Data-driven Linear Predictor based on Maximum Likelihood Nonnegative Matrix Decomposition for Batch Cultures of Hybridoma Cells

Guilherme A. Pimentel [*] **Laurent Dewasme** [*] **Alain Vande Wouwer** [*]

[*] *Systems, Estimation, Control and Optimization Group (SECO),*
University of Mons, Belgium
(e-mail:<guilherme.araujopimentel, laurent.dewasme,
alain.vandewouwer>@umons.ac.be)

Abstract: This paper presents an original design of low-rank linear predictors of nonlinear process state variables based on nonnegative matrix decomposition (NMD). Therefore, this predictor is data-driven and does not require an accurate model description of the process. In addition, measurement errors are considered, conferring maximum likelihood (ML) properties to the estimator and resulting in a maximum likelihood nonnegative matrix decomposition (MLNMD) formulation. The latter is validated in simulation with a model developed by the authors, describing monoclonal antibody (MAb) production from sequential batch hybridoma cell cultures that are further validated with real-life experimental data. To this end, two available experimental data sets are used for direct and cross-validation, highlighting the good predictive properties of the method.

Keywords: nonnegative matrix decomposition, maximum likelihood, measurements errors, hybridoma cells cultures.

1. INTRODUCTION

The nonnegative matrix decomposition (NMD) (also called nonnegative matrix factorization and positive matrix factorization) is a recent methodology that decomposes a nonnegative matrix (i.e. matrix composed by zeros and positive values) into two low-rank nonnegative matrices: one called the basis matrix and the other the mixing matrix. One of the first applications of NMD in the field of engineering was achieved by Lee and Seung (1999) in image processing. The NMD is able to extract patterns or unmix signals involved in a data set (Cichocki et al., 2009) and has been successfully applied to processes that present the intrinsic property to produce nonnegative data. Typical examples of NMD are the unsupervised setting in image and natural language processing (Donoho and Stodden, 2003; Mysore, 2012), music instrument recognition (Smaragdis et al., May 2014), text database classification (Luong and Nayak, 2019) and recommender system (Koren et al., 2009) to mention a few. It has also been successfully exploited in a variety of applications in computational biology (Devarajan, 2008), which includes endogenous metabolite discovery (Bartel et al., 2013), cellular identity (Gao and Welch, 2020), genes and clustering samples (Liu et al., 2018) and multi-omics biological data (Wang et al., 2018).

To be suitable for all these applications, several different NMD algorithms have been proposed. One of the first mathematical descriptions of the nonnegative factor model with optimal utilization of the error estimates of data values was presented by Paatero and Tapper (1994). Thereafter, several extensions and simplifications have been proposed as the multiplicative update algorithms, proposed in Lee and Seung (2001), *semi-NMD*, which removes the non-negative constraints on the data, *sparse-NMD* that is used to reduce the non-uniqueness of solutions and enhance interpretability of the NMD results, *kernel-NMD* where the optimization of the model is dimension-free, *orthogonal-NMD* that imposes the orthogonality constraint to enhance sparsity and *weighted-NMD* for data sets with missing elements, see Cichocki et al. (2009) and Berry et al. (2007) as well as references therein.

Although there are other applications of NMD, such as dimensionality reduction and preprocessing, this paper will consider the approach from a modeling perspective. Differently from maximum likelihood principle component analysis (MLPCA) (Bernard and Bastin, 2005; Mailier et al., 2012; Dewasme et al., 2017), which extracts from the measurement data sets the number of biochemical reactions and a stoichiometric basis, the NMD derives, in an unstructured way, the fundamental compounds involved in these reactions. This can be seen as a complementary and easy-to-use tool to be combined with the MLPCA when a priori knowledge on the components involved in the reaction rates is not available.

The resulting maximum likelihood NMD (MLNMD) formulation is assessed when applied to the hybridoma cell batch culture process, producing monoclonal antibodies (MAbs). This process complexity, the limited amount of data, and measurement noise levels hamper the process of modeling for monitoring, prediction, and control. Also, the measurements of some of the compounds involved in the production of MAbs are costly and time-consuming, requiring great effort for data acquisition. A basic approach in modeling for monitoring and control is to represent the bioprocess as a macroscopic model (Bastin and Dochain, 1990), as proposed in the literature for the MAb production process (Nolan and Lee, 2011; Amribt et al., 2014; Dewasme et al., 2017; Yilmaz et al., 2020).

This paper proposes an original maximum likelihood nonnegative matrix decomposition (MLNMD) algorithm which accounts for the measurement noise in the estimation of the decomposed matrices. To validate the MLNMD, first, we select the model proposed by Dewasme et al. (2017) to generate noisy data. Furthermore, based on the analysis of the decomposition, two experimental data sets are used to validate and cross-validate the proposed low-rank linear predictor.

