Bioinformatics, 33(18), 2017, 2897–2905
doi: 10.1093/bioinformatics/btx224
Advance Access Publication Date: 18 April 2017
Original Paper

OXFORD

Data and text mining

A method for learning a sparse classifier in the presence of missing data for high-dimensional biological datasets

Kristen A. Severson, Brinda Monian, J. Christopher Love and Richard D. Braatz*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 17, 2016; revised on March 16, 2017; editorial decision on April 10, 2017; accepted on April 13, 2017

Abstract

Motivation: This work addresses two common issues in building classification models for biological or medical studies: learning a sparse model, where only a subset of a large number of possible predictors is used, and training in the presence of missing data. This work focuses on supervised generative binary classification models, specifically linear discriminant analysis (LDA). The parameters are determined using an expectation maximization algorithm to both address missing data and introduce priors to promote sparsity. The proposed algorithm, expectation-maximization sparse discriminant analysis (EM-SDA), produces a sparse LDA model for datasets with and without missing data.

Results: EM-SDA is tested via simulations and case studies. In the simulations, EM-SDA is compared with nearest shrunken centroids (NSCs) and sparse discriminant analysis (SDA) with k-nearest neighbors for imputation for varying mechanism and amount of missing data. In three case studies using published biomedical data, the results are compared with NSC and SDA models with four different types of imputation, all of which are common approaches in the field. EM-SDA is more accurate and sparse than competing methods both with and without missing data in most of the experiments. Furthermore, the EM-SDA results are mostly consistent between the missing and full cases. Biological relevance of the resulting models, as quantified via a literature search, is also presented.

Availability and implementation: A Matlab implementation published under GNU GPL v.3 license is available at http://web.mit.edu/braatzgroup/links.html.

Contact: braatz@mit.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The recent '-omics' revolution in the biomedical sciences, fueled by the decreasing cost of high-throughput technologies and an increased desire for large numbers of measurements for valuable clinical samples, has led to the prevalence of wide datasets—i.e. datasets with many more measurements per sample than samples. These datasets can be generated by technologies such as microarrays and RNA-Seq, ChIP-Seq and proteomic and metabolomic techniques (e.g. mass spectrometry, multiplexed molecular assays). Such methods are gaining widespread popularity due to their potential to unearth new molecular targets for diagnosis and treatment, and due to the possibility of discovering combinations of molecular features that contribute to a disease state.

However, having many more measurements than samples leads to ill-conditioned datasets and can introduce statistical inference challenges. Two common problems arise when attempting to build models from wide datasets: the dataset is not full rank, which limits the applicable numerical approaches, and a high-dimensional model

may be difficult to interpret. Both of these issues have led to interest in learning sparse models, where the number of predictors in the final model is a subset of the training dataset.

Because of the prevalence of this problem, there are many techniques for learning a sparse model. In this work, we focus on classification models, which are of particular interest in the biomedical field due to the goal of stratifying classes of patients (e.g. healthy versus not healthy) or treatment conditions (e.g. treated versus untreated). One way to learn a sparse classification model is via the nearest shrunken centroids (NSCs) approach (Tibshirani *et al.*, 2002). This method finds a subset of predictors by penalizing, or 'shrinking', the class centroids. This technique was shown by Wang and Zhu (2007) to be equivalent to applying an ℓ_1 penalty to the class means. This approach is easy to implement and has a nice visual explanation. One limitation is that the method is required to assume a diagonal structure for the covariance matrix to avoid ill-conditioning. Another approach is sparse discriminant analysis (SDA) (Clemmensen *et al.*, 2011). This technique simultaneously performs model fitting and feature selection, finding the k discriminant vectors β_k by solving

$$\begin{aligned} &\text{minimize}_{\beta_k,\theta_k} \; ||\mathbf{Y}\theta_k - \mathbf{X}\beta_k||^2 + \gamma\beta_k\Omega\beta_k + \lambda||\beta_k||_1 \\ &\text{subject to} \;\; \frac{1}{n}\theta_k^\top\mathbf{Y}^\top\mathbf{Y}\theta_k = 1 \\ &\qquad\qquad \theta_k^\top\mathbf{Y}^\top\mathbf{Y}\theta_l = 0, \forall l < k \end{aligned}$$

where $\mathbf{Y}$ is an $n \times K$ matrix of indicator variables of the class, $\mathbf{X}$ is an $n \times p$ data matrix, Ω is a positive-definite matrix, and λ and γ are non-negative tuning parameters. This minimization is then solved iteratively.

A limitation of both of these approaches is their ability to handle missing data. Missing data are common in biological and social data. For example, technical issues may invalidate some results of an assay, a person may drop out of a longitudinal study, a hospital may only run some diagnostic tests given the time and availability of medical equipment, or respondents may skip certain questions in a social survey (García-Laencina *et al.*, 2010). In the UC Irvine Machine Learning Repository, over 20% of the datasets have missing values. Simple techniques to handle missing data involve complete case analysis, where samples with missing data are ignored, or mean imputation, where the missing data are filled in using the observed data mean. These techniques waste data and/or introduce bias.

To address these limitations, the literature contains a significant amount of work on data imputation, particularly for microarray datasets. Troyanskaya *et al.* (2001) did one of the first studies and found that k-nearest neighbors (KNNs) significantly improved on complete case analysis and mean imputation. More complex techniques have been presented by Bo *et al.* (2004), Kim and Park (2007), Kim *et al.* (2004), Oba *et al.* (2003), Ouyang *et al.* (2004), Sehgal *et al.* (2005) and Wang *et al.* (2006). Brock *et al.* (2008) surveyed these results to help practitioners decide which methods to use. The work presented here is fundamentally different than any of these techniques because it performs missing data imputation and model building simultaneously. This simultaneous approach allows for consistent assumptions in the imputation and model-building phases, and decreases the number of algorithm decisions the analyst must make. The work of Blanchet and Vignes (2009) also considers simultaneous model building and handling of missing data, but does not support a sparse model, which is a key feature of the proposed methodology. To tackle the two issues simultaneously, an expectation-maximization (EM) procedure is proposed.

