Analysis of Fluxomic Experiments with Principal Metabolic Flux Mode Analysis

Sahely Bhadra, Juho Rousu

Abstract In the analysis of metabolism, two distinct and complementary approaches are frequently used: Principal component analysis (PCA) and stoichiometric flux analysis. PCA is able to capture the main modes of variability in a set of experiments and does not make many prior assumptions about the data, but does not inherently take into account the flux mode structure of metabolism. Stoichiometric flux analysis methods, such as Flux Balance Analysis (FBA) and Elementary Mode Analysis, on the other hand, are able to capture the metabolic flux modes, however, they are primarily designed for the analysis of single samples at a time, and assume the stoichiometric steady state of the metabolic network.

We will discuss a new methodology for the analysis of metabolism, called Principal Metabolic Flux Mode Analysis (PMFA), which marries the PCA and stoichiometric flux analysis approaches in an elegant regularized optimization framework. In short, the method incorporates a variance maximization objective form PCA coupled with a stoichiometric regularizer, which penalizes projections that are far from any flux modes of the network. For interpretability, we also discuss a sparse variant of PMFA that favours flux modes that contain a small number of reactions. PMFA has several benefits: (1) it can be applied to large metabolic network in efficient way as PMFA does not enumerate elementary modes, (2) The method is more robust to the steady-state violations than competing approaches, and (3) can compactly capture the variation in the data by a few factors. This chapter will describe the detailed steps how to do the above task on experimental data from fluxomic and gene expression measurements.

Key words: Principal component analysis, Metabolic Flux analysis, Sparsity

Sahely Bhadra
Indian Institute of Technology, Palakkad, e-mail: sahely@iitpkd.ac.in

Juho Rousu
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University e-mail: juho.rousu@aalto.fi

1 Introduction

In the context of transcriptomics and fluxomics, Principal Component Analysis (PCA) has been widely applied (Yao *et al.*, 2012; Barrett *et al.*, 2009), where a principal component (PC) identifies linear combinations of genes or enzymatic reactions whose activity changes explain a maximal fraction of variance within the set of samples under analysis. The main goals of PCA in fluxomic data analysis are (i) to identify which parts of the metabolism retain the main variability in flux data and (ii) to relate them to the samples, i.e, behaviour of the organism for particular experimental condition.

However, in the context of analysing metabolic networks, PCA has a few limitations (Folch-Fortuny *et al.*, 2016) as depicted in Figure 1(b): PCA considers reactions independently without considering any other structure or relationship among reactions, including stoichiometric relations implied by metabolic pathways. PCA simply extracts a set of reactions that are important to describe sample variance. Moreover, the principal components output by PCA are known to be generally dense, thus including most of the variables, which precludes their interpretation of pathways of any kind.

This chapter discusses a method called Principal Metabolic Flux Mode Analysis (PMFA) (Bhadra *et al.*, 2017), which aims to rectify the deficiencies of the PCA approach. PMFA finds metabolic flux modes that explain the variance in experiments consisting of fluxomic or gene expression data collected from heterogeneous environmental conditions, without requiring a fixed set of predefined pathways to be given. The method can be seen as a cross between Principal Component Analysis (PCA), and stoichiometrix flux analysis: It combines the variance maximization objective of PCA coupled with a stoichiometric regularizer, which penalizes projections that are far from any flux modes of the network.

The benefit of the approach for modelling and biological interpretation is that the sample variance captured by PMFA can be expressed in terms of metabolic pathways or flux modes (Figure 1(c)). Let us first briefly review the PCA and Flux Balance Analysis methods, which are frequently used to analyse data arising from metabolic systems, before describing PMFA.

1.1 Principal Component analysis

Principal component analysis (PCA) is one of the most frequently applied statistical methods in systems biology (Ma and Dai, 2011; Yao *et al.*, 2012; Barrett *et al.*, 2009). PCA is used to reduce the dimensionality of the data while retaining most of the variation in the data-set (Shlens, 2014). This reduction is done by identifying directions, i.e. linear combination of variables, called principal components, along which the variation in the data is maximal. By using a few such components, each sample can be represented by relatively few variables compared to thousands of

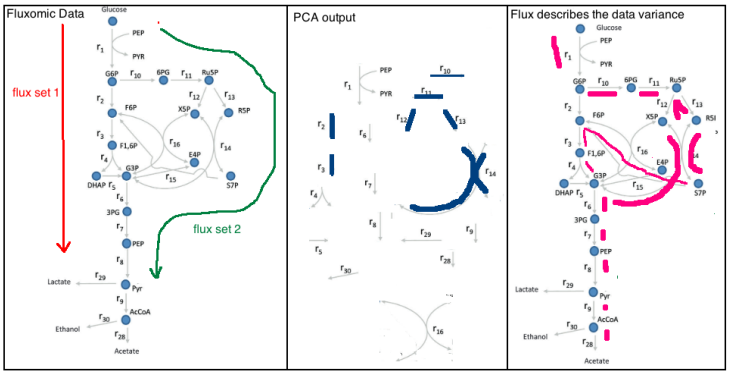

Fig. 1 While doing differential analysis of data given by (a), PCA considers reaction independently and extract the reaction which are important to describe sample variance (as shown in (b)). PMFA can be used for differential analysis of fluxomic data to extract interpretable pathways which are responsible for maximum sample variance (as shown in (c)).

features. It also helps us to distinguish between biologically relevant variables and noise.