This paper is organized as follows. Section 2 revisits the macroscopic hybridoma cell model, which is used to analyze the decomposition results and validate the MLNMD. Section 3 presents one of the well-established algorithms of NMD - the alternating least-squares (ALS) -, and introduces the MLNMD algorithm. The analysis and the validation of MLNMD using simulation data sets from hybridoma cell batch cultures are presented in Section 4. In Section 5, the experimental data is decomposed by the MLNMD algorithm and the linear predictor is designed, while a second experimental data set cross-validates the proposed method. Section 6 points to open problems and concludes the paper.

2. DYNAMIC MODEL OF HYBRIDOMA CELL CULTURES

The reaction scheme presented in Dewasme et al. (2017), macroscopically describing hybridoma cell catabolism through three metabolic mechanisms (substrate oxidation and overflow, and biomass decay), reads:

(a) Substrate oxidation:

$$k_{31}G + k_{41}Gn \xrightarrow{\varphi_1} X_v + k_{61}MAb. \tag{1}$$

(b) Substrate overflow:

$$k_{32}G + k_{42}Gn \xrightarrow{\varphi_2} X_v + k_{52}L. \tag{2}$$

(c) Biomass death

$$X_v \xrightarrow{\varphi_3} X_d + k_{63}MAb. \tag{3}$$

where X_v, X_d, G, Gn, L and MAb are the concentration of the viable biomass, dead biomass, glucose, glutamine, lactate and monoclonal antibodies (MAb), respectively, φ_j are reaction rates and k_i are the stoichiometric parameters of the process.

From the reaction scheme, the corresponding mass balance equations can be written as follows:

$$\frac{dX_v}{dt} = \varphi_1 + \varphi_2 - \varphi_3, \tag{4a}$$

$$\frac{dX_d}{dt} = \varphi_3, \tag{4b}$$

$$\frac{dG}{dt} = -k_{31}\varphi_1 - k_{32}\varphi_2, \tag{4c}$$

$$\frac{dGn}{dt} = -k_{41}\varphi_1 - k_{42}\varphi_2, \tag{4d}$$

$$\frac{dL}{dt} = k_{52}\varphi_2, \tag{4e}$$

$$\frac{dMAb}{dt} = k_{61}\varphi_1 + k_{63}\varphi_3, \tag{4f}$$

where the reaction rates are discontinuous functions of the form:

$$\varphi_1 = min(\mu_G, \mu_{Gmax}), \tag{5a}$$

$$\varphi_2 = max(0, (\mu_G - \mu_{Gmax})), \tag{5b}$$

$$\varphi_3 = \mu_{dmax}\frac{K_{Gd}}{K_{Gd} + G}\frac{K_{Gnd}}{K_{Gnd} + Gn}X_v, \tag{5c}$$

where

$$\mu_G = \mu_{max1}\frac{Gn}{K_{Gn} + G_n}X_v, \;\; \mu_{Gmax} = \mu_{max2}X_v. \tag{6}$$

This model will be considered as a reference in the following result analyses.

3. NONNEGATIVE MATRIX DECOMPOSTION (NMD) METHODS

3.1 NMD

The NMD methods can be split into two main branches: (i) Exact NMD and (ii) Approximate NMD. Exact NMD is an essential tool for linear algebra, and it is closely related to the smallest r such that X admits an Exact NMF size r. On the other hand, Approximate NMD, the subject of this paper, takes into account the statistical properties of the measurement errors and is more suitable for practical applications where an Exact NMD is unlikely (see discussion about Exact and Approximate NMD in Gillis (2020)). For simplicity, the Approximate NMD is referred to as NMD in the text.

One widely used algorithm to obtain the NMD models is the alternating least squares (Berry et al., 2007), which can be considered as a maximum likelihood method if all measurement error standard deviations have the same normal distribution (i.e. independently and identically distributed (i.i.d)).

The NMD decomposes a nonnegative matrix $X \in \mathbf{R}_+^{m\times n}$ in two matrices $W \in \mathbf{R}_+^{m\times r}$ (basis matrix) and $H \in \mathbf{R}_+^{r\times n}$ (mixing matrix), where r, selected by the user, is the basis dimension of the decomposition represented as follows:

$$X_{m\times n} = W_{m\times r} \cdot H_{r\times n}, \tag{7}$$

where m is the number of collected samples (observations), n is the number of the process compounds, which, in this study, are the number of extracellular measurement components (i.e. biomass, metabolites, and substrate concentrations).

The W and H matrices are chosen to minimize the objective function that is defined as in Berry et al. (2007):

$$J = \frac{\|X - W \cdot H\|_F}{\sqrt{n \cdot m}}, \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm. The decomposition is obtained by the solution of a bilinear problem as there are two matrices to be determined. In the NMD, the minimization of the bilinear problem uses the alternating least squares (ALS) algorithm, which can be represented by the following steps.