The EM framework is a way to handle instances of missing data by iteratively updating the expected complete data log-likelihood and the maximum likelihood estimate of the model parameters (Dempster *et al.*, 1977). Although EM is a local optimization technique, the likelihood can only improve at each step and the method has been applied to many problems. The challenge of using EM is to choose an appropriate model for the data. In this work, we build on probabilistic principal component analysis (PPCA), a technique that uses EM to find the principal subspace, by adding sparsity-inducing priors. This method allows the learning of a subset of predictors, even in the presence of missing data. The resulting classifier is a linear discriminant analysis (LDA) model.

The proposed expectation-maximization sparse discriminant analysis (EM-SDA) algorithm addresses the intersection of these ideas to be able to tackle high-dimensional datasets that may have missing elements. The proposal is foremost meant to be able to handle expected characteristics of biological datasets, which include correlation amongst the measurements (protein, genes, etc.) and missing elements. The model assumes a symmetric distribution, which can typically be approximated via an appropriate scaling. Scaling becomes more difficult in the presence of missing data so this model is most well-suited to high-throughput assays where many measurements are performed using the same instrument and therefore the data have similar scaling. Critical care or clinical trial datasets may be more challenging to work with because of the variety of scales; however, if past information and/or intuition of scaling are available, this method would also be appropriate. As is often important in biological settings, the resulting predictions are probabilities, which are useful when more than a yes or no answer is preferred. Because the model is generative, it is also able to make predictions on new samples that also have missing elements by performing imputation.

Section 2 introduces the proposed methodology, including procedures for cases with and without missing data. Section 3 presents simulation and case studies using synthetic and real datasets. Section 4 contains discussion and conclusions.

2 Approach

2.1 Background

PCA is a widely used technique for dimensionality reduction. PCA constructs a linear projection that maximizes the variance in a lower dimensional space (Hotelling, 1933) and is also the optimal rank a approximation of a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ for $a < p$ based on the least-squares criterion (Pearson, 1901). An alternative view of PCA as a generative latent variable model (Roweis, 1998; Tipping and Bishop, 1999) is

$$\mathbf{x}_i = \mathbf{W}\mathbf{t}_i + \mu + \epsilon_i \tag{1}$$

where $\mathbf{x}_i$ is the p-dimensional observations, $\mathbf{t}_i$ is the a-dimensional latent variables, $\mathbf{W} \in \mathbb{R}^{p\times a}$ is the factor loadings, μ is a constant whose maximum likelihood estimator is the mean of the data, and ϵ_i is the error. The corresponding distributional assumptions are

$$\begin{aligned} \mathbf{t}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_a) \\ \epsilon_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_p) \\ \mathbf{x}_i|\mathbf{t}_i &\sim \mathcal{N}(\mathbf{W}\mathbf{t}_i + \mu, \sigma^2\mathbf{I}_p) \\ \mathbf{x}_i &\sim \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}_p) \end{aligned}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. The model parameters $\theta = [\mathbf{W}, \mu, \sigma^2]$ are found using an EM approach (Dempster *et al.*, 1977), which is computationally more expensive than solving the

PCA problem directly using the singular value decomposition but has the benefit of being able to handle missing data (Ilin and Raiko, 2010; Marlin, 2008; Yu *et al.*, 2012). Generally, EM is only guaranteed to converge to a local maximum of the likelihood (Dempster *et al.*, 1977); however, Tipping and Bishop (1999) show EM must converge to a global maximum for the PPCA problem.

2.2 Motivation

Let $\mathbf{x}_i \in \mathbb{R}^p$ be a vector of measurements for observations $i = 1, \ldots, n$. Let $y_i \in \{0, 1\}$ be the class label of sample i, which is observed. The classification problem is to perform supervised training to learn a model to predict the class of a new sample. To solve this problem, a LDA model is used. Because the number of samples, n, may be less than the dimension of the sample, p, the model is required to be sparse. Often when LDA is applied to datasets of this type, one of two simplifying assumptions is made: the covariance matrix has a diagonal structure, as in NSC, or a regularization penalty of the form $\lambda \mathbf{I}_p$ is added, as in SDA. The use of EM allows for a structured covariance approximation (Marlin, 2008). Under the generative latent variable model described by Equation (1), the marginal covariance is $\mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}_p$. Specifying this covariance requires estimating $pa + 1 - a(a-1)/2$ parameters: pa parameters for $\mathbf{W}$ where $a \ll p$ and 1 parameter for σ^2. The $a(a-1)/2$ term is because $\mathbf{W}$ is scaled to have orthogonal columns, each with unit length, which restricts the degrees of freedom (Bishop, 2007). An estimation of the covariance matrix in the full data space requires the estimation of $p(p+1)/2$ parameters, therefore using the latent variable model greatly decreases the number of parameters that need to be estimated. A relaxation of the diagonal matrix constraint is desirable because the data are often known to be correlated but with too few measurements to reliably estimate the full covariance. An example is gene microarray data in which genes that participate in a pathway are expected to be correlated (Witten and Tibshirani, 2011).

The LDA model makes distributional assumptions about the data, specifically

$$Y \sim \text{Binomial}(\pi)$$
$$X|Y = c \sim \mathcal{N}(\mu^c, \Sigma),$$

which is specified fully by the prior probability π, class means μ^c, and the shared covariance Σ. Here, the uninformative prior of $\pi = 0.5$ is used but the model could be extended to incorporate prior class information. The method described here learns the class means and covariance to build the classifier. The dataset is modeled in a latent space using PPCA (Roweis, 1998; Tipping and Bishop, 1999) and a sparsity-induced prior is used for the means (Figueiredo, 2003; Park and Casella, 2008).

