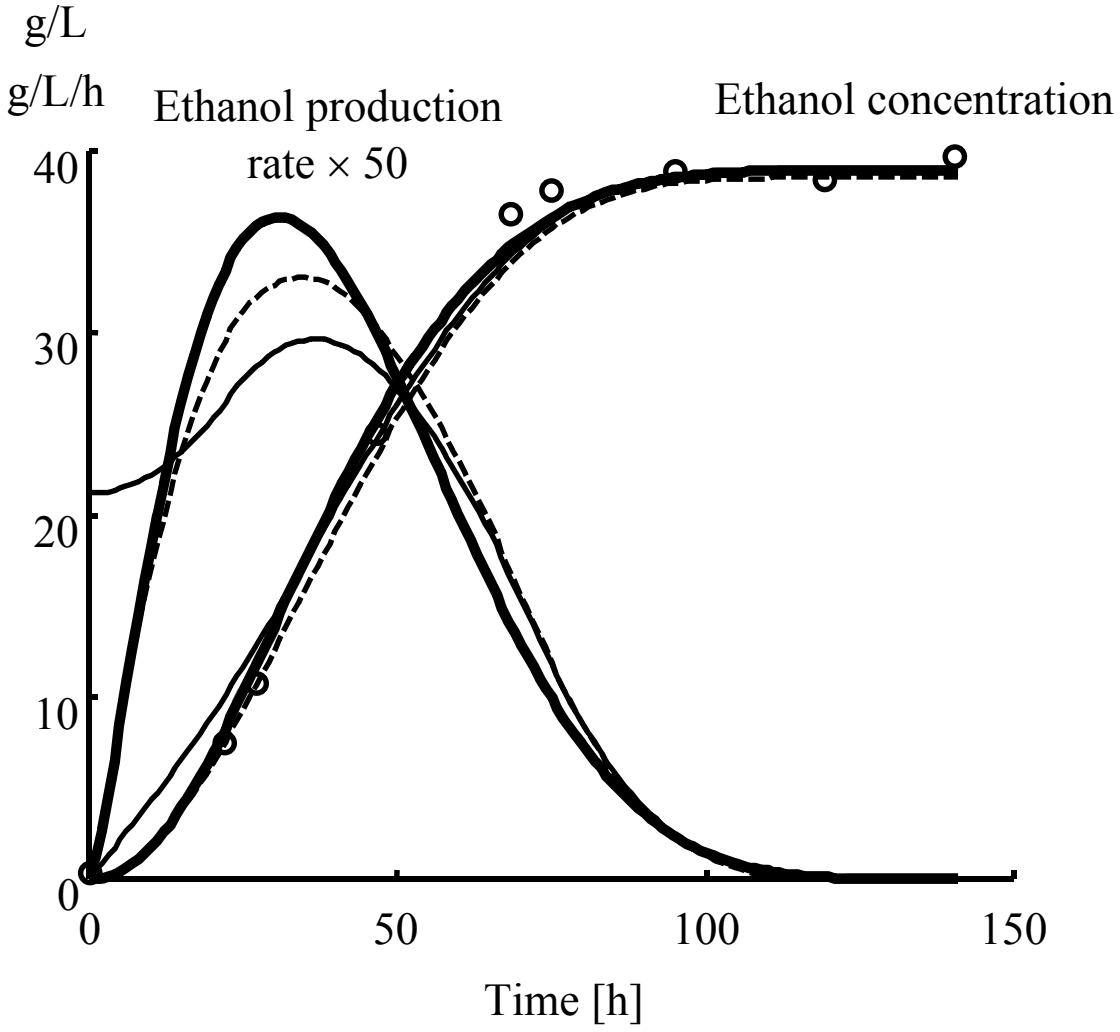
g/L
g/L/h
Ethanol production
rate × 50
Ethanol concentration
40
30
20
10
0
0
50
100
150
Time [h]

