Journal of Computational Biology: http://mc.manuscriptcentral.com/liebert/jcb

A model-based approach to gene clustering with missing observation reconstruction in a Markov Random Field framework

Journal:	*Journal of Computational Biology*
Manuscript ID:	JCB-2008-0078.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Blanchet, Juliette; INRIA Rhone-Alpes, MISTIS; SFL Vignes, Matthieu; RRI - University of Aberdeen, BioSS
Keyword:	computational molecular biology, STATISTICS, GENE NETWORKS, GENE EXPRESSION
Abstract:	The different measurement techniques that interrogate biological systems provide means for monitoring the behaviour of virtually all cell components at different scales and from complementary angles. However data generated in these experiments are difficult to interpret. A first difficulty arises from high-dimensionality and inherent noise of such data. Organizing them into meaningful groups is then highly desirable to improve our knowledge of biological mechanisms. A more accurate picture can be obtained when accounting for dependencies between components (e.g. genes) under study. A second difficulty arises from the fact that biological experiments often produce missing values. When it is not ignored, the latter issue has been solved by imputing the expression matrix prior to applying traditional analysis methods. Although helpful, this practice can lead to unsound results. We propose in this paper a statistical methodology that integrates individual dependencies in a missing data framework. More explicitly, we present a clustering algorithm dealing with incomplete data in a Hidden Markov Random Field context.

A model-based approach to gene clustering with missing observations reconstruction in a Markov Random Field framework

Juliette Blanchet [a,1], Matthieu Vignes [b,*]

[a] *INRIA Rhône-Alpes, 655 av. de l'Europe, 38330 Saint Ismier Cedex, France*

[b] *Biomathematics and Statistics Scotland at the RRI, Bucksburn, Aberdeen, AB21 9SB, Scotland, UK*

Abstract

The different measurement techniques that interrogate biological systems provide means for monitoring the behaviour of virtually all cell components at different scales and from complementary angles. However data generated in these experiments are difficult to interpret. A first difficulty arises from high-dimensionality and inherent noise of such data. Organizing them into meaningful groups is then highly desirable to improve our knowledge of biological mechanisms. A more accurate picture can be obtained when accounting for dependencies between components (*e.g.* genes) under study. A second difficulty arises from the fact that biological experiments often produce missing values. When it is not ignored, the latter issue has been solved by imputing the expression matrix prior to applying traditional analysis methods. Although helpful, this practice can lead to unsound results.

We propose in this paper a statistical methodology that integrates individual dependencies in a missing data framework. More explicitly, we present a clustering algorithm dealing with incomplete data in a Hidden Markov Random Field context.

Preprint submitted to Journal of Computational Biology *6th November 2008*

This tackles the missing value issue in a probabilistic framework and still allows us to reconstruct missing observations *a posteriori* without imposing any pre-processing of the data. Experiments on synthetic data validate the gain in using our method and real biological data analysis present its potential to extract biological knowledge.

Key words: Biological interaction network; Gene clustering; Markov Random Field; Mean field-like approximation; Missing data.

1 Introduction

A vast continuously increasing amount of functional data is now available thanks to recent high-throughput techniques: whole-genome sequences, gene expression or localization, mass-spectrometry analysis, *etc.*. However, at present, these complex data are difficult to interpret because of such features as their high-dimensionality, their inherent noise or even bias, the absence of standardized representation. Organizing data into meaningful structures is highly desirable as a first step in unsupervised exploration of the large number of genes. Most biological mechanisms involve groupings of genes, gene products or proteins (*e.g.* enzymes) that act in a coordinated manner. Many clustering algorithms have been proposed over the last decade to decipher the message contained in DNA microarray data (see a list in the introduction of Kim et al. (2007)). In particular, Yeung et al. (2001) proposed a Gaussian mixture model to tackle this issue. This latter method and many others have the drawback to

* Corresponding author.
Email addresses: juliette.blanchet@inrialpes.fr (Juliette Blanchet), matthieu@bioss.ac.uk (Matthieu Vignes).
[1] Present address: SLF, Fluelastrasse, 11, 7260 Davos Dorf, Switzerland

consider gene measurements to be independent. Hence we proposed in a previous publication (Vignes and Forbes (2007)) an extension of this approach to account for individual features (*e.g.* microarray data) and dependencies between genes in a united framework based on *Hidden Markov Random Fields* (HMRF).

All clustering methods above use a full matrix of expression data as an input. An unfortunate feature of microarray experiments and other high-throughput technologies is that they often produce multiple missing values (McLachlan et al. (2004)). Most of the time, these missing entries appear because of various experimental issues (Troyanskaya et al. (2001), Bo et al. (2004)): dust or scratches on the slide, corrupted images, difficulties in measuring fluorescence intensity, systematic error of the robot that drops the probes, problem with precise gene spotting on the array...

A common practice –*case deletion*– is to remove genes and/or arrays from the analysis to end up with a fully-observed matrix on which classical approaches can be applied. However this approach can lose important information. Up to 90% of genes (rows) or experimental conditions (columns) can be affected (Ouyang et al. (2004)). It can also conceal interactions in a network. An alternative approach is to replace the missing values by zeros or by column/row means. Such a naive filling-in strategy is a particular case of *single imputation*. It is known to cause spurious estimation of summary statistics. The subsequent clustering results can be misleading (Little and Rubin (2002)).

Several more sophisticated methods have been proposed since the pioneering work of Troyanskaya et al. (2001). Most of them propose single imputation methods to transform the data matrix into a full matrix as needed by subse-

quent classical statistical analysis (see Troyanskaya et al. (2001), Oba et al. (2003), Bo et al. (2004), Ouyang et al. (2004)...). More recently, promising approaches make use of multiple imputation (Sehgal et al. (2005)) or iterative alternate blended clustering and missing values estimation (Kim et al. (2007)). Hu et al. (2006) proposed to improve classical estimation procedures by incorporating a large reference microarray dataset to define a general context for each gene. Nevertheless, none of these approaches take into account relationships imposed by the biological system between genes.