2.3 Problem formulation

The data are assumed to be modeled as

$$\mathbf{x}_i^c = \mathbf{W}\mathbf{t}_i + \mu^c + \epsilon_i \tag{2}$$

where $\mathbf{x}_i^c, \mu^c, \epsilon \in \mathbb{R}^p$, $\mathbf{t}_i \in \mathbb{R}^a$, $\mathbf{W} \in \mathbb{R}^{p \times a}$, i represents the experiment index, and c represents the class of the observation. This PPCA formulation is typical with the small change that $\mu^c = \mu + \Delta^c$ where μ is the total mean and Δ^c is the class-specific deviation. Therefore the distributions of $\mathbf{x}$ are

$$\mathbf{x}_i|\mathbf{t}_i, y_i = c \sim \mathcal{N}\left(\mathbf{W}\mathbf{t}_i + \mu + \Delta^c, \sigma^2 \mathbf{I}_p\right)$$
$$\mathbf{x}_i|y_i = c \sim \mathcal{N}\left(\mu + \Delta^c, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}_p\right)$$

The class superscripts are dropped for convenience but the analysis assumes that all observations have class-specific means and shared covariance. A prior is set for Δ as

$$\Delta|\mathbf{T} \sim \mathcal{N}(0, \mathbf{T})$$
$$\tau_j \sim \text{Gamma}\left(1, \frac{\gamma^2}{2}\right)$$

where $\mathbf{T} = \text{diag}(\tau_j)$, which is chosen because (Figueiredo, 2003; Murphy, 2012)

$$\begin{aligned} p(\Delta_j|\gamma) &= \int_0^\infty \mathcal{N}(\Delta_j; 0, \tau_j)\text{Ga}\left(\tau_j; 1, \frac{\gamma^2}{2}\right) d\tau_j \\ &= \frac{\gamma}{2}\exp(-\gamma|\Delta_j|) = \text{Laplace}\left(\Delta_j; 0, \frac{1}{\gamma}\right) \end{aligned} \tag{3}$$

and the Laplace distribution is known to lead to sparse solutions (Murphy, 2012).

2.4 Expectation maximization

In EM, the algorithm alternates between calculating the expected complete-data log-likelihood and the maximum likelihood estimate of the parameters. For this problem, the parameters are $\theta = [\mathbf{W}, \mu, \Delta, \sigma^2]$ and the missing data are $[\mathbf{t}_i, \tau]$ (in Section 2.6, this set is augmented to include missing observations from the dataset). The observed data are $\mathbf{x}_i$ and the hyperparameter for the prior on τ, γ. The derivation of the algorithm is provided in the Supplementary Information.

To implement the result, the E-step requires the calculation of the expectations:

$$\langle \mathbf{t}_i \rangle = \left(\sigma^2 \mathbf{I}_a + \mathbf{W}^\top \mathbf{W}\right)^{-1} \mathbf{W}^\top (\mathbf{x}_i - \mu - \Delta) \tag{4a}$$

$$\langle \mathbf{t}_i \mathbf{t}_i^\top \rangle = \sigma^2 \left(\sigma^2 \mathbf{I}_a + \mathbf{W}^\top \mathbf{W}\right)^{-1} + \langle \mathbf{t}_i \rangle \langle \mathbf{t}_i \rangle^\top \tag{4b}$$

$$\left\langle \frac{1}{\tau_j} \right\rangle = \frac{\gamma}{|\Delta_j|} \tag{4c}$$

And the M-step requires the update equations:

$$\mu^{\text{new}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i - \Delta - \mathbf{W}\langle \mathbf{t}_i \rangle \tag{5a}$$

$$\Delta^{\text{new}} = \mathbf{T}\left(\sigma^2 \mathbf{I}_p + \mathbf{T}\right)^{-1} \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i - \mu - \mathbf{W}\langle \mathbf{t}_i \rangle \tag{5b}$$

$$\mathbf{W}^{\text{new}} = \left[\sum_{i=1}^n (\mathbf{x}_i - \mu - \Delta)\langle \mathbf{t}_i \rangle^\top\right]\left[\sum_{i=1}^n \langle \mathbf{t}_i \mathbf{t}_i^\top \rangle\right]^{-1} \tag{5c}$$

$$\begin{aligned} \sigma^{2\,\text{new}} = \frac{1}{dn}\sum_{i=1}^n \text{trace}\Big[&\mathbf{x}_i \mathbf{x}_i^\top - 2(\mu + \Delta)\mathbf{x}_i^\top + \mathbf{W}\langle \mathbf{t}_i \mathbf{t}_i^\top \rangle \mathbf{W}^\top \\ &+ 2(\mu + \Delta - \mathbf{x}_i)\langle \mathbf{t}_i \rangle^\top \mathbf{W}^\top + (\mu + \Delta)(\mu + \Delta)^\top\Big] \end{aligned} \tag{5d}$$

where

$$\mathbf{T} = \text{diag}(|\Delta_j|/\gamma). \tag{6}$$

The algorithm alternates between the E- and M-steps until a convergence criterion is satisfied based on the change in the negative log-likelihood (NLL) of the observed data. The change in NLL is the typical convergence criterion, but is rather expensive to calculate, which may motivate another criterion such as the change in the parameters or a fixed number of steps. Additionally, because the NLL decreases at each step, the NLL could be calculated intermittently to

reduce computational cost without risk of moving away from the optimum. Once the algorithm converges, the learned parameters are used to train the classifier.

Cross-validation should be used to select the value of the latent dimension, a, and the parameter governing sparsity, γ. Additional details on the convergence and the selection of the tuning parameters can be found in the Supplementary Information.

2.5 Classification model

LDA models are specified by two parameters, $\mathbf{w} \in \mathbb{R}^k$ and the scalar b, where k is the dimension of the vector of means whose class-specific deviations are non-zero,

$$\mathbf{w} = \widehat{\Sigma}^{-1}(\mu_1 - \mu_2), \tag{7}$$

$\widehat{\Sigma}$ is the marginal covariance of the discriminating variables which can be read from the full covariance matrix, and

$$b = -\frac{1}{2}\mu_1^\top \widehat{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2^\top \widehat{\Sigma}^{-1}\mu_2 \tag{8}$$

Predictions are then made from

$$\widehat{y} = \mathbf{w}^\top \mathbf{x}_i + b. \tag{9}$$

A value of $\widehat{y} > 0$ indicates class 1, otherwise class 2 is indicated. Its value can be converted back into a probability measure

$$P(y_i = 1|\mathbf{X} = \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i - b)} = \frac{1}{1 + \exp(-\widehat{y})} \tag{10}$$

To decrease the bias of the estimator, the EM procedure can be used for model selection, and the final model is trained without a penalty term. Whether or not this step is possible depends on data availability.