Figure 4

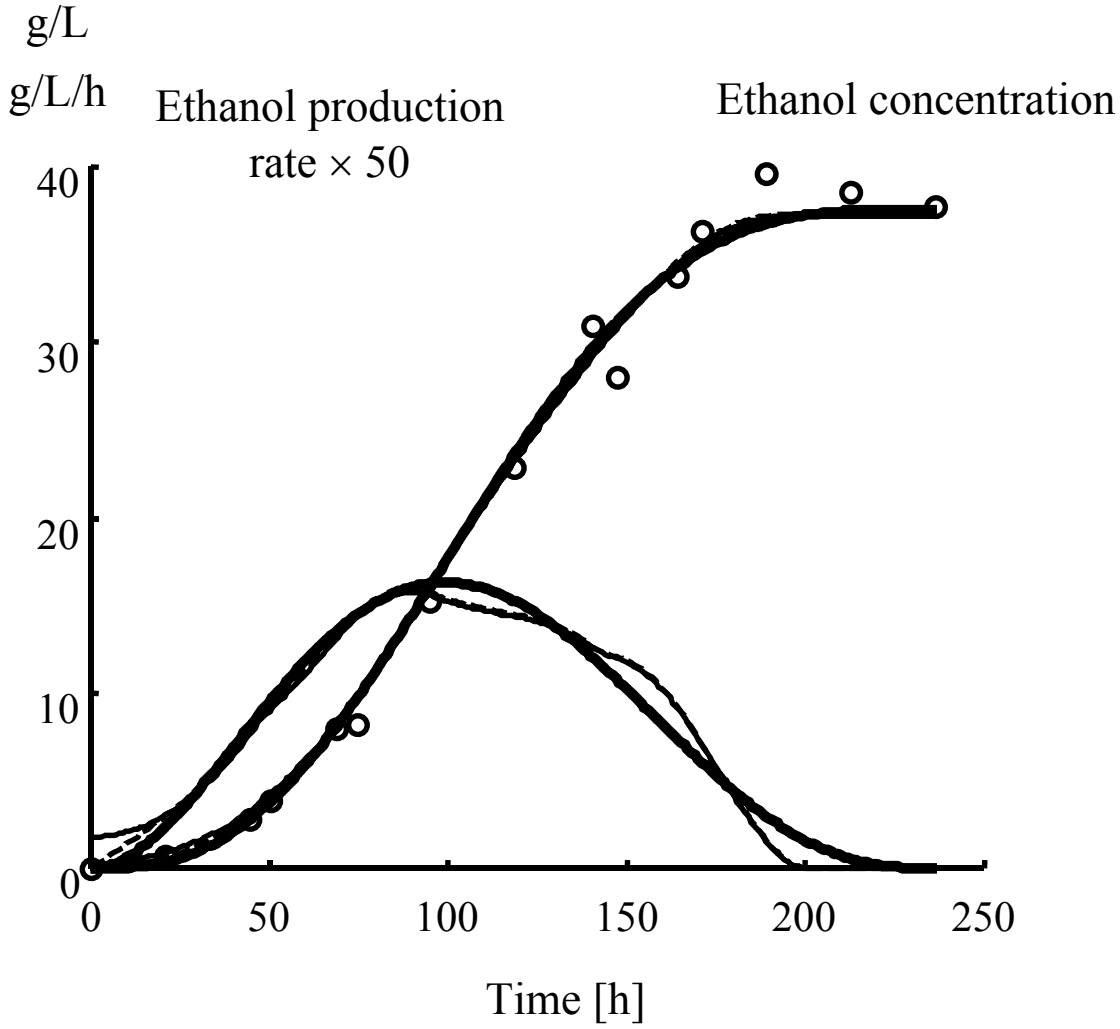
g/L
g/L/h
Ethanol production
rate × 50
Ethanol concentration
40
30
20
10
0
0
50
100
150
200
250
Time [h]

Figure 5

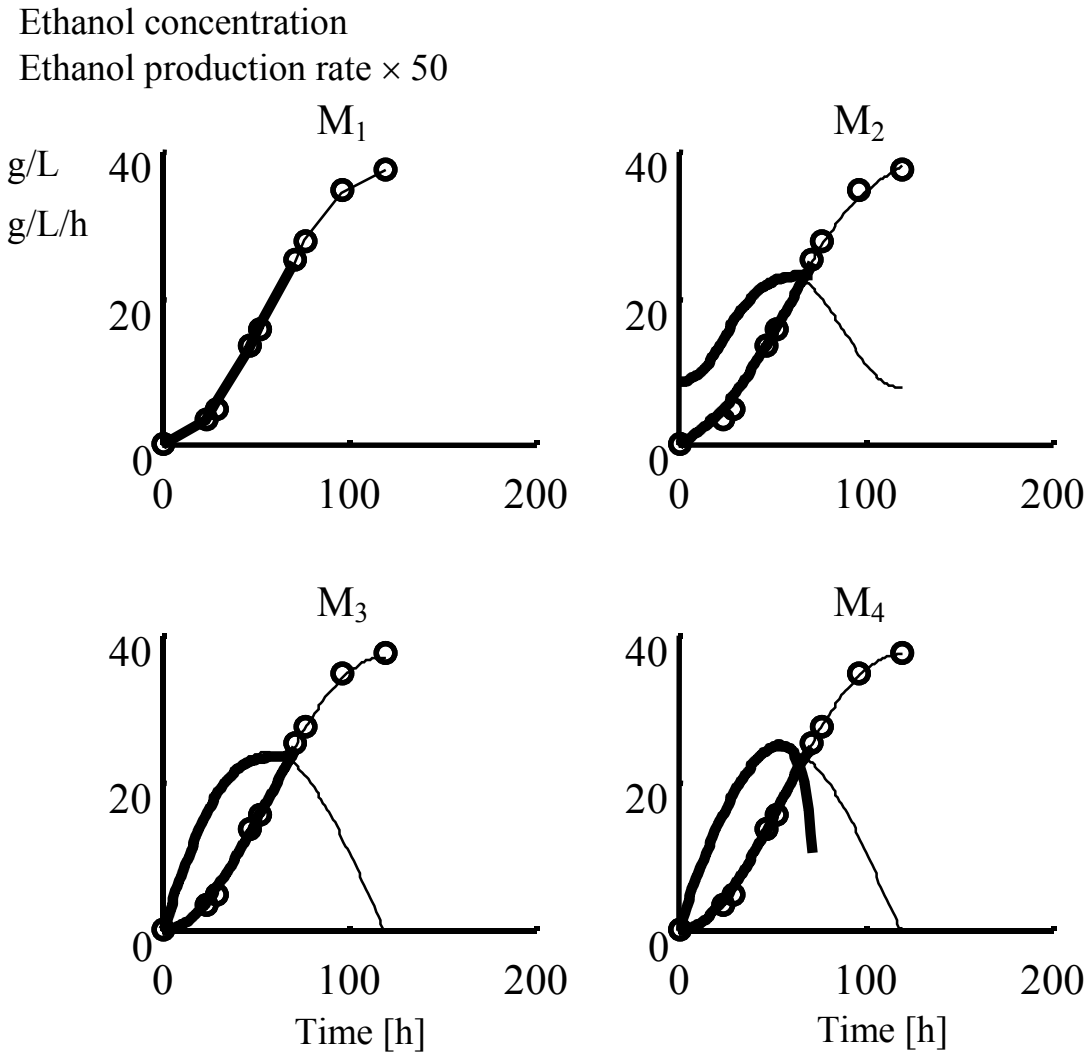
Ethanol concentration
Ethanol production rate × 50
M1
M2
M3
M4
g/L
g/L/h
40
20
0
0
100
200
Time [h]