We propose to tackle both issues of clustering and missing data imputation in a statistical framework. To our knowledge, these two issues have never been tackled simoultaneously for dependent data. The clustering methodology has already been presented in a previous work (Blanchet and Vignes (2007)). In this paper, its efficiency is highlighted on varied datasets. The methodology is also expanded to the important issue of missing value reconstruction. Instead of imputing values prior to the analysis, our integrated approach makes the best use of the statistical framework we consider. Estimation of missing values is made *a posteriori*, based on the network, the observed individual data and the clustering pattern. We are hence able both to quantitatively compare the quality of missing data estimations and to assess the biological significance of our results in regards of approaches cited above. Given the huge amount of such algorithms, we tested algorithms reported to work well (Brock et al. (2008)) and for which we were able to retrieve the corresponding algorithms. We emphasize that our model can be useful in a great range of applications for clustering biological entities of interest such as genes, proteins, metabolites in post-genomics studies. It requires individual possibly incomplete measurements taken on these entities related by a relevant interaction network. Hence

our method is neither organism- nor data-specific.

The present paper is organized as follows: the statistical model is presented in Section 2 with the EM-like estimation procedure, the classification framework and the reconstruction of missing observations. It provides *a posteriori* probabilities of entity (*e.g.* gene) classification given the observed data. This can be seen as a confidence measure of assignment. Experiments on synthetic data are reported in Section 3 while results on real yeast cell-cycle data combined with a network of interacting proteins are presented in Section 4.

2 Markovian model for clustering and imputation with missing data

2.1 Model

In what follows, we assume that values are *Missing At Random* (MAR, Little and Rubin (2002)). The fact that a datum is missing is not related to its actual unobserved value. A particular case of MAR is when the missingness process does not depend at all on the data, as for example when a dust on the slide produces missing values. Data are then said to be *Missing Completely At Random* (MCAR, Little and Rubin (2002)). An advantage of MAR hypothesis is that maximum likelihood can be estimated independently on the missingness process (Little and Rubin (2002)). In real applications, the MAR assumption might not be true as regards the phenomenon generating missing values. Just think of censorship issues due to machine limits of detection. Data are then said to be *Not Missing At Random* (NMAR). Methods based on MAR assumption can however produce satisfactory results if observed values contain enough

information to predict missing values with a likelihood approach. Simulations in Section 3 show that inferences made by our model under MAR assumption lead to satisfactory results even on NMAR data.

We present in this section a statistical model for clustering and imputing incomplete dependent data. We refer to the entities of interest (pixels in Section 3, genes in Section 4) as *sites*, which we assume to be in interaction. These interactions can be due to spatial proximity as for the pixel images of Section 3, or due to biological relationships as for the genes of Section 4. We further assume that experiments conducted on these sites create incomplete data. We denote $\mathcal{S}$ the set of N sites and $\mathbf{x} = \{\mathbf{x}_i \in \mathbb{R}^D\}$ the $N \times D$ matrix of observations, for which some entries are missing. For each $i \in \mathcal{S}$, we write $o_i \subset [\![1, D]\!]$ the indices corresponding to the observed values x_{id} and m_i the complementary indices for missing values ($o_i \cup m_i = [\![1, D]\!]$). We shall denote $\mathbf{x}_i^{o_i} = \{x_{id}, d \in o_i\}$ the vector of observed data at site i, $\mathbf{x}_i^{m_i} = \{x_{id}, d \in m_i\}$ the vector of missing data at site i, $\mathbf{x}^o = \{\mathbf{x}_i^{o_i}, i \in \mathcal{S}\}$ the set of observed data and $\mathbf{x}^m = \{\mathbf{x}_i^{m_i}, i \in \mathcal{S}\}$ the set of missing data. We address the issue of clustering *i.e.* distinguishing meaningful groups in a dataset. In other words, each site $i \in \mathcal{S}$ has to be assigned one of the K labels $z_i \in [\![1, K]\!]$. Dependencies between sites are maintained by an interaction network defining a neighbourhood structure. The model we consider is a *Hidden Markov Random Field* (HMRF), meaning that the hidden labels (or clusters) follow a Markov Random Field distribution. This is the generalization of the one-dimensional *Hidden Markov Chains* (also referred to as Hidden Markov Models, HMM) to higher dimensions, needed to deal with a graph of interactions. In this paper, we restrict to the widely-used Potts model for which the joint Markovian (or Gibbs) distribution of labels $\mathbf{Z} = \{Z_i, i \in \mathcal{S}\}$ is:

$$P_G(\mathbf{z}) = W^{-1} \exp(\beta \sum_{i \sim j} 1_{z_i = z_j}) \quad (1)$$

where $i \sim j$ denotes neighbours sites in the network (*i.e.* linked by an edge). Note that the distribution $P_G(\mathbf{z})$ above depends on a single parameter β controlling the "smoothness" of the classification: the higher β, the more likely two neighbouring sites are to be assigned to the same cluster.

We eventually assume that data are independent conditionally on classes, this is to say that $P(\mathbf{x}|\mathbf{z}) = \prod_{i \in \mathcal{S}} P(\mathbf{x}_i|z_i)$. In this paper, class-dependent distributions are considered to be Gaussian: $P(.|Z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$.

2.2 Parameter estimation

For sake of clarity, we denote $\theta_k = (\mu_k, \Sigma_k)$ the parameters of the k-th Gaussian distribution. The single parameter of the Potts distribution is, as in equation (1), denoted by β. Without *a priori* knowledge, the full set of parameters $\Psi = (\theta_1, \cdots, \theta_K, \beta)$ is unknown and has to be estimated.