2.6 Extension to missing data

Missing data are typically described by three categories: missing completely at random, missing at random and not missing at random (Rubin, 1976). Each of these categories has a precise definition; however, robust tests do not exist to determine which mechanism is applicable to a particular scenario and instead auxiliary problem information is used to inform which model applies. Our analysis focuses on the types of missingness that we have observed in practice for high-throughput biological assays.

First, randomly missing measurements throughout the dataset, perhaps due to inappropriate sample handling or image corruption, are considered. No pattern to the missingness is assumed for this case (Fig. 1a). Second, missing measurements that are subject to a pattern are considered. Patterned missingness may represent local scratches or, as is sometimes the case in clinical settings, that some patients provided smaller samples, i.e. less volume of blood, and a rank-ordered list of assays are performed until there is no sample remaining (Fig. 1b). Finally censoring is considered, where values that meet a certain threshold are missing. An example could be species concentrations that are below a limit of detection (Fig. 1c). If censoring is a known issue, other imputation techniques, such as those that generate low or high values based on prior information, may be more appropriate as the validity of the inference is no longer guaranteed (Little and Rubin, 2002). However, the example remains relevant as the analyst may not realize that censoring is occurring.

To account for the introduction of missing data, let $\mathbf{x}_i \in \mathbb{R}^p$ be a vector of measurements for observations $i = 1, \ldots, n$ which may have elements that are missing. Any observation $\mathbf{x}_i$ can be permuted such that $\mathbf{x}_i = \left[\mathbf{x}_i^o;\ \mathbf{x}_i^m\right]$. The superscript notation denotes the elements of the ith observation which are missing (m) and observed (o). These elements are a function of the observation, i.e. $m = m(i)$, however this explicit dependence is dropped for simplicity.

The joint distribution is augmented to include these missing elements, and the derivation of the joint distribution is described in the Supplementary Information. The complete data log-likelihood does not change in this scenario but the E- and M-steps change because of the new distribution with which the expectation is taken with respect to:

$$\langle \mathbf{t}_i \rangle = \left(\sigma^2 \mathbf{I}_a + \mathbf{W}^{o\top}\mathbf{W}^o\right)^{-1}\mathbf{W}^{o\top}\left(\mathbf{x}_i^o - \mu^o - \Delta^o\right) \tag{11a}$$

$$\langle \mathbf{t}_i \mathbf{t}_i^\top \rangle = \sigma^2\left(\sigma^2 \mathbf{I}_a + \mathbf{W}^{o\top}\mathbf{W}^o\right)^{-1} + \langle \mathbf{t}_i \rangle \langle \mathbf{t}_i \rangle^\top \tag{11b}$$

$$\left\langle \frac{1}{\tau_j} \right\rangle = \frac{\gamma}{|\Delta_j|} \tag{11c}$$

$$\langle \mathbf{x}_i^m \rangle = \mathbf{W}^m \langle \mathbf{t}_i \rangle + \mu^m + \Delta^m \tag{11d}$$

$$\begin{aligned}\langle \mathbf{x}_i^m \mathbf{x}_i^{m\top} \rangle = {} & \sigma^2\left(\mathbf{I}_m + \mathbf{W}^m\left(\sigma^2 \mathbf{I}_a + \mathbf{W}^{o\top}\mathbf{W}^o\right)^{-1}\mathbf{W}^{m\top}\right) \\ & + \langle \mathbf{x}_i^m \rangle \langle \mathbf{x}_i^m \rangle^\top\end{aligned} \tag{11e}$$

$$\langle \mathbf{x}_i^m \mathbf{t}_i^\top \rangle = \sigma^2 \mathbf{W}^m\left(\sigma^2 \mathbf{I}_a + \mathbf{W}^{o\top}\mathbf{W}^o\right)^{-1} + \langle \mathbf{x}_i^m \rangle \langle \mathbf{t}_i \rangle^\top \tag{11f}$$

In the M-step, the parameters are updated by

$$\mu^{\text{new}} = \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i \rangle - \mathbf{W}\langle \mathbf{t}_i \rangle - \Delta \tag{12a}$$

$$\Delta^{\text{new}} = \mathbf{T}\left(\sigma^2 \mathbf{I}_p + \mathbf{T}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i \rangle - \mathbf{W}\langle \mathbf{t}_i \rangle - \mu \tag{12b}$$

$$\mathbf{W}^{\text{new}} = \left[\sum_{i=1}^{n}\langle \mathbf{x}_i \mathbf{t}_i^\top \rangle - (\mu + \Delta)\langle \mathbf{t}_i \rangle^\top\right]\left[\sum_{i=1}^{n}\langle \mathbf{t}_i \mathbf{t}_i^\top \rangle\right]^{-1} \tag{12c}$$

$$\begin{aligned}\sigma^{2\,\text{new}} = \frac{1}{dn}\sum_{i=1}^{n}\text{trace}\Big[& \langle \mathbf{x}_i \mathbf{x}_i^\top \rangle - 2\langle \mathbf{x}_i \mathbf{t}_i^\top \rangle \mathbf{W}^\top - 2(\mu + \Delta)\langle \mathbf{x}_i \rangle^\top \\ & + 2(\mu + \Delta)\langle \mathbf{t}_i \rangle^\top \mathbf{W}^\top + \mathbf{W}\langle \mathbf{t}_i \mathbf{t}_i^\top \rangle \mathbf{W}^\top + (\mu + \Delta)(\mu + \Delta)^\top\Big]\end{aligned} \tag{12d}$$

Note that $\langle \mathbf{x}_i \rangle$ is a concatenation of the expectations for the missing elements and the observed values. Building the final model follows the same approach as in the full data case. Re-estimation of the parameters may or may not be reasonable in this case, depending on how much data are missing. In the event of missing data in the test case, the generative model can be used to impute the relevant elements.