We assume $\mathbf{X} \in \mathbb{R}^{N \times N_r}$ be the data matrix of N_r reactions in N samples, with each entry corresponding to the flux, i.e. the rate of the reaction, through a particular reaction in a particular experiment. We assume throughout the paper that all variables have been centered to have zero empirical mean. The empirical covariance matrix is then given by $\Sigma = \frac{1}{N}\mathbf{X}^T\mathbf{X}$. Denoting $\Sigma_1 = \Sigma$, the 1^{st} principal component (PC) $\mathbf{w}_1$ can be found by solving

$$\mathbf{w}_1 = \underset{w \in \mathbb{R}^{Nr}}{\arg\max}\, \mathbf{w}^T \Sigma_1 \mathbf{w}, \;\; s.t. \, \|\mathbf{w}\|_2 = 1 \tag{1}$$

Above, $\|\mathbf{w}\|_2 = \sqrt{\mathbf{w}^T\mathbf{w}}$ is the l_2 norm of the vector $\mathbf{w}$. The second PC can be found by applying Eq.(1) on updated the covariance matrix using deflation as $\Sigma_2 = (1 - \mathbf{w}_1\mathbf{w}_1^T)\Sigma_1(1 - \mathbf{w}_1\mathbf{w}_1^T)$ (Mackey, 2009).

The weights, also called the loadings, of the principal component $\mathbf{w} \in \mathbb{R}^{N_r}$ can be interpreted as the importance of reactions in explaining the variance in fluxomic data. The principal components are generally dense, containing most of the reactions of the metabolic network. Sparse PCA (Zou *et al.*, 2006) aims to increase the interpretabilty of PCA by finding principal components that have a small number of non-zero weights through solving the following optimization problem

$$\max_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w} - \gamma \|\mathbf{w}\|_1, \;\; s.t. \, \|\mathbf{w}\|_2 = 1 \tag{2}$$

where γ is a user defined hyper-parameter which controls the degree of sparsity on PC. However, the principal components extracted by neither method represent

metabolic flux modes, and will not in general adhere to the thermodynamic constraints on reaction directions.

1.2 Flux balance analysis (FBA)

Flux balance analysis (FBA) (Orth *et al.*, 2010) is a mathematical method for simulating metabolism in genome-scale reconstructions of metabolic networks. FBA is designed be used to find a flux distribution, in a stoichiometrix steady state, that maximizes a given objective (e.g. growth).

The metabolic balance of the metabolic system is described using the exchange stoichiometric matrix $\mathbf{S} \in \mathbb{R}^{N_m \times N_r}$ (Raman and Chandra, 2009) which contains transport reactions for inflow of nutrients and output flow of products, but does not contain any external metabolites (as they cannot be balanced). Rows of this matrix represent the N_m internal metabolites, columns present the N_r metabolic reactions including transport reactions and each element $\mathbf{S}_{m,r}$ shows participation of the m^{th} metabolite in the r^{th} reaction: $\mathbf{S}_{m,r} = 1$ (or -1) indicates that reaction r produces (or consumes) the metabolite m. The value $\mathbf{S}_{m,r} = 0$ indicates metabolite m is not involved in the reaction r. For a flux vector $\mathbf{w}$, $\mathbf{Sw}$ gives the change of metabolic concentration for all metabolites. The metabolic steady-state is assured by imposing a constraint $\mathbf{Sw} = 0$.

FBA solves the following optimization problem

$$\max_{\mathbf{w}} c^T \mathbf{w} \quad s.t. \, \mathbf{Sw} = 0 \text{ and } l \leq \mathbf{w} \leq u, \tag{3}$$

that calls for a finding a combination of reaction rates ($\mathbf{w}$) that adhere to stoichiometric steady state as well as upper (u) and lower bounds (l), and maximize the objective given by the combination of coefficients c and the reaction rates $\mathbf{w}$. Typically, the objective is taken as maximization of biomass production, and in this case c is equals to a row in the stoichiometric matrix corresponding to biomass production.

Simulations performed using FBA are computationally inexpensive and can calculate steady-state metabolic fluxes for large models (over 2000 reactions) in a few seconds on modern personal computers. However, as the experimental data is not directly represented in the optimization problem (3), FBA cannot be efficiently used to understand the variability between samples.

1.3 Principal Metabolic Flux Mode Analysis (PMFA)

Here we describe the Principal Metabolic Flux Mode Analysis (PMFA) approach, that combines the PCA and stoichiometric modelling views of metabolism. It finds metabolic flux modes that explain the variance in gene expression or fluxomic data

collected from heterogeneous environmental conditions without requiring a fixed set of predefined pathways to be given. Here each principal component, called *principal metabolic flux mode* (PMF), is found by selecting a set of reactions which represents a metabolic flux mode which is approximately in steady state and explains most of the data variability. In addition, we present a sparse variant, called Sparse Principal Metabolic Flux Mode analysis (SPMFA), to further help the interpretation of the principal components.

To obtain meaningful solutions of steady state flux distributions as PC loading one can impose two additional constraints in PCA formulation:

1. the weights associated with irreversible reactions should always be positive, i.e., $w_{ir} \geq 0$, where *ir* is an index of an irreversible reaction.
2. System is in a steady state, where the internal metabolite concentrations do not change, i.e. the metabolite producing and consuming fluxes cancel each other out: $\mathbf{Sw} = 0$.