The method we propose is a maximum likelihood-based approach. The principle is to choose the most likely parameters Ψ for the observed data. Bayesian techniques would offer an alternative way to draw inference from the likelihood function. Such methods are not considered here. We use the Expectation-Maximisation (EM) algorithm (Dempster et al. (1977)). At iteration (q), a current estimate $\Psi^{(q-1)}$ is available and the algorithm maximizes the function Q defined, in a missing data framework, as:

$$Q(\Psi|\Psi^{(q-1)}) \equiv \mathbb{E}[\log P(\mathbf{x}^o, \mathbf{X}^m, \mathbf{Z}|\Psi)|\mathbf{x}^o, \Psi^{(q-1)}] \quad (2)$$

to get updated $\Psi^{(q)}$. It is worth stressing that expectation in equation (2)

is not only taken over unknown labels $\mathbf{Z}$ (as in the classical fully-observed data case), but also over missing values $\mathbf{X}^m$. An EM algorithm with incomplete data has already been studied for *Independent Mixture Model* (IMM) (see Little and Rubin (2002)), as well as for *Hidden Markov Chain Model* (see Celeux and Durand (2007) for a recent reference). To our knowledge, it has never been studied for any HMRF model. Due to the more complex dependence structure, expectation of equation (2) is not explicitely tractable for an HMRF model, as both the normalizing constant W of equation (1) and the conditional probability $P(\mathbf{z}|\mathbf{x})$ cannot be computed exactly. Approximations are then required to make the algorithm tractable.

In this paper, we propose to use a mean field-like approximation of the Markovian *a posteriori* distribution $P_G(\mathbf{z}|\mathbf{x})$ similar to the distribution proposed in Celeux et al. (2003) in the framework of complete data clustering. It was shown to be more efficient than the most widely-used clustering approaches on both simulated and real data (see Celeux et al. (2003) and Vignes and Forbes (2007)). This suggests good properties of convergence; local convergence of a very similar algorithm has been proven in Forbes and Fort (2007). The algorithm developed here extends the procedure to the missing data framework. Informally, the idea of our algorithm when considering a particular site i is to neglect the fluctuations of the neigbouring sites by setting them to fixed values $\tilde{z}_j, j \in N_i$ (means for example). The untractable Markovian distribution $P_G(\mathbf{z})$ is then approximated by the tractable factorized distribution $\prod_{i \in \mathcal{S}} P_G(z_i|\tilde{z}_{N_i})$ where $\tilde{z}_{N_i}$ denotes the set $\{\tilde{z}_j, j \in N_i\}$. Due to conditional independence, $P(\mathbf{x}, \mathbf{z})$ is also approximated as a factorized distribution and equation (2) becomes tractable. Values $\tilde{z}_i$ being *a priori* unknown, mean field-like approximations lead to an iterative EM-like algorithm repeating two steps. In what

follows, values for the $\tilde{z}_i$'s are simulated, as recommended in Celeux et al. (2003) in a complete data framework. More precisely, starting with parameters $\Psi^{(0)}$, at iteration (q),

(1) For each site $i \in \mathcal{S}$, simulate from the observed data $\mathbf{x}_i^{o_i}$ and the current parameter estimate $\Psi^{(q-1)}$ a configuration $\tilde{z}_i^{(q)}$ *i.e.* values for the Z_i's.

(2) Apply one step of the EM algorithm on the factorized model resulting from the mean field-like approximation to get updated estimates $\Psi^{(q)}$ of the parameters.

This procedure will be referred in what follows as *"SFmiss algorithm"* standing for **S**imulated **F**ield algorithm with **miss**ing values. This algorithm accounts for the Markovian structure of the data while factorizing the distribution on which EM is tractable.

In the E-step, *a posteriori* probabilities are computed for all $i \in \mathcal{S}$ and $k \in [\![1, K]\!]$ by

$$\tilde{t}_{ik}^{(q)} = P_G(Z_i = k | \mathbf{x}_i^{o_i}, \tilde{z}_{N_i}^{(q)}) \tag{3}$$

The difference with the complete data case of Celeux et al. (2003) is that conditioning in (3) involves the observed data $\mathbf{x}_i^{o_i} \in \mathbb{R}^{|o_i|}$, and not the whole vector $\mathbf{x}_i \in \mathbb{R}^D$. As in Celeux et al. (2003), the conditioning also includes neighbours through the $z_{N_i}^{(q)}$ term.

In the M-step, parameters $\Psi = (\theta_1, \cdots, \theta_K, \beta)$ are updated. The updating of the Markovian parameter β remains unchanged as compared to the complete data case (see Celeux et al. (2003)). No analytical expression is available but the optimal $\beta^{(q)}$ is unique and can easily be obtained numerically. Unlike β, the updating of the Gaussian class-dependent parameters $\theta_k = (\mu_k, \Sigma_k)$, $k \in$

$[\![1, K]\!]$, differs from the complete data case. Denote $\Sigma_k^{o_i o_i} = \{(\Sigma_k)_{st}, s \in o_i, t \in o_i\}$, $\Sigma_k^{o_i m_i} = \{(\Sigma_k)_{st}, s \in o_i, t \in m_i\} = (\Sigma_k^{m_i o_i})^T$ and $\Sigma_k^{m_i m_i} = \{(\Sigma_k)_{st}, s \in m_i, t \in m_i\}$. Then $P(X_i^{m_i}|\mathbf{x}_i^{o_i}, \theta_k)$ is a Gaussian distribution with mean η_{ik} and covariance Γ_{ik} defined as:

$$\begin{aligned} \eta_{ik} &= \mu_k^{m_i} + \Sigma_k^{m_i o_i}(\Sigma_k^{o_i o_i})^{-1}(\mathbf{x}_i^{o_i} - \mu_k^{o_i}) \\ \Gamma_{ik} &= \Sigma_k^{m_i m_i} - \Sigma_k^{m_i o_i}(\Sigma_k^{o_i o_i})^{-1}\Sigma_k^{o_i m_i}. \end{aligned} \tag{4}$$

At iteration (q) the component $s \in [\![1, D]\!]$ of μ_k is updated as:

$$(\mu_k^s)^{(q)} = \frac{\sum_i \tilde{t}_{ik}^{(q)}(r_i^s x_i^s + (1 - r_i^s)\eta_{ik}^{s\,(q)})}{\sum_i \tilde{t}_{ik}^{(q)}} \tag{5}$$

with $r_i^s = 1$ if variable x_i^s is observed, 0 otherwise. Compared with the complete data case, equation (5) simply replaces the missing variable x_i^s by the conditional mean $(\eta_{ik}^s)^{(q)}$ of the distribution $P(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)})$.