1. Given an $m \times n$ data matrix X_0, subtract the offset of each column vector of the measurements, in way that each set of data starts or ends in zero $(X = X_0 - min(X_0))$;
2. Initialize W randomly or by using a specific deterministic strategy;
3. Estimate H from the matrix equation $W^TWH = W^TX$ by solving
$$\min_H J = \frac{\|X - W \cdot H\|_F}{\sqrt{n \cdot m}}, \text{ fixed } W; \tag{9}$$
4. Set all the negative elements of H to zero;
5. Estimate W from the matrix equation $HH^TW^T = HX^T$ by solving
$$\min_W J = \frac{\|X^T - H^T \cdot W^T\|_F}{\sqrt{n \cdot m}}, \text{ fixed } H; \tag{10}$$
6. Set all negative elements of W to zero;

7. Test $X - W \cdot H < tolValue$ or if the maximum iteration number *maxIter*, set by the user, is exceeded, terminate. Otherwise return to Step 3.
8. Add the offset values from Step 1 to $W \cdot H$.

In this paper the nonnegative matrix decomposition is called by the function $[W,H] = nnmd(X,r)$ (as embedded in MatLab (Berry et al., 2007)), where r is the decomposition dimension.

3.2 MLNMD

As aforementioned, the maximum likelihood nonnegative matrix decomposition (MLNMD) is a modeling method accounting for the measurement errors during NMD pattern extraction. This is usually achieved by minimizing the weighted residual of the sum of distances of the data with some r-dimension model, corresponding to :

$$S^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2}, \qquad (11)$$

where $\hat{x}$ is the estimate value of the measurement and σ_{ij}^2 are the measurement variances. A straightforward implementation of the MLNMD algorithm, inspired from the combined works of Wentzell et al. (1997) and Mailier et al. (2012) is presented in Algorithm 1.

Algorithm 1. MLNMD

—

1. Given an $m \times n$ data matrix X_0, subtract the offset of each column vector of the measurements, in way that each set of data starts or ends in zero ($X = X_0 - min(X_0)$).
2. Given an $m \times n$ data matrix X with no offset and a corresponding $m \times n$ matrix Q of measurement error variances, use NMD to obtain the initial approximation to the MLNMD. The NMD is truncated to rank r.

$$[W,H] = nnmd(X,r). \qquad (12)$$

3. Transpose X and Q and compute the maximum likelihood estimates using H.

$$X = X^T, \quad Q = Q^T, \quad \Sigma_i = diag(q_i), \qquad (13)$$

$$\tilde{x}_i = H(H^T \Sigma_i^{-1} H)^{-1} H^T \Sigma_i^{-1} x_i, \qquad (14)$$

where x_i is a column vector of the transposed matrix X, and the cost function is computed as

$$S_1^2 = \sum_{i=1}^{m} (x_i - \tilde{x}_i)^T \Sigma_i^{-1} (x_i - \tilde{x}_i) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(x_{ji} - \tilde{x}_{ji})^2}{\sigma_{ji}^2}. \qquad (15)$$

4. Compute the NMD of $\tilde{X}$ from step 3 and obtain new H

$$[W,H] = nnmd(\tilde{X},r). \qquad (16)$$

5. Repeat step 3 to estimate the MLNMD in the origin space.

$$X = X^T, \quad Q = Q^T, \quad \Phi_j = diag(q_j), \qquad (17)$$

$$\tilde{x}_j = H(H^T \Phi_j^{-1} H)^{-1} H^T \Phi_j^{-1} x_j, \qquad (18)$$

$$S_2^2 = \sum_{j=1}^{n} (x_j - \tilde{x}_j)^T \Phi_j^{-1} (x_j - \tilde{x}_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(x_{ji} - \tilde{x}_{ji})^2}{\sigma_{ji}^2}. \qquad (19)$$

6. Compute the NMD of $\tilde{X}$ to obtain a new estimate of the MLNMD solution in the original space.

$$[W,H] = nnmd(\tilde{X},r). \qquad (20)$$

7. Compute the convergence parameter λ or maximum predefined iteration value.

$$\lambda = (S_1^2 - S_2^2)/S_2^2. \qquad (21)$$

If λ is less than the convergence limit or if the maximum iteration number *maxIter*, set by the user, is exceeded, terminate. Otherwise return to Step 3.