Considering (1) and (2) the modified optimization problem for doing PCA with structural constraint is as following

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T \Sigma \mathbf{w} \\ s.t. \quad & \mathbf{Sw} = 0 \text{ (stoichiometric steady state)} \\ & \mathbf{w}_{ir} \geq 0 \text{ (irreversible reactions can have only positive flux)} \\ & \|\mathbf{w}\|_2 = 1 \end{aligned} \tag{4}$$

The constraint $\|\mathbf{w}\|_2 = 1$ restricts the spurious scaling up of the weights in the solution. Here, $\mathbf{Sw} = 0$ is a hard constraint and in practise imposes too much restriction, due to noise in the data, or when the data does not actually arise from steady-state conditions, e.g. given transients or perturbations of the fluxes during the experiment. Numerically, the steady state constraint amounts to a set of linear equations of size $N_M \times N_R$ which makes the problem (Eq.(4)) also computationally hard to solve. Hence instead of considering this hard constraint on the PC loadings we introduce a soft constraint which penalizes the deviation from the steady state. Our aim is to find a flux which optimizes a combination of (1) maximal explained sample variance $\mathbf{w}^T \Sigma \mathbf{w}$ and (2) minimal deviation from a steady-state condition, expressed in the l_2 norm: $\|\mathbf{Sw} - 0\|_2^2 = \|\mathbf{Sw}\|_2^2$. This entails solving the following optimization problem:

$$\boxed{\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T \Sigma \mathbf{w} - \lambda \|\mathbf{Sw}\|_2^2 \\ s.t. \quad & \mathbf{w}_{ir} \geq 0 \\ & \|\mathbf{w}\|_2 = 1 \end{aligned}} \tag{5}$$

Here λ imposes the degree of hardness of the steady-state constraint. For $\lambda = 0$ the Eq.(5) produces loadings similar to PCA with the exception of the reaction direc-

tionality constraint. The model will be henceforth denoted as PMFA$^{(l_2)}$. If desirable, we can make our model to disregard reaction directionality simply by dropping the inequality constraints $w_{ir} > 0$. By dropping the directionality constraint PMFA gives fluxes corresponding to a metabolic network where all reactions are reversible.

The l_2 norm on $\mathbf{Sw}$ in Eq.(5) has the tendency to penalize large steady state deviations in individual metabolites, at the cost of favoring small deviations in many metabolites. This is probably the desired behaviour in case the data comes from conditions where there is no subsystems that is considerably farther from steady state than other parts of the system. In order to capture the opposite scenario, where a small subset of metabolites have large deviation from steady state, one can use l_1 norm regularizer on $\mathbf{Sw}$. The l_1 norm regularizer $\|\mathbf{Sw}\|_1$ in Eq.(5) puts the emphasis of pushing most of the steady-state deviations to zero, whilst allowing a few outliers, metabolites that markedly deviate from steady state. Using l_1 regularizer and a trade-off parameter λ we get to solve the following optimization problem:

$$\begin{array}{ll} \max\limits_{\mathbf{w}} & \mathbf{w}^T \Sigma \mathbf{w} - \lambda \|\mathbf{Sw}\|_1 \\ s.t. & \mathbf{w}_{ir} \geq 0 \\ & \|\mathbf{w}\|_2 = 1 \end{array} \tag{6}$$

Here λ imposes the degree of hardness of the steady-state constraint. Similarly to Eq.(5) for $\lambda = 0$ the Eq.(6) also produces loadings similar to PCA with selective non-negative constraint. The model will be hence forth denoted as PMFA$^{(l_1)}$. Note that the solution of PMFA$^{(l_2)}$ is more stable than the solution of PMFA$^{(l_1)}$.

1.3.1 Sparse principal metabolic flux mode analysis

The above formulation of PCA with stoichiometric constraint still suffers from the fact that each principal component is a linear combination of all possible reaction activities, thus it is often difficult to interpret the results. This problem can be avoided by a variant of PMFA, the sparse principal metabolic flux mode analysis (SPMFA) using an l_1 regularizer (Tibshirani, 1996) on $\mathbf{w}$ to produce modified principal components with sparse loadings.

$$\begin{array}{ll} \max\limits_{\mathbf{w}} & \mathbf{w}^T \Sigma \mathbf{w} - \lambda \|\mathbf{Sw}\|_* \\ s.t. & \mathbf{w}_{ir} \geq 0 \\ & \|\mathbf{w}\|_1 = C \end{array} \tag{7}$$

where $\|\cdot\|_*$ can be any of the l_2 and l_1 norm and C is a used defined hyper-parameter which controls the degree of sparsity in principal metabolic flux (PMF) loadings. Similarly to PMFA, Sparse PMFA can also be made consider all reaction reversible by dropping the directionality constraints $w_{ir} \geq 0$.

1.3.2 Analysis of metabolic subsystems

One can apply our method to study differential flux modes only in a subsystem of metabolic network (e.g. central carbon metabolism, redox subsystem, lipid metabolism) by restricting the covariance matrix in objective function to the fluxes in the subsystem, while keeping the stoichiometric regularizer the same as before. Similarly, when some flux measurements are missing, one can change the covariance matrix in the objective function to exclude the fluxes that are missing.

For example, to study the variation within the redox subsystem, let $\mathbf{X}_{rdx}$ contain the columns of $\mathbf{X}$ corresponding to reactions containing redox co-factors, and let $\mathbf{w}_{rdx}$ represent the corresponding part of $\mathbf{w}$. We will consider $\Sigma_{rdx} = \frac{1}{N}\mathbf{X}_{rdx}^T\mathbf{X}_{rdx}$ for finding variance maximizing directions. Hence need to solve

$$\begin{aligned} \max_{w} \quad & \mathbf{w}_{rdx}^T \Sigma_{rdx} \mathbf{w}_{rdx} - \lambda \|\mathbf{S}\mathbf{w}\|_* \\ s.t. \quad & \mathbf{w}_{ir} \geq 0 \text{ and } \|\mathbf{w}\|_2 = 1 \end{aligned} \tag{8}$$

Similarly we can also apply SPMFA on metabolic subsystem.