Similarly, the component $s, t \in [\![1, D]\!]$ of Σ_k is updated as:

$$(\Sigma_k^{st})^{(q)} = \frac{\sum_i t_{ik}^{(q)}(S_{ik}^{st})^{(q)}}{\sum_i t_{ik}^{(q)}}$$

with for all $i \in S$, $k \in [\![1, K]\!]$, $s, t \in [\![1, D]\!]$,

$$\begin{aligned} (S_{ik}^{st})^{(q)} &= r_i^s r_i^t (x_i^s - \mu_k^{s\,(q)})(x_i^t - \mu_k^{t\,(q)}) \\ &+ r_i^s (1 - r_i^t)(x_i^s - \mu_k^{s\,(q)})(\eta_{ik}^{t\,(q)} - \mu_k^{t\,(q)}) \\ &+ (1 - r_i^s) r_i^t (\eta_{ik}^{s\,(q)} - \mu_k^{s\,(q)})(x_i^t - \mu_k^{t\,(q)}) \\ &+ (1 - r_i^s)(1 - r_i^t)\{(\eta_{ik}^{s\,(q)} - \mu_k^{s\,(q)})(\eta_{ik}^{t\,(q)} - \mu_k^{t\,(q)}) + \Gamma_{ik}^{st\,(q)}\} \end{aligned}$$

It is worth stressing that, because of the $\Gamma_{ik}^{st\,(q)}$ term in the last factor, this is not equivalent to replacing a missing variable x_i^s by the mean $(\eta_{ik}^s)^{(q)}$ of the

conditional distribution $P(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)})$. This is consistent with the remark that mean imputation technique lowers the estimated variance (Little and Rubin (2002)).

2.3 A posteriori classification and imputation

Running q_{max} steps of the SFmiss algorithm leads to estimates $\Psi^{(q_{max})}$ of the model parameters and to configurations $\tilde{z}_i^{(q_{max})}$, $i \in \mathcal{S}$, for the mean field-like approximation. This quantities can then be used to both cluster sites and impute missing data. Due to the factorization of $P(\mathbf{z}|\mathbf{x})$ resulting from mean field-like approximation, MAP (*Maximum A Posteriori*) and MPM (*Maximum Posterior Marginal*) classification rules are equivalent and consist in classifying a site $i \in \mathcal{S}$ in :

$$\hat{z}_i = \arg\max_{k \in [\![1,K]\!]} P(Z_i = k|\mathbf{x}_i^{o_i}, \tilde{z}_{N_i}^{(q_{max})}, \beta^{(q_{max})}) = \arg\max_{k \in [\![1,K]\!]} \tilde{t}_{ik}^{(q_{max})} \tag{6}$$

Classification rule (6) involves therefore (i) the observed data $\mathbf{x}_i^{o_i} \in \mathbb{R}^{|o_i|}$ (and not the whole vector $\mathbf{x}_i \in \mathbb{R}^D$ as in the complete data case) and (ii) the neighbours through the additional $\tilde{z}_{N_i}^{(q_{max})} = \{\tilde{z}_j^{(q_{max})}, j \in N_i\}$ term. It accounts therefore explicitely for dependencies between sites. This is a clear advantage of our HMRF model over IMM.

Missing observations can also be *a posteriori* reconstructed. MAP (or MPM) rule leads to impute missing observations $\mathbf{x}_i^{m_i}$ for sites $i \in \mathcal{S}$ by the most likely values conditionally on observed $\mathbf{x}_i^{o_i}$ and on class $\hat{z}_i$:

$$\hat{\mathbf{x}}_i^{m_i} = \arg\max_{\mathbf{x}_i^{m_i}} P(\mathbf{x}_i^{m_i}|\mathbf{x}_i^{o_i}, \hat{z}_i) = \eta_{i\hat{z}_i}. \tag{7}$$

Equation (7) can be seen as a mean imputation. It differs nevertheless from

the classical pre-processed mean imputation in several points:

(i) It is performed *a posteriori*, and not as a pre-processing. Therefore it does not artificially biaised the parameter estimation (in particular the Σ_k's, see Little and Rubin (2002)) required for classification.

(ii) Relationships between sites are taken into account though the classification $\hat{z}_i$ which, as seen previously (equation (6)), involves the neighborhood structure.

(iii) The mean is not computed over all sites, but only over sites belonging to the same cluster and therefore sharing information (related biological functions for example, as in Section 4).

3 Illustration on synthetic data

The purpose of this section is to illustrate the differences between our method and standard imputation methods, and to emphasize the general aspects of the former with respect to the latter, for both classification and imputation issues. From among several exercises we have performed, we present here some results related to a 4-class synthetic image. Data were obtained as follows. Starting from the synthetic image, a noisy image is generated by considering that observations belonging to the k-th class (for $k = 1, ..., 4$) are realizations of a 4-dimensional ($D = 4$) non diagonal Gaussian distribution, with mean $\mu_k = (k, k, k, k)^T$ and covariance matrix Σ_k with diagonal terms equal to 0.5 and non diagonal terms to 0.2. We then consider two ways of producing missing data. The first one removes randomly a given proportion of data (MCAR case). The second one removes a given proportion of the highest and the lowest data (left and right censorship, NMAR case).

The classification results obtained respectively by our method and by the Markovian Simulated-Field algorithm (referred to as SF, Celeux et al. (2003)) with various prior imputations are shown in Figure 1. Imputation techniques considered are filling in with zeros (ZERO+SF), with column means (MEAN+SF), or using standard imputation methods such as K-Nearest Neighbors (Troyanskaya et al. (2001), KNN+SF), Bayesian Principal Component Analysis (Oba et al. (2003), BPCA+SF) or Support Vector Regression (Wang et al. (2006), SVR+SF). The Local Least Square Impute method (Kim et al. (2005)) gave poor results on our data and are not reported here. It appears that the SFmiss algorithm performs better than tested imputation methods, although the underlying model is the same: an HMRF model. To assess the gain in using a Markovian model, we also compare with the IMM, with parameters estimated by the EM algorithm with incomplete data (refered to as EMmiss, see Little and Rubin (2002)).