3 Case study

3.1 Simulation

EM-SDA is first tested by application to synthetic data. In all cases, the dataset has 100 'experiments' and 2000 'measurements' where half of the experiments are assigned to class 1 and the other half are assigned to class 0. Details concerning how the datasets were generated can be found in the Supplementary Information.

In the first application, the true model has a 2D decision boundary, so that it can be visualized easily. No missing data is added and

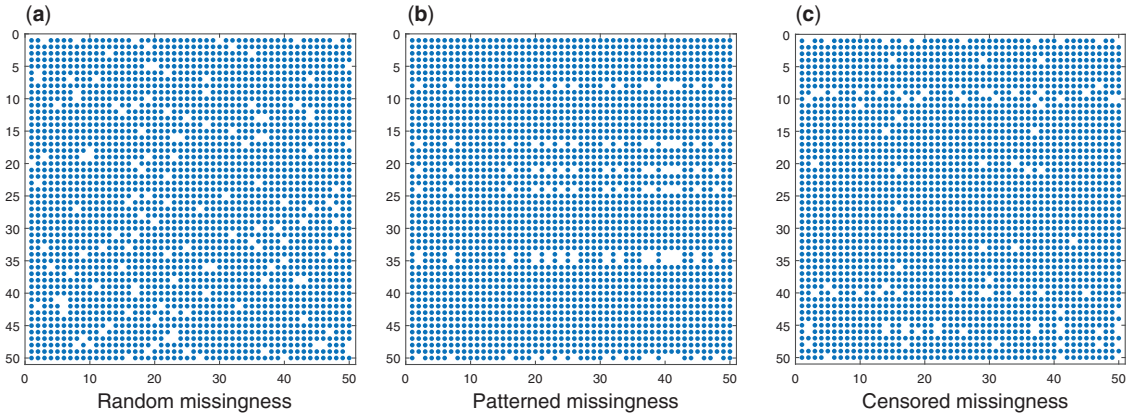

Fig. 1. Examples of the various types of missingness patterns considered from one of the simulation datasets with 5% missing data

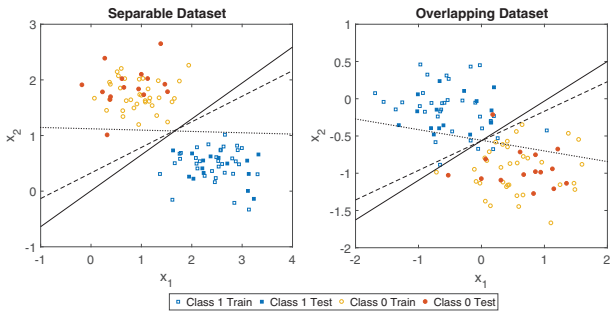

Fig. 2. The results of the 2D simulation. The true, learned and de-biased decision boundaries are the solid, dotted and dashed lines, respectively. In both cases, the correct two discriminating variables are discovered by EM-SDA (Color version of this figure is available at *Bioinformatics* online.)

two cases are considered: separable and overlapping classes. The results of learning the decision boundary using EM-SDA and 5-fold cross validation are shown in Figure 2. In both cases, EM-SDA correctly identifies the two discriminating variables from the 2000 measurements. For the separable case, NSC is unable to differentiate between a 1D and 2D model, whereas EM-SDA always correctly chooses the 2D model. SDA is less successful in finding the true discriminating variables. In the separable case, the model has three variables, one true and two spurious, and in the overlap case, the model has nine variables, two true and seven spurious. EM-SDA is better able to handle these cases where the measurements are correlated.

The second part of the simulation study focuses on the missingness mechanism (random, patterned and censored) and level (i.e. percent of data missing). Five datasets were generated and 20 discriminating variables were randomly chosen. Overlap was specified such that the true LDA model would achieve at least 95% accuracy but never 100%, i.e. the data are not separable. Missingness is introduced using random, patterned and censored assumptions into each of the cases at 5 and 15%. Cases without missing data are also tested. To train the model, 70 of the experiments are used with a 5-fold cross-validation strategy. The remaining 30 experiments are used as held-out test set. In all instances, test error refers to the model error as applied to samples not used during the training phase.

Table 1 compares the results to the NSC and SDA approaches combined with KNNs imputation. The area under the receiver operator curve (AUC) for the test data, the number of true discriminant variables, and the number of false discriminant variables selected by the model were chosen as the appropriate evaluation metrics. The best scores possible are 1, 20 and 0, respectively. In nearly all cases, the NSC method had the lowest AUC, the lowest number of true dimensions, and the fewest false dimensions. EM-SDA and SDA had similar AUC results, with EM-SDA having significantly better performance for censored data with high proportion of missing data. EM-SDA also found fewer spurious predictors than SDA in most cases.

3.2 Applications

To assess performance on real data, EM-SDA is applied to three publicly available biomedical datasets. The first, Golub *et al.* (1999), is a landmark study classifying two types of leukemia, acute myeloid leukemia and acute lymphoblastic leukemia using microarray-based gene expression data. The dataset has 72 samples, which are pre-assigned as train and test, and 7129 measurements per sample. The second dataset, Ramilo *et al.* (2007), also utilizes gene expression microarrays, but for a different application: classifying patients with acute infections of different pathogens, specifically *Escherichia coli* infection and *Staphylococcus aureus* infection. The dataset contains 59 samples, each with 211 measurements. Finally, a third dataset, Higuera *et al.* (2015) was chosen because it uses a different technology and reports missing data, unlike the first two. This dataset classifies rescued and failed learning in trisomic mice based on protein expression levels from reverse phase protein arrays. The dataset has 240 samples each with 77 protein measurements. Further details and data processing steps for all three datasets can be found in the Supplementary Information.

For the gene microarray datasets, missing data were artificially introduced. Troyanskaya *et al.* (2001) cite many possible reasons for missing data in microarrays such as insufficient resolution, image corruption, or scratches and dust on the slide. All of the presented missingness mechanisms could be applicable to microarray datasets, and therefore all three were tested. The protein expression dataset has missing data due to technical artifacts (Higuera *et al.*, 2015). 2.4% of the data is missing; however, of the 77 protein measurements, only 9 have missing data and therefore the dataset follows the patterned assumption and only the patterned mechanism was tested.