2 Materials

We demonstrate the PMFA methods through two datasets: a simulation case study on *Pichia pastoris* metabolic network, and an expreimental study on *Saccharomyces cerevisiae* metabolic network. The details of the datasets are given in the following.

2.1 Datasets

Saccharomyces cerevisiae experimental case study

We use the metabolic network for *Saccharomyces cerevisiae* proposed by Hayakawa *et al.* (2015) and 13C isotopic tracer based fluxome data used in (Stosch *et al.*, 2016; Hayakawa *et al.*, 2015; Frick and Wittmann, 2005) to demonstrate the methods. The network describes the central cytosolic and mitochondrial metabolism of S. cerevisiae, comprising glycolysis, the pentose phosphate pathway, anaplerotic carboxylation, fermentative pathways, the TCA cycle, malic enzyme and anabolic reactions from intermediary metabolites into anabolism (Stosch *et al.*, 2016).

The network contains 42 compounds (30 of which are internal metabolites, which can be balanced for growth) and 47 reactions of which 39 are intracellular. The objective in this case study is to evaluate the performance of PMFA Eq.(5) on fluxome data and compare it with PEMA and PCA. For PEMA we have used 1182 EMs provided by Stosch *et al.* (2016).

This dataset is available at `https://github.com/aalto-ics-kepaco/PMFA/tree/master/Data/SaccharomycesFluxomicData.mat`. Table 1 describes its elements.

Table 1 Description of *Saccharomyces cerevisiae* fluxomic data

Matrix name	Size	Description
StoichiometricMatrix	42x47 double	Stoichiometric information matrix for all reactions
rxnE	47x7 double	Fluxomic data
metNames	42x1 cell	Name of metabolites
rxnNames	47x1 cell	Name of reaction
ExternalmetaboliteID	1x12 double	ID of extra cellular metabolites
EMs	47x1182 double	Elementary Modes
L	47x1 double	lower bound for reaction flux ($L_r = 0$ for irreversible and $L_r = -1$ for reversible reactions)

2.2 Scripts

Matlab software for PMFA and SPMFA are available in `https://github.com/aalto-ics-kepaco/PMFA`. Both PMFA and SPMFA can be applied on fluxomic and transcriptomic data.

Table 2 List of scripts required for using PMFA.

Script name	Description
CentralizedExpression.m	To centralized expression/fluxomic data
PCA.m	To find principal components (PC) of a expression/fluxomic data
SPCA.m	To find sparse PC of a expression/fluxomic data
PMFA_L2.m	To find PMF by minimizing squared norm of steady-state deviations of intracellular metabolites
PMFA_L1.m	To find PMF by minimizing l_1 norm of steady-state deviations of intracellular metabolites
SPMFA_L2.m	To find sparse PMF by minimizing squared norm of steady-state deviations of intracellular metabolites
SPMFA_L1.m	To find sparse PMF by minimizing l_1 norm of steady-state deviations of intracellular metabolites
Deflation.m	To deflate a covariance matrix of the variability explained by a PMF.

3 Finding principal flux modes

3.1 Data centralization

PCA, SPCA, PMFA, and SPMFA aim at explaining the main variability in data using a few PCs.

If the original data have non-zero mean, the first principal component typically heavily biased towards the sample mean, and fails to capture any variability between the samples.

Hence before applying any of the methods, we need to centralize the expression and fluxomic data.

Ec= CentralizedExpression(Einput,axis)

- Input :
  - Einput: Expression/fluxomic matrix
  - axis : Centralization should be done according to this axis
- output :
  - Ec: Centralized expression//fluxomic matrix

Example in Matlab for centralizing fluxomic matrix of *Saccharomyces cerevisiae* such that for every reaction the sample mean of the expression/flux is zero.

```
>> load('../Data/SaccharomycesFluxomicData.mat');
>> Ec= CentralizedExpression(saccharomyces.rxnE,2);
>> mean(Ec,2)      % this will produce a zero vector
```

3.2 Principal component analysis

Principal component analysis(PCA) as given by Eq.(1) and Sparse PCA corresponding to Eq.(2) are implemented in *PCA.m* and *SPCA.m*

function W = PCA(E,num)

- Input :
  - E: Expression/fluxomic matrix

  - num : The number of PCs to be extracted

- Output :
  - W: Each column of this matrix represents PC loadings

Example in Matlab for finding the first 3 PC loadings for *Saccharomyces cerevisiae* fluxomic data:

```
>> load('../Data/SaccharomycesFluxomicData.mat')
>> W = PCA(saccharomyces.rxnE,3)
```

function W = SPCA(Einput,gamma, num)

- Input :
  - Einput: expression/fluxomic matrix
  - gamma : User-defined parameter which indicates the degree of required sparsity in PC loadings. It corresponds to γ in Eq.(2)
  - num : The number of PCs to be extracted
- Output :
  - W: Columns of this matrix represent sparse PC loadings

Example in Matlab for finding the first 3 sparse PC loadings for *Saccharomyces cerevisiae* fluxomic data:

```
>> load('../Data/SaccharomycesFluxomicData.mat')
>> W = SPCA(saccharomyces.rxnE,1,3)
```

3.3 Finding Principal Metabolic Fluxes with PMFA

Principal Flux Mode Analysis as described in Section 1.3. is solved by the following scripts. The script *PMFA_L2.m* solves PMFA$^{(l_2)}$ with l_2 regularization on the stoichiometric constraint Eq.(5) while *PMFA_L1.m* solves PMFA$^{(l_1)}$ with l_1 regularization on stoichiometric constraint Eq.(6).