[Figure 1 about here.]

[Figure 2 about here.]

As compared with IMM, it appears that taking dependencies into account (through the use of Markovian models) improves the results significantly. Furthermore, SFmiss algorithm provides a way of modelling uncertainty over missing observation values, leading to better classifications and imputations. It can also be noted that our algorithm performs well even when the correct underlying model is not set as an hypothesis: the synthetic image is not a realization of a Potts model and censored data are NMAR! The censored data case seems to be more difficult than MCAR case but SFmiss provides reasonably good classifications for high percentages of missing values (up to 60%, see Figure 1

and visualization on Figure 2).

As mentioned in Section 2.3, in addition to providing a classification, our algorithm has the ability to reconstruct -or impute- missing data. Figure 3 displays imputation errors for the methods mentioned above. These errors are measured by the normalized Root Mean Squared Error (RMSE): if $\hat{\mathbf{x}}$ is the imputed data matrix, *i.e.*, an estimate of the complete data matrix $\mathbf{x}$, the RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\text{mean}\{(\mathbf{x} - \hat{\mathbf{x}})^2\}}{\text{mean}\{\mathbf{x}^2\}}},$$

where $\mathbf{x}^2$, for example, is component-wise.

[Figure 3 about here.]

Results presented in Figure 3 confirm that SFmiss offers a substantial improvement as concerns imputation issue. Note that, as average measures of error, RMSE for KNN, BPCA and SVR imputation techniques are similar although corresponding imputed component-wise values can be quite different, leading to different classification as reported in Figure 1.

4 Experiments on yeast cell-cycle data

4.1 Individual Data

Data of Spellman et al. (1998) on *Saccharomyces cerevisiae* that focuses on the identification of cell-cycle regulated genes were used. These data are expression profiles from yeast cultures synchronized by different methods. The full data set consists of a 77 dimensional vector for each of the 6179 genes. The initial

dataset has 5% overall missing entries. In the following, we will focus on the *cdc*28 experiment initially performed by Cho et al. (1998) (Spellman et al. (1998) used this data along with their own for their analysis). Yeast were synchronized by stopping them in late G1 phase of the cell cycle. 17 time points (dimensions) were collected every 10 minutes so nearly two cell cycles have occurred.

4.2 Interaction Network

Available biological networks contain a significant amount of information that should not be ignored to provide optimal statistical analysis of the machinery of the cell. Our aim is to build a graph with a biological entity (gene) at each node. An edge will stand for a confirmed link between two entities: interactions between genes, gene products, complexes of proteins, families, metabolic pathways,...

For network data, we use the release 7 of STRING (von Mering et al. (2007)), a consistent database of known and predicted protein-protein interactions. It gathers information from a wide variety of different sources: genomic context, literature knowledge, physical interactions, *etc.* The current version contains 401 948 curated interactions for 5611 genes of *Saccharomyces cerevisiæ*. Note that two or more interactions may occur between the same couple of genes, because different kinds of interactions are considered. We selected the intersection between the set of 800 genes identified as cell-cycle regulated in Spellman et al. (1998) and those contained in the STRING database. The resulting graph consists of 612 nodes (genes) and 3530 edges accounting for one or more interaction(s), which are given equal weight (see Section 5 for a discussion on

this aspect).

Unlike with synthetic data, the correct number of clusters is unknown. Bayes Information Criterion (BIC, Schwarz (1978)), a penalized likelihood that accounts for the complexity of the model, is widely used to tackle this issue. In our HMRF setting, BIC is untractable and we used a mean-field approximation as described in Forbes and Peyrard (2003). We allowed the number of clusters to range from 2 to 12. $K = 9$ was the selected number of clusters (data not shown).

4.3 Results and discussion

We performed imputation and 9-class clustering of the yeast data using SFmiss and several other algorithms for comparison. This section aims at assessing the quality of the produced clusters, a difficult task as there is no consensus criterion to rely on. We illustrate the gain in using our approach on some specific biological features.

We first check whether the output clusters of our model are well-suited to summarize biological knowledge compared to other algorithms. A general trend is that the SFmiss algorithm gathers interacting genes better than other algorithms. More precisely, genes clustered together by SFmiss have more internal connections than those clustered together by other imputation methods (although they rely on SF algorithm that takes the network into account). It reveals that the way SFmiss deals with missing observations is certainly more appropriate. This is consistent with the spatial parameter β of our model: β is estimated to 0.41, which means that the neighbourhood plays a significant

role.

The clusters reliability can be quantified using *Gene Ontology* (GO, The Gene Ontology Consortium (2000)) terms representativeness focusing on the biological process under study: yeast cell cycle. The more GO terms present in the data set are shared by genes in the same cluster, the more *sensitive* the method is. The more the 9 clusters isolate different parts of GO, the more *specific* the method is.

[Table 1 about here.]

For each GO category, a test is performed to determine whether the category is over-represented in each cluster. Under-representation can be tested as well but its analysis is not presented here for brevity reasons. P-values in Table 1 are computed with the FDR correction of Benjamini and Hochberg (1995) which is widely used and has proven its efficiency. Very low P-value indicates that the tested GO term is over-represented in the analyzed cluster, and therefore that the algorithm successfully grouped genes sharing this biological feature. For clarity, we only compare in Table 1 our SFmiss algorithm to IMM with missing data (EMmiss, see Little and Rubin (2002)), and to HMRF model with prior KNN imputation (SF+KNN). Other tested imputation methods (mean imputation, BPCA, SVR...) did not give better results. Apart from few exceptions (as cluster $k = 8$), P-values of Table 1 suggest that SFmiss algorithm performs better at grouping genes with similar annotations than other algorithms, *i.e.* is more sensitive.