8. Add the offset obtained in the Step 1 to final $\tilde{X}$.

—

4. SIMULATION RESULTS AND DECOMPOSITION ANALYSIS

Model (4) is used to generate an 8-day simulated experiment with a sampling period of 0.1 days. The corresponding parameters are presented in Table 1. Also, for each measurement is added an uncorrelated white-noise with 5% of relative standard deviation. Thus, the generated data is nonnegative with $m = 81$ samples from $n = 6$ different component concentrations, i.e viable biomass X_v, dead biomass X_d, glucose G, glutamine Gn, lactate L and monoclonal antibodies MAb.

4.1 Selection of the r-dimension

In order to select the r-dimension, the latter is increased until the rank of H is reached. In the current case, the maximum admissible r-dimension is 4. Figure 1 presents the fitting root-mean-square errors of NMD for increasing values of r.

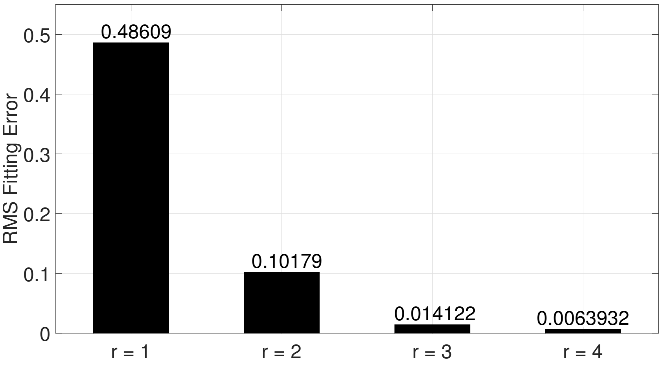
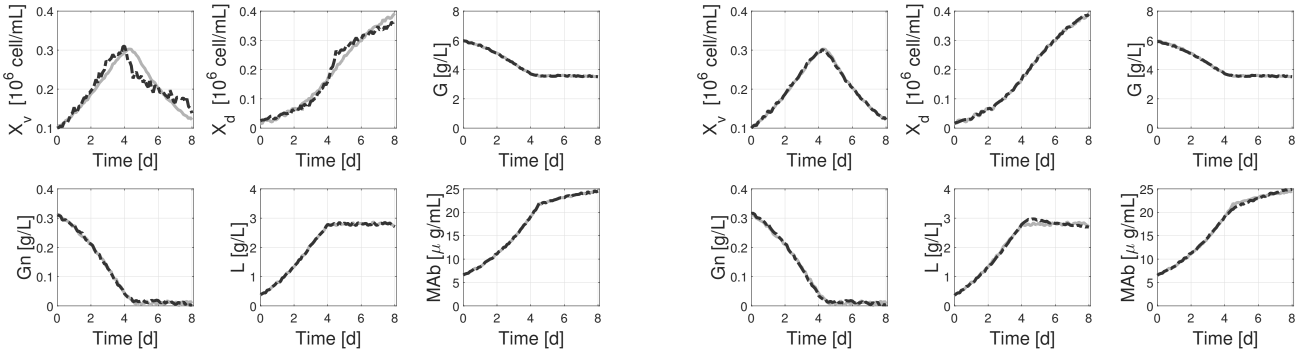


Fig. 1. NMD r-dimensions ans its root-mean-square errors.

Analyzing the basis matrix W we find out that from $r = 1$ to $r = 3$, there is a direct link between the basis signals, contained in basis matrix W, and the measured compounds from the process - this will be clear in the next section. Despite the root-mean-square error being the smallest with $r = 4$, the obtained basis signals do not contain dynamic relations to the gathered process measurements. Hence, to keep the relation between the basis matrix W and the bioprocess measurements, the select decomposition dimension is $r = 3$.

4.2 Numerical analysis: NMD vs. MLNMD

Analyzing Figure 2a, with the selected basis dimension $r = 3$, the reconstruction of the measurements by the decomposition $\hat{X} = W \cdot H$ badly fits, especially for viable and dead biomasses, despite a quite accurate fitting regarding the other compound concentrations. This insufficient behavior of the NMD motivates the inclusion of a maximum likelihood criterion to improve the quality of the estimates.



(a) NMD validation of the macroscopic hybridoma cells data set.

(b) MLNMD validation of the macroscopic hybridoma cells data set.

Fig. 2. Validations of the simulation data. (a) NMD and (b) MLNMD results. Continuous gray lines are the measurements from the data set (X) and dashed black lines are the decomposition validation ($\hat{X}$).

Table 1. Simulation parameters, obtained from Dewasme et al. (2017).