To fit the models, both the latent dimension and the value of the regularization parameter γ must be chosen. A 5-fold cross-validation strategy was used to determine the values for these hyperparameters. The values were chosen by considering the NLL of the validation set, the dimension of the final model, and the prediction

Table 1. Results for a simulation study in which EM-SDA is compared with NSC and SDA

	EM-SDA			NSC			SDA		
	0%	5%	15%	0%	5%	15%	0%	5%	15%
Test AUC									
Full	0.95 (0.04)			0.80 (0.23)			0.97 (0.01)		
Random		0.94 (0.07)	0.92 (0.08)		0.79 (0.20)	0.79 (0.23)		0.96 (0.03)	0.98 (0.01)
Patterned		0.94 (0.05)	0.97 (0.09)		0.77 (0.22)	0.77 (0.17)		0.95 (0.03)	0.95 (0.04)
Censored		0.97 (0.05)	0.91 (0.08)		0.79 (0.24)	0.77 (0.27)		0.95 (0.04)	0.73 (0.30)
Number of true dimensions found									
Full	6.8 (1.64)			6.0 (1.58)			5.4 (3.78)		
Random		6.4 (1.52)	5.8 (1.30)		4.6 (1.52)	4.6 (2.61)		5.2 (4.49)	5.4 (1.95)
Patterned		6.8 (1.10)	6.0 (2.55)		5.4 (2.88)	4.6 (1.82)		5.6 (2.70)	5.2 (3.96)
Censored		5.8 (1.64)	5.0 (1.73)		3.6 (2.30)	4.2 (2.49)		5.4 (2.79)	2.4 (0.89)
Number of false dimensions found									
Full	2.6 (3.78)			1.6 (2.07)			4.2 (3.03)		
Random		2.2 (2.59)	2.8 (1.79)		0.8 (1.10)	2.8 (5.72)		2.0 (3.03)	6.0 (5.15)
Patterned		2.2 (3.35)	3.6 (2.88)		2.0 (2.35)	0.6 (0.89)		2.6 (1.82)	2.4 (3.36)
Censored		4.2 (4.49)	4.2 (4.76)		0.4 (0.55)	5.8 (6.26)		6.0 (5.79)	9.6 (8.96)

Note: In all analyses, the SDA results are generated by running the public code, available at http://www.imm.dtu.dk/projects/spasm/ (Sjöstrand et al., 2012) and the NSC results are generated by running the public R package PAMR (Tibshirani et al., 2002). For the missing data cases, the benchmark algorithms are combined with KNN imputation. The table contains the average for the five trials and standard deviation, in parenthesis.

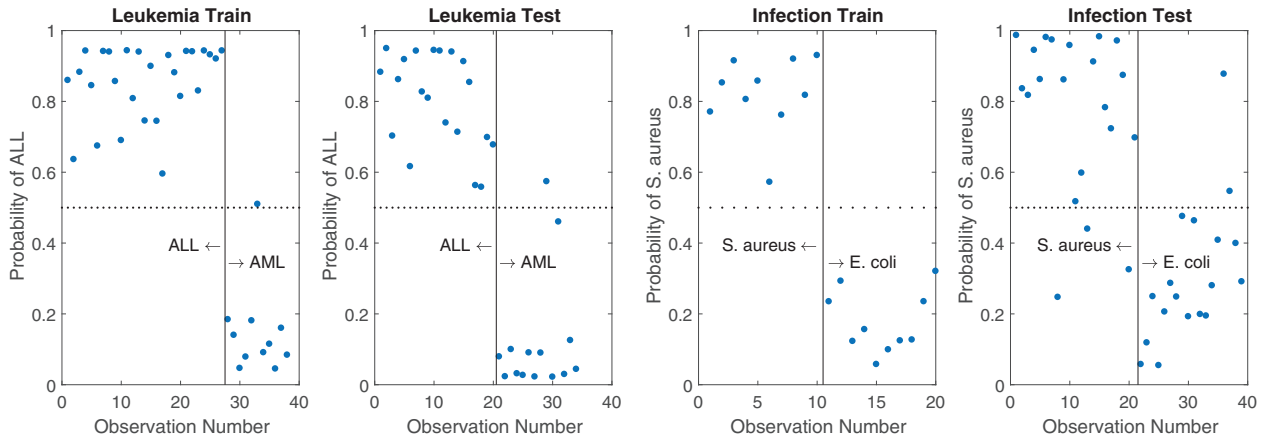

Fig. 3. The EM-SDA model predictions for the full dataset cases. The dotted line shows the decision boundary at 50%. In the leukemia problem, there is one misclassified point in each of the training and testing datasets. In the infection problem, there are zero and five misclassified points in the training and testing datasets, respectively

error. Here, a strong preference is given towards sparsity. The Supplemental Information provides additional details on the cross validation procedure.

To compare with EM-SDA, both imputation and classification algorithms must be chosen. As in the simulation study, NSC and SDA were selected as the classification algorithms. Using the results of Brock *et al.* (2008) which surveys the imputation literature for microarray data, the imputation benchmarking algorithms were chosen as KNN, Bayesian PCA (BPCA) (Oba *et al.*, 2003) and local least squares (LLS) (Kim *et al.*, 2004). Mean imputation is also included as a baseline technique. For the majority of cases considered, complete case analysis is not reasonable and is not presented here.

The model is then applied to the test data. The results for the full datasets are shown in Figure 3 and compared with the originally proposed model and the NSC and SDA approaches in Table 2.

For both leukemia and infection, EM-SDA improved the classification accuracy on the test data while using a smaller subset of genes.

Table 2. Results for the non-missing (full) cases for the leukemia and infection problems

Application	Method	Train error	Test error	Number of genes
Leukemia	Golub et al. (1999)	3/38	4/34	50
	NSC	1/38	2/34	21
	SDA	0/38	2/34	5
	EM-SDA	1/38	1/34	4
Infection	Ramilo et al. (2007)	1/20	6/39	30
	NSC	1/20	6/39	26
	SDA	0/20	11/39	9
	EM-SDA	0/20	5/39	7

Note: The results of Golub et al. (1999), Ramilo et al. (2007) and Tibshirani et al. (2002) are directly from their publications.