Another nice feature of our SFmiss algorithm is that it does not only summarize known biological knowledge but can give directions for putative functions on components of living organisms based on the clustering results. For

a detailed example, genes $ydl105w$, $yer111c$, $ykr077w$, $yjl196c$, $ylr212c$ and $ynl082w$ are all classified in cluster 1 by SFmiss, whereas there are dispatched in various clusters by EMmiss. But all of them are brought into play during cell cycle processes: mitotic spindle complex repair ($ylr212c$) or G1/S transition of the mitotic cycle ($yer111c$) for example. $ykr077w$ is annotated as a putative transcription activator. Our method suggests that this annotation is fully coherent and that this gene plays a key role either as a cell cycle regulator or as a regulated gene of the process.

We can also illustrate the advantage of accounting for missing values in a united fashion as compared to prior filling-in with Troyanskaya et al. (2001) KNNimpute. Genes $ybl002w$, $ygl093w$ and $ypl269w$ belong to SFmiss cluster 4 and are dispatched in various clusters by KNN+SF. Their annotations are making sense when compared with those of their cluster and confirm a possible functional description of the cluster: chromatin assembly, required for accurate chromosome segregation localized to the nuclear side of the spindle pole body and required for cytoplasmic microtubule orientation in yeast (polarization) respectively.

The interpretation of clusters when compared to temporal classes of the cell cycle (G1, S, S/G2, G2/M and M/G1, Spellman et al. (1998)) emphasizes the specificity of the SFmiss algorithm, observed to a lesser extent in clusters resulting from other algorithms. Cluster 0 is almost entirely included in Spellman et al. (1998) G2/M group and cluster 1 is in G1 just like cluster 5. Cluster 2 include genes regulated in late G2, M and early G1 phases (quite broad, certainly a reason why no specific function is highlighted in this cluster). Cluster 3 is similar but with an earlier start. Cluster 8 is focused on M-regulated genes. Cluster 4 shows its temporal peak in S phase. Lastly, cluster 6 has many genes

from early G2 to M. These interpretations are corroborated if we investigate the expression profiles for each meaningful cluster. Examples of such additional evidences are given in Figure 4. These profiles are very similar to those obtained in Figure 4 (C, D) by Cho et al. (1998) for annotated genes.

[Figure 4 about here.]

Last but not least we would like to present another major advantage of our approach: it responds with a much greater level of stability than other tested methods when the number of observed data decreases. This is illustrated in Figure 5. Additional missing values were generated under MCAR. We then compared (i) the new classification with the initial one to assess stability of the classification (Figure 5, left panel), (ii) the new imputed values with the initial ones to assess stability of the imputation using RMSE (Figure 5, right panel).

[Figure 5 about here.]

Apart from SFmiss, all algorithms show dramatical instability when the rate of added missing value increases above 6%. Note that a 11% difference with the initial classification corresponds to one cluster which is fully lost. This suggests that these algorithms have an unsatisfactory behaviour when they are facing datasets even with "as few as" nearly 10% of total missing data in the favourable MCAR case. On the contrary, the SFmiss algorithm shows an interesting stability towards the rate of missing value. Its performance impairs significantly above 25% of overall missing data which is quite acceptable as regards real encountered situations.

5 Conclusions and future work

Missing data can bring many difficulties in data analysis simply because most data analysis procedures were not designed for them. This is particularly true in the context of post-genomic data integration. Data absence is usually a nuisance, not the focus of inquiry. We presented a comprehensive integrated statistical tool for modelling individual measurements that have a network-dependant structure. We overcame the conceptual and computational challenges and demonstrated the good features of our method on both synthetic and real biological datasets.

Our results prompt further studies. It would be interesting to analyze a dataset on a whole-genome scale. We restricted our analysis to genes with prior knowledge for validation purpose. Another prospect would be to take into account the missingness mechanism to improve performances on NMAR generated data. A possibility would be to consider the missingness mechanism as a third process and to use the recent triplet Markov field model of Blanchet and Forbes (2008) that would have to be extended to the framework of incomplete observations. A final plan is to account for missing edges as we did for missing individual measurements; biological interaction data are known to be incomplete or noisy. A first step would be to consider confidence levels for interactions as their reliability vary a lot when reported by two-hybrid screening for example. This feature is being developed in our software.

Acknowledgements

The work was partially funded by the Scottish Government.

Supplementary Materials

The SpaCEM3 software used in this study is available at `http://spacem3.gforge.inria.fr/` along with datasets.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false dicovery rate-a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Blanchet,J. and Vignes,M. (2007) Combined expression data with missing values and gene interaction network analysis: a Markovian integrated approach, 366–373, in *Proceedings of the 7th IEEE BIBE*.

Blanchet,J. and Forbes,F. (2008) Triplet Markov fields for supervised classification of complex structure data, *IEEE PAMI*, to appear.

Bo,T.H., Dysvik,B. and Jonassen,I. (2004) LSimpute: accurate estimation of missing values in microarray data with least square methods, *Nucleic Acids Research* **32,** e34.

Brock,G.N., Shaffer,J.R., Blakesley,R.E., Lotz,M.J., Tseng,G.C. (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, *BMC Bioinformatics*, **9**:12.

Celeux,G., Forbes,F. and Peyrard,N. (2003) EM procedures using mean-field like approximations for Markov-model based image segmentation, *Pattern Recognition* **36,** 131–144.

Celeux,G. and Durand,J.B. (2007) Selecting hidden Markov model state number with cross-validated likelihood, *Computational Statistics*, to appear.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wod-

icka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell* **2,** 65–73.

Dempster,A.P., Laird,N.M., Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Forbes,F. and Peyrard,N. (2003) Hidden Markov random field model selection criteria based on mean field-like approximations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25,** 1089–1101.

Forbes, F. and Fort, G. (2007) Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields, *IEEE Transactions on Image Processing* **16,** 824-837.

Hu,J., Li,H., Waterman,M.S. and Zhou,X.J. (2004) Integrative missing value estimation for microarray data, *BMC Bioinformatics* **7,** 449.