Parameters	Values	Parameters	Values
μ_{max1}	$0.484\ d^{-1}$	k_{31}	3.12
μ_{max2}	$0.319\ d^{-1}$	k_{32}	15.2
K_{Gn}	$0.0089\ g/L$	k_{41}	0.624
K_{Gd}	$1.58\ g/L$	k_{42}	1.22
K_{Gnd}	$1.33\ g/L$	k_{52}	23.9
μ_{dmax}	$0.866\ d^{-1}$	k_{61}	43.5
K_G	$0.100\ g/L$	k_{63}	14.2
$X_v(0)$	$0.100\ cells/ml$	$X_d(0)$	$0.0151\ cells/ml$
$G(0)$	$5.99\ g/L$	$Gn(0)$	$0.303\ g/L$
$L(0)$	$0.360\ g/L$	$MAb(0)$	$6.53\ \mu g/ml$

The same data set generated from model (4) simulation is used to obtain the maximum likelihood decomposition $[\widetilde{W}, \widetilde{H}] = mlnnd(X_0, Q, r)$, where Q is the matrix of measurement error variances and $\widetilde{(\cdot)}$ denotes the maximum likelihood solution. Figure 2b shows the validation of the MLNMD in dashed black lines. The maximum likelihood solution for the bilinear problem is expressed by the following linear relation:

$$\underbrace{\begin{bmatrix} \widetilde{X_v} \\ \widetilde{X_d} \\ \widetilde{G} \\ \widetilde{Gn} \\ \widetilde{L} \\ \widetilde{MAb} \end{bmatrix}}_{\widetilde{X}} = \underbrace{\begin{bmatrix} 0 & \mathbf{0.0313} & 0 \\ \mathbf{0.0209} & 0.0026 & 0.0006 \\ 0.0000 & 0 & \mathbf{0.9915} \\ 0.0001 & 0.0011 & \mathbf{0.1303} \\ \mathbf{0.1206} & \mathbf{0.2597} & 0.0003 \\ \mathbf{0.9925} & \mathbf{0.9652} & 0 \end{bmatrix}}_{\widetilde{H}} \cdot \underbrace{\begin{bmatrix} \widetilde{w}_1 \\ \widetilde{w}_2 \\ \widetilde{w}_3 \end{bmatrix}}_{\widetilde{W}} + \Phi_{offset}, \quad (22)$$

where $\Phi_{offset} = [min(X_v)\ \ min(X_d) min(G)\ \ min(Gn)\ \ min(L)\ \ min(MAb)]^T$, which can be obtained directly from the data set and, for the sake of clarity, the values of matrix $\widetilde{W}$ are represented in Figure 3. Originally $\widetilde{X}$, $\widetilde{H}$ and $\widetilde{W}$ should be expressed by its transpose, but we suppressed the transpose symbol for notation simplification.

Evaluating the decomposition results represented in (22) and Figure 3, it can be easily seen one of the fundamental properties of the nonnegative matrix decomposition: the addition of sparsity in the decomposed matrices (Cichocki et al., 2009). To significantly highlight this characteristic in (22), the most significant values of each row in $\widetilde{H}$ are represented in bold. Note that some values are extremely small and not zeros due to noise and mainly the selected value of r-dimension. Nonetheless, analyzing this matrix, it is clear that the measurements of viable biomass X_v can be approximated by the element in row 1 and column 2, multiplied by all the elements of the second row of $\widetilde{W}$, i.e. $X_v \approx \widetilde{X}_v = \widetilde{h}_{12} \cdot \widetilde{w}_2$.

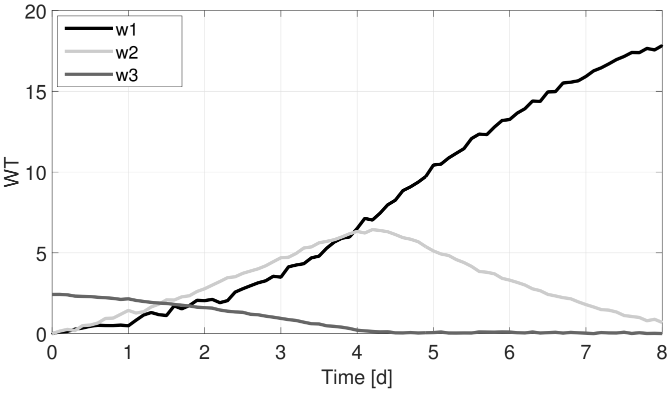


Fig. 3. MLNMD $\widetilde{W}$ base signals, $r = 3$.

Likewise, it is straightforward to conclude that the dynamical behavior of $\widetilde{w}_2$, represented by the light gray line in Figure 3, is very close to the dynamical behavior of the X_v measurements, presented in Figure 2b. In the same way, the dead biomass measurements can be approximated by ($X_d \approx \widetilde{X}_d = \widetilde{h}_{21} \cdot \widetilde{w}_1$) and both glucose and glutamine can be approximated by only using the dynamics of their respective element located in column 3 of $\widetilde{H}$ multiplied by $\widetilde{w}_3$.