Both scripts can be used to also find principal flux modes with respect to a subsystem of metabolic network as described in Section 1.3.2. Both take reaction expression/fluxomic matrix corresponding to the defined subsystems along with a list of

indices of these reactions in the stoichiometric matrix of the whole system. For the steady state constraint both methods use the **exchange stoichiometric matrix** that contains all reactions (intra-cellular and transport reactions) in the whole metabolic network but only inter-cellular metabolites as this allows consumption and production of extra-cellular metabolites through the principal flux modes.

function [W,TotalrunTime] = PMFA_L2(Einput,S,lambda,L,U,num,ID)

- Input :
  - Einput : The expression/flux data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of samples*
  - S : Exchange stoichiometric matrix, containing all reactions in whole metabolic network but only inter-cellular metabolites. The size of this matrix is *number of metabolites* $\times$ *number of reactions*
  - lambda : User-defined regularization parameter which indicates the degree of penalization of steady-state violations in the PMF loadings. It corresponds to λ in equation Eq.(5)
  - L: Vector containing lower bounds for fluxes in the reactions
  - U: Vector containing upper bounds for fluxes in the reactions (Default = vector of all ones)
  - num: How many principal flux modes are to be computed (Default = 1)
  - ID: If we consider the analysis of a subsystem then ID contains list of index of target reactions in stoichiometrix matrix (Default = index of all reactions in the metabolic network)
- Output :
  - W: Columns of this matrix represent the PMF loadings
  - TotalrunTime: Total CPU time taken by PFMA

Example in Matlab for finding the first 3 PMF loadings for *Saccharomyces cerevisiae* fluxomic data when $\lambda = 1$:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% to find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.

>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)

% Find the first 3 PMF loadings when λ = 1
```

```
>> [W,TotalTime] = PMFA_L2(saccharomyces.rxnE, ...
    S,1,saccharomyces.L,saccharomyces.U,3)

 % Find the first 3 rev-PMF loadings when lambda = 1
 % Here we set the lower bound for all reactions at
 % negative one

 >> L = -1*ones(N,1)
 >> [W,TotalTime] = PMFA_L2(saccharomyces.rxnE, ...
    S,1,L,saccharomyces.U,3)
```

For this data set optimum value for λ is 5. For $\lambda = 5$ PMF loadings for this data is available in `https://github.com/aalto-ics-kepaco/PMFA/tree/master/SuplementaryResult/PMFsaccharoResultandAnalysis`.

	PMFA$^{(l_2)}$ $\lambda = 5$			PCA			PMFA$^{(l_2)}$ $\lambda = 7$		
Principal components	PMF1	PMF2	PMF3	PC1	PC2	PC3	PMF1	PMF2	PMF3
Fraction of sample variance	0.94	0.95	0.96	0.97	0.99	1.00	0.71	0.72	0.72
Metabolites changes ($\|Sw\|_2^2$)	0.27	0.28	0.05	0.28	0.38	1.05	0.09	0.01	0.00

Table 3 Comparing variance captured and changes of intra-cellular metabolites by PMFA$^{(l_2)}$ for optimum λ and by PCA

The total percentage of variance captured by up to 1st, 2nd and 3rd PMFs are 94.11, 94.99 and 95.76. The ℓ_2-norm of steady-state deviations in intracellular metabolites of the PMF are $0.27, 0.28$ and 0.05. Figure 3 shows the comparison of optimal PFM with PCA. With increase of λ value the resultant PMFs captured lesser variance but on the other hand they are very close to steady state fluxes.

function [W,TotalrunTime]=PMFA_L1(Einput,S,lambda,L,U,num,ID)

- Input :
  - Einput : Expression/fluxomic data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of samples*
  - S : Stoichiometric matrix with all reactions in whole metabolic network but with only intracellular metabolites. The size of this matrix is *number of metabolites* $\times$ *number of reactions*
  - lambda : User-defined regularization parameter which indicates the degree of penalization of steady-state violations in the PMF loadings. It is corresponding to λ in equation Eq.(6).
  - L: Vector containing lower bounds of fluxes in the reactions
  - U: Vector containing upper bounds of fluxes in the reactions (Default = vector of all ones)

  - num: How many principal flux modes are to be computed (Default = 1)
  - ID: If we consider the analysis of a subsystem then ID contains list of index of target reactions in stoichiometrix matrix (Default = index of all reactions in the metabolic network).

- Output :
  - W: Columns of this matrix represent the PMF loadings
  - TotalrunTime: Total cpu time taken by PFMA.

Example in Matlab for finding the first 3 PMF loadings for *Saccharomyces cerevisiae* fluxomic data when $\lambda = 1$:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% Find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.

>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)

% Find the first 3 PMF loadings when λ = 1

>> [W,TotalTime] = PMFA_L1(saccharomyces.rxnE, ...
    S,1,saccharomyces.L,saccharomyces.U,3)

 % Find the first 3 rev-PMF loadings when lambda = 1
 % Here we set the lower bound for all reactions at
 % negative one

 >> L = -1*ones(N,1)
 >> [W,TotalTime] = PMFA_L1(saccharomyces.rxnE, ...
    S,1,L,saccharomyces.U,3)
```

3.4 Finding Sparse Principal Metabolic Fluxes with SPMFA

Sparse Principal Flux Mode Analysis as described in Section 1.3.1 is solved by the following scripts. The script *SPMFA_L2.m* solves SPMFA$^{(l_2)}$ with l_2 regularization on the stoichiometric constraint while *SPMFA_L1.m* solves SPMFA$^{(l_1)}$ with l_1 regularization on stoichiometric constraint.