Figure 6

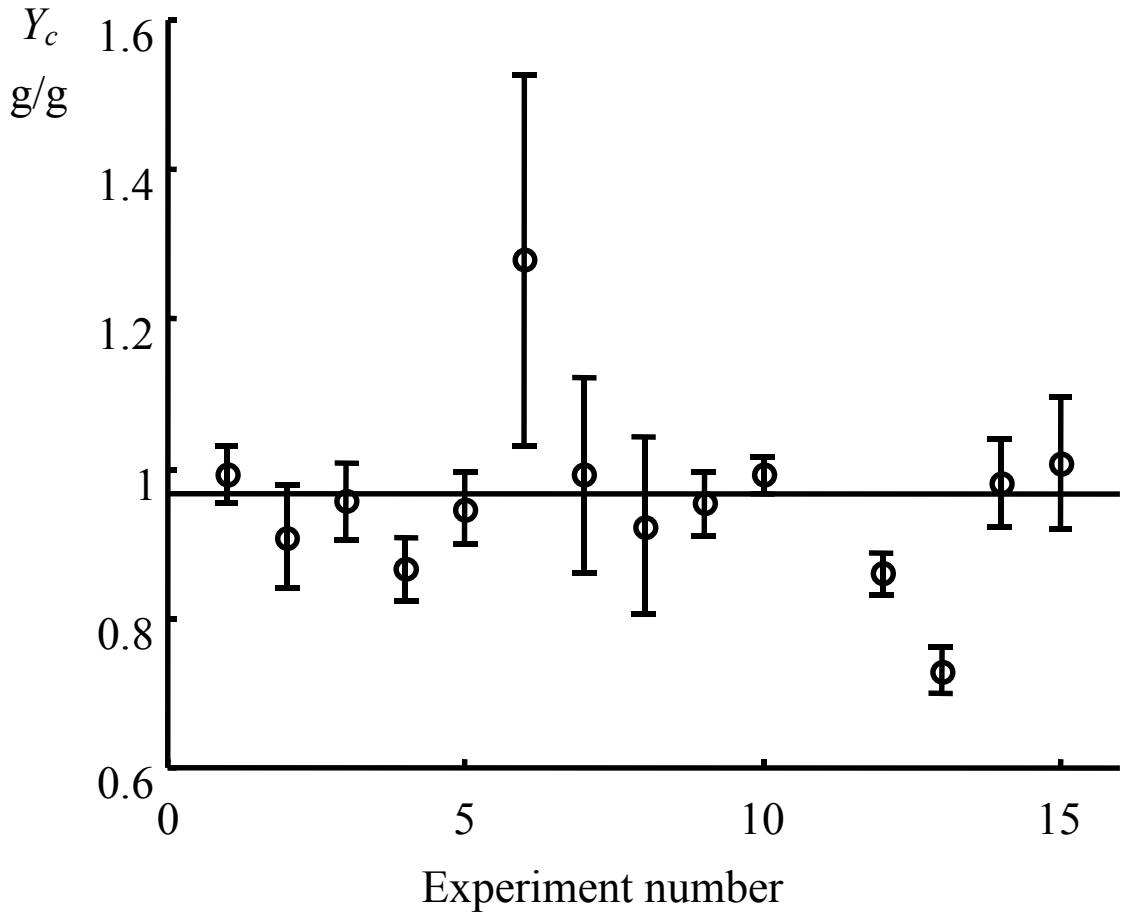

Figure 7