As previously mentioned, the analysis of the matrix $\widetilde{H}$ is corroborated by the comparison of the dynamical behavior of $\widetilde{W}$ (Figure 3) and the dynamical behavior of $X_v \to \widetilde{w}_2$, $X_d \to \widetilde{w}_1$ and both $G, Gn \to \widetilde{w}_3$ (Figure 2b). Furthermore, taking into account the remaining compounds, the Lactate L and monoclonal antibodies $MAbs$ can be reconstructed by the linear combinations of $\widetilde{w}_1$ and $\widetilde{w}_2$.

Moreover, the analysis of the sparsity of $\widetilde{H}$ revealed the linear interconnections between the pairs $(X_v, X_d) \to (L, MAb)$ and the mutual relation of $G \leftrightarrow Gn$. The NMD exposed the fundamental compounds involved in reactions (1), (2) and (3). This also discloses the fundamental measurements required to reconstruct all the other process compounds. This is the subject of the next section, where an experimental data set (HB1) is used for the design of the linear predictor, cross-validated with a second experimental data set (HB2).

Remark 1. As previously mentioned, NMD derives, in an unstructured way, the fundamental compounds involved in the process reactions. Therefore, this can be seen as a complementary and easy-to-use tool to be combined with the MLPCA, for instance, when a priori knowledge of the components involved in the reaction rates is not available.

5. MLNMD: EXPERIMENTAL DATA AND LINEAR PREDICTOR DESIGN

5.1 MLNMD: Experimental Data Validation

Two batch cultures of a hybridoma strain (called, HB1 and HB2) were performed in 200 mL T-flasks. At the initial time of each batch, biomass is kept in the reactor, while the metabolites (lactate, ammonia, and monoclonal antibodies) are withdrawn and the substrate concentrations (glucose and glutamine) are set to prescribed values (respectively ranging between 6 and 7 g/L, and 0.3 and 0.4 g/L). The culture time is approximately 7 days for HB1 and 9 days for HB2. Measurements are taken once every day, for more information about culture medium and protocol of measurements, see Dewasme et al. (2017).

In Figure 4a, the experimental data from HB1 and the corresponding confidence intervals, obtained from Dewasme et al. (2017), are represented by the error bars. The sample time, as mention before, is 1 day and the batch durations are 7 and 9 days for HB1 and HB2, respectively. Thus, it generates a data set with $m = 7$ (HB1) and $m = 9$ (HB2) samples, both with six component concentrations ($n = 6$).

The HB1 data set is applied to the MLNMD Algorithm 1, with the decomposition basis dimension of three, $r = 3$. The MLNMD of HB1 data is presented in (23) and Figure 4b, and its validation is presented in Figure 4a in black dashed lines.

$$\underbrace{\begin{bmatrix} \bar{X}_v \\ \bar{X}_d \\ \bar{G} \\ \bar{Gn} \\ \bar{L} \\ \bar{MAb} \end{bmatrix}}_{\bar{X}_{(HB1)}} = \underbrace{\begin{bmatrix} 0.0033 & \mathbf{0.0257} & 0 \\ \mathbf{0.0182} & 0.0003 & 0 \\ 0 & 0.0557 & 0.9915 \\ 0 & 0.0018 & \mathbf{0.1304} \\ \mathbf{0.1547} & \mathbf{0.1310} & 0 \\ \mathbf{0.9878} & \mathbf{0.9895} & 0 \end{bmatrix}}_{H_{(HB1)}} \cdot \underbrace{\begin{bmatrix} w_{(HB1),1} \\ w_{(HB1),2} \\ w_{(HB1),3} \end{bmatrix}}_{W_{(HB1)}} + \Phi_{offset}, \tag{23}$$

where the elements with $\bar{\cdot}$ are the MLNMD of the experimental data set HB1 and $\Phi_{offset} = [min(X_v)\ min(X_d)\ min(G)\ min(Gn)\ min(L)\ min(MAb)]^T$ are provided by the original data set.

Similarly to the results of Section 4, the sparsity of matrix $H_{(HB1)}$ reveals the linear relations between the pairs $(X_v, X_d) \rightarrow (L, MAb)$ and the mutual relation of $G \leftrightarrow Gn$. Also, (23) and Figure 4b shows that the dynamical behavior of $w_{(HB1),1}$, $w_{(HB1),2}$ and $w_{(HB1),3}$ are linked to the dynamic behavior of $\bar{X}_d$, $\bar{X}_v$ and both $\bar{G}$ and $\bar{Gn}$, respectively. These results initiate the design of a linear low-rank predictor for the hybridoma cell cultures.