Similarly to PMFA, SPMFA can also find differential flux modes only in a subsystem of metabolic network as described in Section 1.3.2.

function [W,TotalrunTime] = SPMFA_L2(Einput,S,lambda,C,L,U,num,ID)

- Input :
  - Einput : The expression/fluxomic data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of samples*.
  - S : Stoichiometric matrix with all reactions in whole metabolic network but with only intracellular metabolites. The size of this matrix is *number of metabolites* $\times$ *number of reactions*
  - lambda : User-defined regularization parameter which indicates the degree of penalization of steady-state violations in the PMF loadings. It is corresponding to λ in equation Eq.(7).
  - C: The parameter controlling the sparsity; PMFs are more sparse for smaller C.
  - L: Vector containing lower bounds of fluxes in the reactions
  - U: Vector containing upper bounds of fluxes in the reactions (Default = vector of all ones)
  - num: How many principal flux modes are to be computed (Default = 1)
  - ID: If we consider the analysis of a subsystem then ID contains list of index of target reactions in stoichiometrix matrix. (Default = index of all reactions in metabolic network)
- Output :
  - W: Columns of this matrix represent the sparse PMF loadings
  - TotalrunTime: Total CPU time taken by PFMA

Example in Matlab for finding the first 3 PMF loadings for *Saccharomyces cerevisiae* fluxomic data when $\lambda = 1$:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% to find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.

>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)

% Find the first 3 PMF loadings when λ = 1 and C = 3

>> [W,TotalTime] = SPMFA_L2(saccharomyces.rxnE, ...
    S,1,3,saccharomyces.L,saccharomyces.U,3)

 % Find the first 3 rev-SPMF loadings when λ = 1
 % Here we set the lower bound for all reactions at
```

```
% negative one

>> L = -1*ones(N,1)
>> [W,TotalTime] = SPMFA_L2(saccharomyces.rxnE, ...
   S,1,3,L,saccharomyces.U,3)
```

function [W,TotalrunTime]=SPMFA_L1(Einput,S,lambda,C,L,U,num,ID)

- Input :
  - Einput : The expression/fluxomic data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of samples*.
  - S : Stoichiometric matrix with all reactions in whole metabolic network but with only intracellular metabolites. The size of this matrix is *number of metabolites* $\times$ *number of reactions*
  - lambda : User-defined regularization parameter which indicates the degree of penalization of steady-state violations in the PMF loadings. It is corresponding to λ in equation Eq.(7).
  - C: The parameter controlling the sparsity; PMFs are more sparse for smaller C.
  - L: Vector containing lower bounds of fluxes in the reactions
  - U: Vector containing upper bounds of fluxes in the reactions (Default = vector of all ones)
  - num: How many principal flux modes are to be computed (Default = 1)
  - ID: If we consider the analysis of a subsystem then ID contains list of index of target reactions in stoichiometrix matrix. (Default = index of all reactions in metabolic network)
- Output :
  - W: Columns of this matrix represent the sparse PMF loadings
  - TotalrunTime: Total cpu time taken by PFMA

Example in Matlab for finding the first 3 sparse PMF loadings for *Saccharomyces cerevisiae* fluxomic data when $\lambda = 1$ and $C = 1$:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% Find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.
```

```
>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)

% Find the first 3 PMF loadings when λ = 1

>> [W,TotalTime] = SPMFA_L1(saccharomyces.rxnE, ...
    S,1,3,saccharomyces.L,saccharomyces.U,3)

 % Find the first 3 rev-PMF loadings when lambda = 1
 % Here we set the lower bound for all reactions at
 % negative one

 >> L = -1*ones(N,1)
 >> [W,TotalTime] = SPMFA_L1(saccharomyces.rxnE, ...
    S,1,3,L,saccharomyces.U,3)
```

3.5 Deflating the Covariance matrix

To obtain a *multi-factor* PMFA model, i.e. a model containing several PMFs jointly representing the data, we follow a approach similar to some PCA algorithms, namely the deflation of the covariance matrix. However, due to additional stoichiometric constraint here we deal with a sequence of non-orthogonal vectors, $[\mathbf{w}_1,\ldots,\mathbf{w}_d]$ hence we must take care to distinguish between the variance explained by a vector and the additional variance explained, given all previous vectors. We have used orthogonal projection for deflating the data matrix (Mackey, 2009). This also maintain positive definiteness of covariance. For every iteration $d+1$ we first transfer already found principal flux modes $W \in \mathbb{R}^{N_R \times d}$ to a set of orthogonal vectors, $\{q_1,\ldots,q_d\}$.

$$q_d = \frac{(I - Q_{d-1}Q_{d-1}^T)\mathbf{w}_d}{\|(I - Q_{d-1}Q_{d-1}^T)\mathbf{w}_d\|} \tag{9}$$

where, $q_1 = \mathbf{w}_1$, and $q_1,\ldots,q_d$ form the columns of Q_d. $q_1,\ldots,q_d$ form an orthonormal basis for the space spanned by $\mathbf{w}_1,\ldots,\mathbf{w}_d$. Then the Schur complement deflation of covariance matrix is done by

$$\Sigma_{d+1} = \Sigma_d - \frac{\Sigma_d q_d q_d^T \Sigma_d}{q_d^T \Sigma_d q_d} \tag{10}$$

function [Covdef,Q] = Deflation(Cov,W)

- Input :

  - Cov : Covariance of Expression/fluxomic data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of reactions in subsystem.*
  - W: Columns of this matrix represent the sparse PMF loadings
- Output :
  - Covdef: Deflated Covariance matrix.
  - Wn: Orthogonal transformation of PMFs. It is Q in Eq.(9)

Example in Matlab for finding the first 2 PMF loadings for *Saccharomyces cerevisiae* fluxomic data using deflation of expression matrix:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% Find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.