The results for the missing data are shown in Tables 3–5. EM-SDA provided about a factor of 10 and a factor of 2 reduction in the sum of training and testing errors for patterned missingness for

Table 3. Results for the trisomic mice classification with patterned missingness (Higuera et al., 2015)

	Patterned 2.4%			Patterned 10%		
Method	Train error	Test error	Number of proteins	Train error	Test error	Number of proteins
NSC	35/120	31/120	15	31/120	31/120	8
SDA	5/120	4/120	9	11/120	8/120	9
EM-SDA	2/120	3/120	14	6/120	5/120	12

Note: The 2.4% case is the original dataset and the 10% case demonstrates the effect of additional missing data. NSC and SDA each use KNN as the imputation technique.

Table 4. Results for the four missing data cases as compared with benchmark approaches for sparse classification for the leukemia classification problem (Golub et al., 1999)

	Random 1.5%			Random 15%			Patterned 18%			Censored 20%		
Method	Train error	Test error	Number of genes	Train error	Test error	Number of genes	Train error	Test error	Number of genes	Train error	Test error	Number of genes
NSC MI	2/38	3/34	7	1/38	3/34	6	2/38	2/34	4	2/38	6/34	9
NSC KNN	1/38	3/34	7	1/38	3/34	5	2/38	2/34	4	3/38	5/34	9
NSC LLS	2/38	2/34	7	4/38	3/34	6	1/38	2/34	5	11/38	14/34	11
NSC BPCA	1/38	2/34	8	1/38	1/34	8	2/38	5/34	5	1/38	6/34	11
SDA MI	0/38	2/34	7	0/38	4/34	7	0/38	2/34	4	19/38	21/34	5
SDA KNN	0/38	2/34	5	0/38	2/34	9	0/38	8/34	5	19/38	14/34	8
SDA LLS	0/38	2/34	7	2/38	2/34	2	2/38	2/34	2	11/38	14/34	9
SDA BPCA	0/38	2/34	6	0/38	2/34	3	0/38	7/34	4	1/38	9/34	3
EM-SDA	0/38	2/34	7	2/38	4/34	5	1/38	1/34	4	4/34	6/34	4

Table 5. Results for the four missing data cases as compared with benchmark approaches for sparse classification for the infection classification problem (Ramilo et al., 2007)

	Random 1.5%			Random 15%			Patterned 15%			Censored 11%		
Method	Train error	Test error	Number of genes	Train error	Test error	Number of genes	Train error	Test error	Number of genes	Train error	Test error	Number of genes
NSC MI	2/20	14/39	2	1/20	9/39	5	1/20	7/39	25	1/20	10/39	13
NSC KNN	2/20	13/39	2	1/20	11/39	3	2/20	15/39	1	1/20	21/39	1
NSC LLS	2/20	12/39	3	1/20	10/39	7	1/20	7/39	24	10/20	18/39	3
NSC BPCA	2/20	15/39	1	1/20	8/39	9	1/20	10/39	15	1/20	8/39	34
SDA MI	0/20	12/39	12	1/20	10/39	12	0/20	14/39	17	0/20	11/39	13
SDA KNN	1/20	14/39	5	1/20	12/39	13	0/20	12/39	15	1/20	13/39	9
SDA LLS	0/20	11/39	12	0/20	9/39	13	2/20	16/39	6	—	—	—
SDA BPCA	1/20	11/39	12	0/20	13/39	5	0/20	13/39	12	0/20	21/39	11
EM-SDA	1/20	7/39	5	2/20	8/39	4	0/20	8/39	5	1/20	10/39	5

trisomic mice, compared with NSC and SDA, respectively (Table 3). EM-SDA also outperformed the other methods for patterned missingness for leukemia (Table 4). For the other missingness patterns, EM-SDA performed similarly to the best of the other methods, often with fewer genes. In the censored case, some methods such as LLS fail to generate reasonable imputations and cannot be used in the modeling phase.

In addition to prediction accuracy, consistency and biological relevance are important to consider. Consistency is defined as the amount of gene overlap between the missing and non-missing cases for a given method. Figure 4 shows the genes that are selected and a relevance metric for the leukemia classification in EM-SDA, SDA with BPCA and NSC with BPCA. Generally, SDA has the most trouble with consistency, although the results are improved when a more advanced imputation technique (BPCA) is used. NSC and EM-SDA have similar success for consistency in the random and patterned cases. For the censored case, the problem is much more challenging. NSC and KNN perform well for the leukemia dataset but fails for the infection dataset. In both censored cases, EM-SDA does well in terms of classification error but does not recover the same set of genes. EM-SDA identifies two genes of high biological relevance—CXCR4 and MPO—for nearly all levels and types of missingness that were missed by SDA and only identified by NSC in one case. EM-SDA also had the highest average score for biological relevance, but did not find the gene with highest individual score of all identified genes, CD33. Similar figures for the infection and trisomic mice problems are available in the Supplementary Information as well as a detailed description of the score calculation.

4 Discussion

The goal of this study was to develop and evaluate a new method for simultaneous imputation and classification of high-dimensional,