5.2 Linear Predictor Design

First, consider a linear predictor in the form

$$Y = K\Xi, \tag{24}$$

where $Y \in \mathbf{R}_+^{n\times m}$ is composed by the predicted compounds (i.e. $Y = \left[\breve{X}_v\ \breve{X}_d\ \breve{G}\ \breve{Gn}\ \breve{L}\ \breve{MAb}\right]^T$), $K \in \mathbf{R}_+^{n\times r}$ and $\Xi \in \mathbf{R}_+^{r\times m}$. Based on the sparsity of the matrix $H_{(HB1)}$ presented in equation (23), the structure of K and Ξ are straightforward as

$$K = \begin{bmatrix} \frac{H_{(HB1),1}}{h_{(HB1),21}} & \frac{H_{(HB1),2}}{h_{(HB1),12}} & \frac{H_{(HB1),3}}{h_{(HB1),43}} \end{bmatrix}, \tag{25}$$

and

$$\Xi = \begin{bmatrix} X_{d\ (HB2)} \\ X_{v\ (HB2)} \\ Gn_{(HB2)} \end{bmatrix}, \tag{26}$$

where $H_{(HB1),1}/h_{(HB1),21}$ means that the first column of $H_{(HB1)}$ is divided by the element of row 2, column 1, of $H_{(HB1)}$ and the values of Ξ are inferred from assumed available measurements of the hybridoma cell culture database (i.e. reported in the HB2 data set). The outputs of the linear predictor (24) are presented in Figure 5, where the experimental data from HB2 and the related confidence intervals, obtained from Dewasme et al. (2017), are represented by the error bars.

Figure 5 shows a good predictive property of the linear predictor as the HB2 data set presents different dynamics and larger culture time (9 days) than HB1. Moreover, the MLNMD low-rank linear predictor has the advantage of being designed based on the data-driven approach only, without proceeding to the dynamic modeling of the process metabolites. It must also be highlighted that the provided predictions depend exclusively on the measurements of the viable biomass, dead biomass, and glutamine concentrations.

6. CONCLUSION

This paper presents a data-driven linear predictor for batch cultures of hybridoma cells. The design is obtained based on the nonnegative matrix decomposition considering possible normally distributed measurement errors, conferring a decomposition that is optimal in a maximum likelihood sense. The resulting straightforward MLNMD algorithm is proposed and validated with the help of two experimental data sets from hybridoma cell batch cultures. The approach validations present satisfactory predictions. Interesting perspectives concern the connections of this unstructured approach with dynamic model identifiability and observability frameworks.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the ProtoDrive project (convention *no. 2010119*) of the Win2Wal program of the Walloon Region (DGO6) and MabDrive project (convention *no. 1410085*), both achieved in collaboration with the CER Groupe. The scientific responsibility rests with its authors.

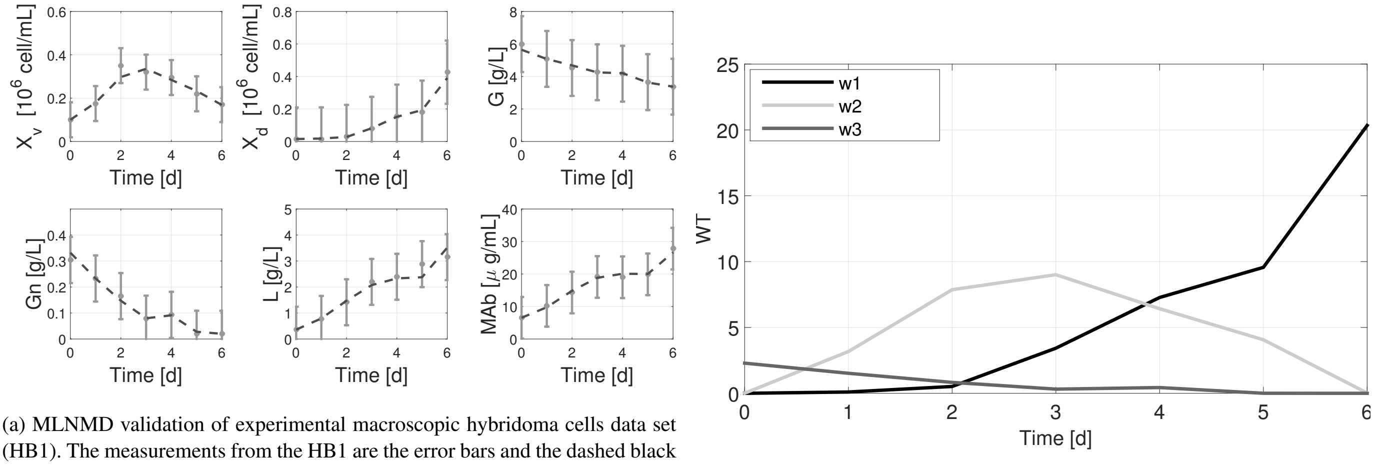
REFERENCES

Amribt, Z., Dewasme, L., Vande Wouwer, A., and Bogaerts, P. (2014). Optimization and robustness analysis of hybridoma cell fed-batch cultures using the overflow metabolism model. *Bioprocess Biosyst Eng*, 37, 1637–1652.