Kim,H., Golub,G.H. and Park,H. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation, *Bioinformatics* **21,** 187–198.

Kim,D.W., Lee,K.Y., Lee,K.H. and Lee,D. (2007) Towards clustering of incomplete microarray data without the use of imputation *Bioinformatics*, **23,** 107–113.

Little,R.J. and Rubin,D.B. (2002) *Statistical analysis with missing data*, New-York: Wiley, second edition.

McLachlan G.J., Do K.A. and Ambroise C. (2004) *Analyzing microarray gene expression data*, New Jersey: Wiley.

Oba,S., Sato,M., Takemasa,I., Monden,M., Matsubara,K. and Ishii,S. (2003) A Bayesian Missing value estimation method, *Bioinformatics* **19,** 2088–2096.

Ouyang M., Welsh W.J., Georgopoulos P. (2004) Gaussian mixture clustering

and imputation of microarray data, *Bioinformatics* **20,** 917–923.

Schwarz G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6,** 131–134.

Sehgal,M.S., Gondal,I. and Dooley,L.S. (2005) Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data, *Bioinformatics* **21,** 2417–2423.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hydridization, *Molecular Biology of the Cell* **9,** 3273–3297.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology, *Nature Genetics* **25,** 25–29.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing values estimation methods for DNA microarrays, *Bioinformatics* **17,** 520–525.

Vignes,M. and Forbes,F. (2007) Gene clustering via integrated Markov models combining individual and pairwise features, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* in press.

von Mering C., Jensen L.J., Kuhn M., Chaffron S., Doerks T., Krüger B., Snel B., Bork P. (2007) STRING 7 - recent developments in the integration and prediction of proteins interactions, *Nucleic Acids Research* **35,** D358–62.

Wang,X., Li,A., Jiang,Z. and Feng H. (2006) Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme, *BMC Bioinformatics* **7,** 32.

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A. and Ruzzo,L. (2001) Model-based clustering and data transformations for gene expression data, *Bioinformatics* **17,** 977–987.

List of Figures

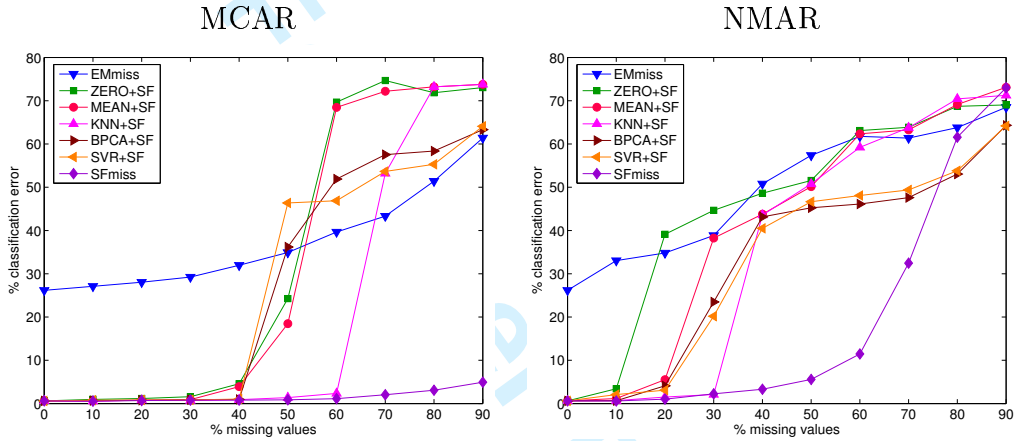


Figure 1. Experiments on image-like simulated data: percentage of misclassified pixels vs. percentage of missing data for (left) randomly missing data (MCAR case) and (right) censored data (NMAR case).

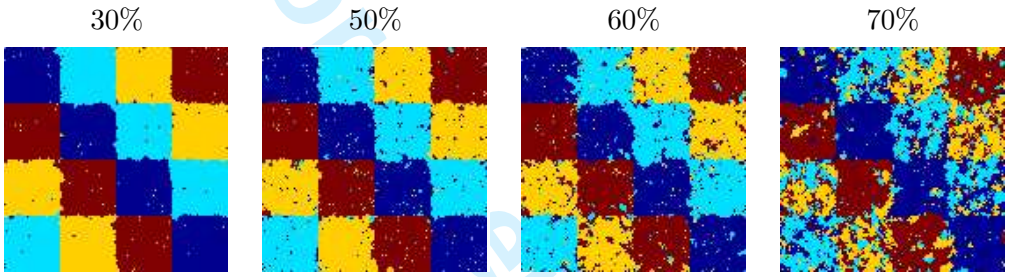


Figure 2. Experiments on image-like simulated data: visualization of the synthetic image results (*i.e.* obtained clusters) for various percent of missing data (30%, 50%, 60% and 70%,) with SFmiss in the NMAR case.

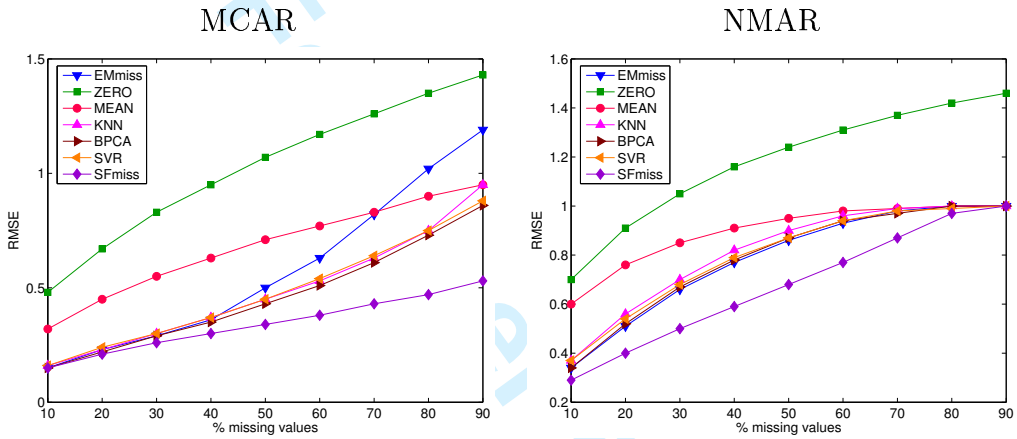
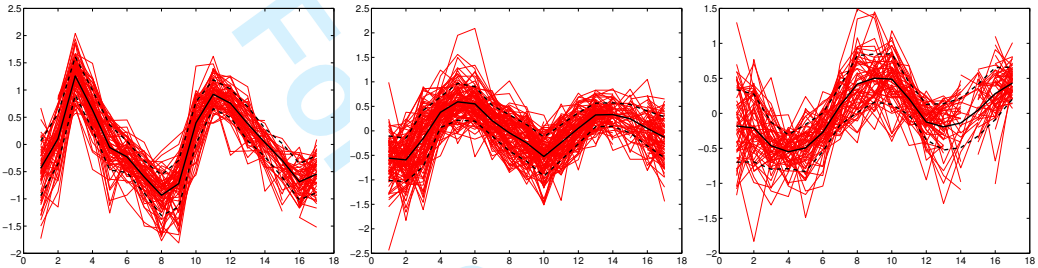