>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)

% Find the first  PMF loadings when λ = 1

>> [W,TotalTime] = PMFA_L2(saccharomyces.rxnE, ...
    S,1,saccharomyces.L,saccharomyces.U,1)

% Data centralization

>> E=CentralizedExpression(saccharomyces.rxnE,2);

% covariance

>> CovE=E*E';

% Find Covariance matrix deflated by the first PMF

>>[Covdef,Q] = Deflation(Cov,W)
```

3.6 Computing the total variance captured by PMFs

To find the total sample variance explained by first few PMFs, we first transfer already found principal flux modes $W \in \mathbb{R}^{N_R \times d}$ to a set of orthogonal vectors, $\{q_1, \ldots, q_d\}$ using Eq.(10). Then we sum up the variance captured by $\{q_1, \ldots, q_d\}$.

The script *varianceCap.m* calculate total cumulative variance captured by upto k^{th} PFMs.

function [v] = varianceCap(E,W)

- Input :
  - E : The expression/fluxomic data for reactions in the defined subsystem. The size of this matrix is *number of reactions in subsystem* $\times$ *number of samples*.
  - W: Columns of this matrix represent the PMF loadings
- Output :
  - v: A vector where the k^{th} element shows the total fraction of sample variance captured by all PMF upto the k^{th} PMF together.

Example in Matlab for finding the variance captured by first 3 PMF loadings for *Pichia pastoris* simulation data:

```
>> load('../Data/SaccharomycesFluxomicData.mat')

% Find Stoichiometric matrix with all reactions
% in whole  metabolic network but with only
% intercellular metabolites.

>> M = size(saccharomyces.StoichiometricMatrix,1)
>> IDin = setdiff([1:1:M],saccharomyces.ExternalmetaboliteID)
>> S = saccharomyces.StoichiometricMatrix(IDin,:)


% Find the first 3  PMF loadings when λ = 1

>> [W,TotalTime] = PMFA_L2(saccharomyces.rxnE, ...
    S,1,saccharomyces.L,saccharomyces.U,3)