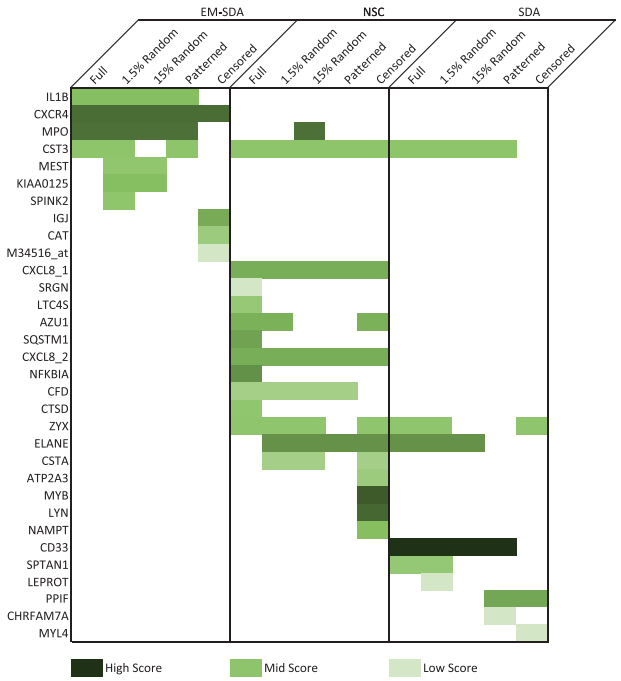

Fig. 4. Overlap and biological significance of the genes that are selected for the various leukemia classification cases. NSC and SDA are combined with BPCA for imputation. Shaded cells indicate that a particular gene was selected and the intensity of the cell represents the leukemia-relevant score based on an independent literature review

correlated data where some measurements may be missing. To achieve these goals, an expectation maximization framework was adopted. The resulting methodology, EM-SDA, was tested using both synthetic and real data for varying levels and mechanisms of missing data. EM-SDA demonstrated low classification error for sparse models in all settings and was shown to be particularly successful when the missingness is patterned.

When compared with the other methods, one advantage of EM-SDA is its ability to handle missing data. Another advantage seen in the case study is that EM-SDA found nearly the same models as if data were not missing. Its use of the structured covariance approximation avoids the non-physical assumption that different measurements are independent. Because the model is generative, it can also be used when test cases have missing elements by imputing the maximum likelihood estimate. A limitation of EM-SDA is its computational cost. For the case where no data are missing, the computational cost per iteration is $\mathcal{O}(np^2)$ and the memory storage is $\mathcal{O}(na^2)$. When data are missing, the computational cost per iteration is $\mathcal{O}(n\tilde{m}^2a)$ and the memory storage is $\mathcal{O}(np^2)$ where $\tilde{m}$ is the maximum number of elements that are missing for any sample. The increase in memory for the missing data case is due to the need to store the expected value of the outer product of the missing data. Expectation maximization is known to be slow to converge. A possible way to speed up convergence would be to use adaptive overrelaxed EM (AEM) (Salakhutdinov and Roweis, 2003). As the fraction of missing data increases, EM is known to take smaller steps, in which case AEM can lead to large speedups (Salakhutdinov and Roweis, 2003).

EM-SDA has been demonstrated to be successful for all of the types of missingness studied. EM-SDA is particularly recommended when the missingness is patterned or if missingness is likely to occur in test samples. EM-SDA is well suited to wide, correlated biological datasets, such as microarray data, RNA-Seq data, patient metadata and proteomic data. As more of these datasets are generated and subjected to rigorous statistical analyses, new models that can both systematically handle missing data and yield simple, interpretable and accurate results will become increasingly valuable.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful comments and suggestions. We would also like to thank Golub *et al.* (1999), Ramilo *et al.* (2007) and Higuera *et al.* (2015) for making their data publicly available.

Funding

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA); US Army Medical Research and Material Command (MRMC); and the Army Research Office (ARO). The views, opinions and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government [grant number N66001-13-C-4025].

Conflict of Interest: none declared.

References

Bishop,C. (2007). *Pattern Recognition and Machine Learning*. Springer, New York.

Blanchet,J. and Vignes,M. (2009) A model-based approach to gene clustering with missing observation reconstruction in a Markov random field framework. *J. Comput. Biol.*, **16**, 475–486.

Bø,T.H. *et al.* (2004) LSimpute: Accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.*, **32**, e34.

Brock,G.N. *et al.* (2008) Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes. *BMC Bioinformatics*, **9**, 12.

Clemmensen,L. *et al.* (2011) Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–38.

Figueiredo,M.A.T. (2003) Adaptive sparseness for supervised learning. *IEEE T. Pattern Anal.*, **25**, 1150–1159.

García-Laencina,P.J. *et al.* (2010) Pattern classification with missing data: a review. *Neural Comput. Appl.*, **19**, 263–282.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Higuera,C. *et al.* (2015) Self-organizing feature maps identify proteins critical to learning in a mouse model of Down syndrome. *PLoS One*, **10**, 1–28.

Hotelling,H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.

Ilin,A. and Raiko,T. (2010) Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.*, **11**, 1957–2000.

Kim,H. *et al.* (2004) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.

Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.

Little,R.J.A. and Rubin,D.B. (2002). *Statisical Analysis with Missing Data*, 2nd edn. John Wiley & Sons, New Jersey.

Marlin,B.M. (2008). *Missing Data Problems in Machine Learning*. PhD Thesis, University of Toronto.

Murphy,K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.

Oba,S. *et al.* (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Ouyang,M. *et al.* (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**, 917–923.

Park,T. and Casella,G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Pearson,K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, **2**, 559–572.

Ramilo,O. *et al.* (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood*, **109**, 2066–2077.

Roweis,S. (1998) EM algorithms for PCA and SPCA. *Adv. Neur. Inf.*, 626–632.

Rubin,D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Salakhutdinov,R. and Roweis,S. (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, DC, pp. 664–671.

Sehgal,M.S.B. *et al.* (2005) Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data. *Bioinformatics*, **21**, 2417–2423.

Sjöstrand,K. *et al.* (2012). *SpaSM: A Matlab Toolbox for Sparse Statistical Modeling*. Technical report, Technical University of Denmark.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *P. Natl. Acad. Sci. USA*, **99**, 6567–6572.

Tipping,M.E. and Bishop,C.M. (1999) Probabilistic principal component analysis. *J. Roy. Stat. Soc. B*, **61**, 611–622.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wang,S. and Zhu,J. (2007) Improved centroids estimation for the nearest strunken centroid classifier. *Bioinformatics*, **23**, 972–979.

Wang,X. *et al.* (2006) Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, **7**, 32.

Witten,D.M. and Tibshirani,R. (2011) Penalized classification using Fisher's linear discriminant. *J. Roy. Stat. Soc. B*, **73**, 753–772.

Yu,L. *et al.* (2012) Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. *J. Struct. Biol.*, **171**, 18–30.