Figure 3. Experiments on image-like simulated data: RMSE vs. percentage of missing data for (left) randomly missing data (MCAR case), (right) censored data (NMAR case) on the synthetic dataset.



SFmiss cluster 1 concerned with G1 phase and DNA replication

SFmiss cluster 4: S phase, chromosome segregation and biosynthesis (*e.g.* S met. proc.)

SFmiss cluster 8 including M phase, polarization and ATP activity

Figure 4. Yeast cell-cycle data: examples of expression profiles for three SFmiss clusters (black solid line is the mean profile and dashed lines indicate standard deviation from the mean).

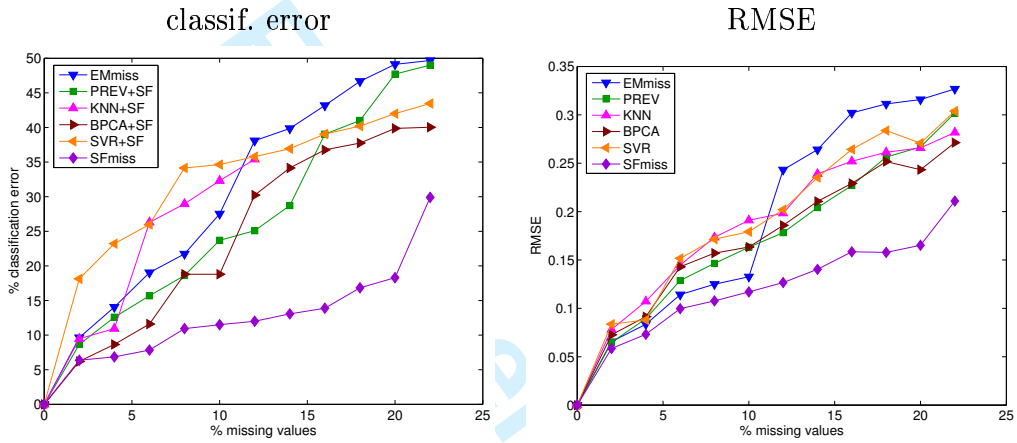


Figure 5. Yeast cell-cycle data. Left: Percentage of error for different algorithms vs. percentage of added (to the inherent approximately 5% in the dataset) missing value. Right: RMSE vs. percentage of added missing value. Algorithms are the same as in Section 3; PREV+SF has prior imputation thanks to an autoregressive model with lag 1 usually suited for time series.

List of Tables

Table 1
Yeast cell-cycle data: some representative GO terms analysis of clusters obtained by tested models and P-values of over-representation. The lowest the P-values, the more isolated the GO terms. For each GO term (row), the best method is indicated by underlined bold P-values.

cluster number	p-values of SFmiss clust.	best p-values among EMmiss clust.	best p-values among KNN+SF clust.
k=3	GO:0006732, coenzyme met. process		
	$\underline{1.1\,10^{-2}}$	> 0.1	> 0.1
k=4	GO:0005819, spindle		
	$\underline{4.6\,10^{-9}}$	$6.7\,10^{-7}$	$2.0\,10^{-6}$
	GO:0006790, sulf. met. process		
	$\underline{1.1\,10^{-4}}$	$2.4\,10^{-4}$	$8.7\,10^{-4}$
	GO:0000278, mitotic cell cycle		
	$\underline{2.2\,10^{-3}}$	$7.7\,10^{-3}$	> 0.1
	GO:0030472, mit. spin. org. & biogen. in nucleus		
	$\underline{5.2\,10^{-3}}$	$8.8\,10^{-3}$	$2.0\,10^{-2}$
k=5	GO:0006974, resp. to DNA dam. stim.		
	$\underline{1.8\,10^{-3}}$	$3.0\,10^{-3}$	$8.0\,10^{-3}$
	GO:0000724, dbl-str. bk rep. via hom. comb.		
	$\underline{1.9\,10^{-2}}$	$2.7\,10^{-2}$	$4.6\,10^{-2}$
	GO:0000030, mannosyltransf. act.		
	$\underline{1.1\,10^{-2}}$	$1.2\,10^{-2}$	$2.7\,10^{-2}$
k=8	GO:0042555, MCM cplx		
	$\underline{3.4\,10^{-4}}$	$8.3\,10^{-4}$	$4.0\,10^{-4}$
	GO:0008026, ATP-dep. helicase act.		
	$5.5\,10^{-4}$	$1.3\,10^{-3}$	**$\underline{4.5\,10^{-4}}$**
	GO:0006268, DNA unwind. replic.		
	$2.8\,10^{-3}$	$6.7\,10^{-3}$	**$\underline{1.1\,10^{-3}}$**
	GO:0042623, ATPase act. coupl.		
	$\underline{4.4\,10^{-3}}$	$1.5\,10^{-2}$	$4.3\,10^{-2}$