% Find total variance captured by PMFs

>> [v] = varianceCap(saccharomyces.rxnE,W)
```

4 Further guidelines

Directionality constraints in PMFA and SPMFA

The benefit of the directionality constraint is that the results are interpretable as flux modes with thermodynamically correct reaction directions. The directionality constraint also has been observed to increase the stability of PMFA. However, insisting on interpretability of flux modes with correct directionality may lose some power of explaining the variance. Hence dropping the directionality constraints may sometimes give further insight on the main sources of variation.

Finding mean flux modes

PMFA is similar in philosophy with the differential expression analyses where genes that vary between experiments are of interest. PMFA is not very well suitable for the analysis of a single sample at a time. If one uses the method for technical or biological replicates, the resulting flux modes will mostly capture the pattern in the noise. Also, the method is not designed to capture the main active flux modes but to capture fluxes that explain differences between different samples. However, it is easy to modify the PMFA objective so that it finds the average flux mode in a set of experiments, essentially replacing the covariance with the mean.

Analysis of non-linear trajectories

PMFs are good for explaining the main linear directions of variance, interpretable as pathways, in the samples but are not expected to fully explain complex non linear trajectories, e.g. time course data.

Finding the optimal models

The objective function of is non-convex Eq.(5), and can be interpreted as difference of two differentiable convex functions. This type of optimization problem is known as Difference of Convex functions (DC) program. We have used the convex-concave procedure (CPP), a local heuristic that utilizes the tools of convex optimization to find local optima of difference of convex functions (DC) programming problems (Lipp and Boyd, 2016). Using the CCP method we solved Eq.(5) by solving following convex approximation (quadratic program) in each iteration t:

$$\begin{aligned} \mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} \quad & \frac{\lambda}{2}\|\mathbf{S}\mathbf{w}^T\|_q - \mathbf{w}^{t^T}\Sigma_E\mathbf{w} \\ s.t. \quad & \mathbf{w}_{ir} \geq 0 \end{aligned} \tag{11}$$

followed by projecting $\mathbf{w}^{t+1}$ on $\|w\|_p = C$. The norms $p,q \in \{1,2\}$ are chosen according to the desired model.

To find a good local optimum, we repeat the above optimization with different random starting points, and take the best local minimum as the solution. In our experiments we used 100 repetitions (Rep=100).

Estimating optimal values for user-defined parameters

The performance of SPCA, PMFA and SPMFA depends on the value of used defined parameters, namely the regularization parameters λ for PMFA and SPMFA, and the level of sparsity C for SPMFA, and γ for SPCA. One should carefully chose those parameters to find correct differential fluxes.

With $\gamma = 0$, SPCA corresponds to normal PCA. With the increase of γ we increase the sparseness in PC loadings and hence increase the interpretability of it but decrease the amount of sample variance described by the PC. Hence too high value in γ is not good. Similarly, The parameter C controls the sparsity in SPMFA. Here with decrease of C the sparsity in loading increases.

The deviation from the steady-state in PMFA and SPMFA is controlled by the regularization parameter $\lambda \geq 0$: high values of λ give low deviation from steady-state and vice-versa. In particular on the fluxomic datasets, relatively heavy regularization can be applied without decrease of variance explained (cf. Figure 2). By change of the regularisation parameter λ, the statistics of PMFA exhibit a continuous transition from fully steady state flux modes ($\|\mathbf{Sw}\|_2^2 = 0$) to the PCA augmented with reaction directionality constraints.

The optimum levels of the can be set by cross-validation maximizing the *fraction of sample variance captured* on test samples

$$\text{Fraction of variance} = \frac{\mathbf{w}^T \Sigma \mathbf{w}}{\text{Trace}(\Sigma)},$$

which is a classic measure used with PCA and related approaches. Above, $\mathbf{w}$ is the PC computed from the training data, and Σ is the co-variance matrix of the test sample. Leave-One-Out (LOO) cross-validation can be used on smaller datasets and less time-intensive techniques, such as 5-fold cross-validation on larger datasets.

For *Saccharomyces cerevisiae* fluxomic data we have selected optimum parameter using LOO cross validation. Figure 2 shows the variance captured by the first PMF on training and test data for various values of λ and also corresponding $\|Sw\|_2$. Optimum λ is chosen to be 5.

PMFA on expression data

To analyze gene expression data with PMFAand SPMFA, one needs to map the gene expression to the corresponding biochemical reactions. One can transfer the expres-

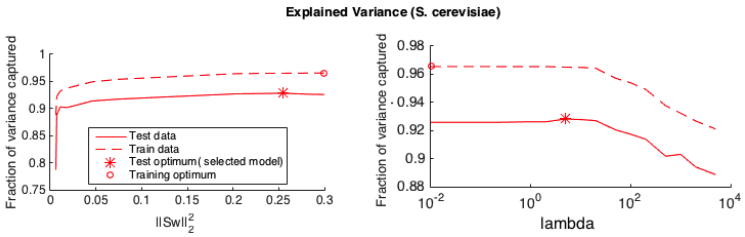

Fig. 2 Figure shows the variance captured by the first PMF on traning and test data for various values of λ and also corresponding $||Sw||_2$. Optimum λ is chosen to be 5

sion matrix from gene to reaction-wise with help of gene rules defined in metabolic network (Jensen *et al.*, 2011; Herrgård *et al.*, 2006). Gene rules are Boolean rules that determine the effect of the expression of regulatory genes on the activity of reactions in the metabolic network.

5 Conclusion

In this chapter we have demonstrated the analysis of fluxomic data with Principal Metabolic Flux Mode Analysis, PMFA. (Bhadra *et al.*, 2017). Through the combination of stoichiometric flux analysis and principal component analysis, the PMFA finds flux modes that explain most of the variation in fluxes in a set of samples. Unlike most stoichiometric modeling methods, PMFA is not tied to the steady-state assumption, but can automatically adapt—by the change of a single regularization parameter—to deviations from the stoichiometric steady-state, whether they are due to measurement errors, biological variation, or other causes. PMFA can also be applied time course and gene expression data. On the other hand, SPMFA that allows us to discover flux modes with a small fraction of reactions activated, thus could be interpreted as pathways. Thus, SPMFA is effective in the analysis large metabolic networks.

References

Barrett, C. L., Herrgard, M. J., and Palsson, B. (2009). Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC systems biology*, **3**(1), 30.

Bhadra, S., Blomberg, P., Castillo, S., and Rousu, J. (2017). Principal metabolic flux mode analysis. *bioRxiv*, page 163055.

Folch-Fortuny, A., Marques, R., Isidro, I. A., Oliveira, R., and Ferrer, A. (2016). Principal elementary mode analysis. *Molecular BioSystems*, **12**(3), 737–746.

Frick, O. and Wittmann, C. (2005). Characterization of the metabolic shift between oxidative and fermentative growth in saccharomyces cerevisiae by comparative 13 c flux analysis. *Microbial cell factories*, **4**(1), 1.

Hayakawa, K., Kajihata, S., Matsuda, F., and Shimizu, H. (2015). 13 c-metabolic flux analysis in s-adenosyl-l-methionine production by saccharomyces cerevisiae. *Journal of bioscience and bioengineering*, **120**(5), 532–538.

Herrgård, M. J., Lee, B.-S., Portnoy, V., and Palsson, B. Ø. (2006). Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in saccharomyces cerevisiae. *Genome research*, **16**(5), 627–635.

Jensen, P. A., Lutz, K. A., and Papin, J. A. (2011). Tiger: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC systems biology*, **5**(1), 147.

Lipp, T. and Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Optimization and Engineering*, **17**(2), 263–287.

Ma, S. and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics*, **12**(6), 714–722.

Mackey, L. W. (2009). Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, **28**(3), 245–248.

Raman, K. and Chandra, N. (2009). Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, **10**(4), 435–449.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Stosch, M. v., de Azevedo, C. R., Luis, M., de Azevedo, S. F., and Oliveira, R. (2016). A principal components method constrained by elementary flux modes: analysis of flux data sets. *BMC bioinformatics*, **17**(1), 200.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Yao, F., Coquery, J., and Lê Cao, K.-A. (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, **13**(1), 1.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, **15**(2), 265–286.