Bartel, J., Krumsiek, J., and Theis, F.J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal*.

Bastin, G. and Dochain, D. (1990). *On-Line Estimation and Adaptive Control of Bioreactors*. Volume 1 of Process Measurement and Control, Elsevier: Amesterdam.

Bernard, O. and Bastin, G. (2005). On the estimation of the pseudo stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Math. Biosci.*, 193, 51–77.



(a) MLNMD validation of experimental macroscopic hybridoma cells data set (HB1). The measurements from the HB1 are the error bars and the dashed black lines are the MLNMD validation.

(b) MLNMD $\overline{W}_{HB1}$ signal ($r = 3$).

Fig. 4. Validation of the experimental data HB1.

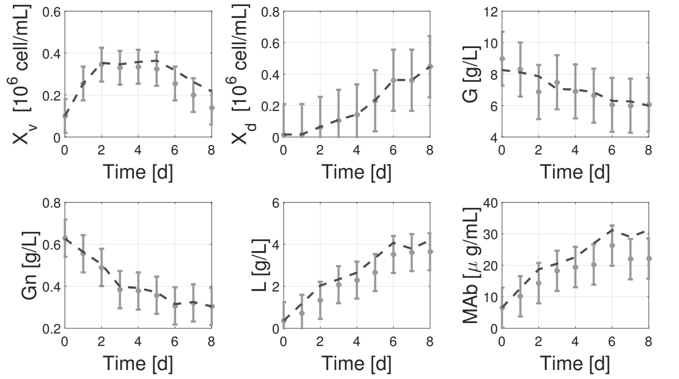


Fig. 5. MLNMD linear predictor cross-validation. The measurements from the HB2 are the error bars and the dashed black lines the linear predictor.

Berry, M.W., Brown, M., Langville, A.N., Paucac, V.P., and Plemmonsc, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52, 155 – 173.

Cichocki, A., Zdunek, R., Phan, A.H., and Amari, S.I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Hoboken, NJ: Wiley.

Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Computational Biology*, 4(7), 1–12.

Dewasme, L., Côte, F., Filee, P., Hantson, A.L., and Vande Wouwer, A. (2017). Macroscopic dynamic modeling of sequential batch cultures of hybridoma cells: an experimental validation. *Bioengineering*, 4, 17, 1 – 20.

Donoho, D. and Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*.

Gao, C. and Welch, J.D. (2020). Iterative refinement of cellular identity from single-cell data using online learning. *bioRxiv*.

Gillis, N. (2020). *Nonnegative Matrix Factorization*. SIAM.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer Society*.

Lee, D. and Seung, H. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13, 556–562.

Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.

Liu, J., Wang, D., Gao, Y., Zheng, C., Xu, Y., and Yu, J. (2018). Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3), 974–987.

Luong, K. and Nayak, R. (2019). *Linking and Mining Heterogeneousand Multi-view Data, Unsupervised and Semi-Supervised Learning*, chapter 9 : Clustering multi-view data using non-negative matrix factorizationand manifold learning for effective understanding: a survey paper, 201–227. Springer Nature Switzerland.

Mailier, J., Remy, M., and Wouwer, A.V. (2012). Stoichiometric identification with maximum likelihoodprincipal component analysis. *Journal of Mathematical Biology*, 67(4), 739–765.

Mysore, G.J. (2012). A block sparsity approach to multiple dictionary learningfor audio modeling. In *International Conference in Machine Learning: Workshop Sparsity, Dictionaries, Projections Machine Learning Signal Processing*.

Nolan, R.P. and Lee, K. (2011). Dynamic model of CHO cell metabolism. *Metabolic Engineering*, 13, 108–124.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 111–126.

Smaragdis, P., Févotte, C., Mysore, G.J., Mohammadiha, N., and Hoffman, M. (May 2014). Static and dynamic source separation using nonnegative factorizations: a unifed view. *IEEE Signal Processing Magazine*, 66–75.

Wang, P., Gao, L., Hu, Y., and Li, F. (2018). Feature related multi-view nonnegativematrix factorization for identifyingconserved functional modules in multiplebiological networks. *BMC Bioinformatics*, 19:394.

Wentzell, P.D., Andrews, D.T., Hamilton, D.C., Faber, K., and Kowalski, B.R. (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics*, 11, 339–366.

Yilmaz, D., Parulekar, S.J., and Cinar, A. (2020). A dynamic EFM-based model for antibody producing cell lines and modelbased evaluation of fed-batch processes. *Biochemical Engineering Journal*, 156, 107494.