Published in final edited form as:
Biometrics. 2014 June ; 70(2): 312–322. doi:10.1111/biom.12149.

A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation

Lin S. Chen,
Department of Health Studies, University of Chicago, 5841 S Maryland Ave, Chicago, Illinois, USA

Ross L. Prentice, and
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, Washington, USA

Pei Wang[*]
Institute of Genomics and Multiscale Biology, Icahn Medical School at Mount Sinai, 1427 Madison Ave, New York, New York, USA

Summary

Missing data rates could depend on the targeted values in many settings, including mass spectrometry-based proteomic profiling studies. Here we consider mean and covariance estimation under a multivariate Gaussian distribution with non-ignorable missingness, including scenarios in which the dimension (p) of the response vector is equal to or greater than the number (n) of independent observations. A parameter estimation procedure is developed by maximizing a class of penalized likelihood functions that entails explicit modeling of missing data probabilities. The performance of the resulting 'penalized EM algorithm incorporating missing data mechanism (PEMM)' estimation procedure is evaluated in simulation studies and in a proteomic data illustration.

Keywords

Expectation-maximization (EM) algorithm; maximum penalized likelihood estimate; not-missing-at-random (NMAR)

1. Introduction

Mass spectrometry (MS) based platforms serve as the workhorse (Faca et al., 2006) in proteomics profiling research. However, properly analyzing proteomics data from MS based experiments remains challenging due to a typical high percentage of missing data and complicated missingness patterns. For example, in our application in Section 5, the missingness rate in one sample is as high as 50%.

[*] Corresponding Author: pwang@fhcrc.org.

Supplementary Materials
Web Appendices and Figures referenced in Sections 2–5, as well as R codes used in the simulations and the application in Section 4–5 are available with this paper at the Biometrics website on Wiley Online Library.

When the proportion of missing values in a dataset is substantial, it is inappropriate to simply ignore the observations with missing values (Rubin, 1976; Little and Rubin, 2002). Various statistical approaches have been proposed for valid inference based on incomplete data (Afifi and Elashoff, 1966; Dempster et al., 1977; Rubin, 1987, 1996; Schafer, 1997). A crucial step involves characterizing the nature of missingness in a study. To do so, Rubin (1976) defined three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Inference based on approaches that ignore the missing data mechanisms could still be valid (Little and Rubin, 2002) for MCAR or MAR, but not for NMAR data.

To properly characterize the missing data mechanism in proteomics data, we first need to obtain a good understanding of the experimental procedures and the related instrument measurement properties. A typical MS based proteomics experiment starts with enzymatically digesting intact proteins into peptides — small segments of amino acid sequences. Then, peptides are introduced into the MS instrument for identification and quantification. In the end, the presence and abundance of proteins are inferred based on the identification and quantification of peptides. Due to the dynamic nature of the MS instrument, many factors in the experiment could contribute to missingness in the final protein abundance data. For example, the MS machine may have trouble detecting weak signals of low-abundance peptides. Or, even if the instrument detects the signal, the peak intensities of low-abundance peptides may be too low to be distinguished from background noise during data processing. Therefore, the lower the abundance, the more likely the peptide will be “missing” in the MS output data. Besides abundance, other physical or chemical properties of a protein could also contribute to its missingness. For instance, a small protein consisting of only a few peptides in general is more likely to be “missing” in MS experiments than a large protein consisting of many peptides. So the probability of a protein being missing also depends on the number of peptides in the protein (or roughly, the size of the protein). Moreover, missing data could be caused by other experimental attributes associated with each protein/peptide. In light of these observations, modeling the missing data mechanism in proteomics data as a function of the abundance to be measured, as well as other pertinent variables, provides an attractive approach to acknowledging the major aspects of these experimental complexities. Note, the abundance-dependent missing data mechanism in proteomics data involves no clear detection threshold, and thus it is more appropriate to utilize a probabilistic missing data mechanism than a censoring model (Little and Rubin, 2002).

In this paper, we focus on the problem of jointly estimating the mean abundance levels of multiple proteins and their covariances, i.e., estimating mean and covariance matrix for incomplete data with abundance-dependent missingness when $p < n$ or $p \geq n$. Addressing this problem could greatly facilitate high dimensional omics data analysis, such as pathway/gene-set based hypothesis testing or discriminant analysis. Here, we propose to model the abundance-dependent missing data mechanism with a class of exponential functions and seek parameter estimates that maximize the joint likelihood of the observed data and the missing events, assuming that the multiple protein abundance levels follow a multivariate Gaussian distribution. In addition, to deal with the high-dimension-low-sample-size issue, we employ a penalized likelihood approach. We impose a Inverse-Wishart penalty on the covariance matrix, which amounts to the conjugate prior for a multivariate Gaussian distribution from a Bayesian perspective. This penalty gives a simple closed-form solution for the maximum penalized likelihood estimates and is computationally efficient.

Since samples are often missing different proteins and thus have different likelihood functions, there is typically no closed-form solution for maximizing the joint log-likelihood function. In order to obtain the maximum likelihood estimates with incomplete data, the EM algorithm and its extensions (Dempster et al., 1977; Meng and Rubin, 1993; Neal and Hinton, 1999) have been widely used in a variety of applications. The EM algorithm gained its popularity because of its easy implementation and its numerical stability due to monotone convergence (Wu, 1983). However, it may converge slowly and may not converge to the global maximum (McLachlan and Krishnan, 1996). To improve its numerical performance and provide more efficient estimation of the parameters of interest, regularization has been introduced to the EM framework. Green (1990) proposed maximizing a penalized likelihood function in the M-step and demonstrated that the corresponding EM algorithm achieves a faster convergence rate. In contrast, Schneider (2001) introduced the penalty to the E-step in each iteration and allowed the penalty to change across iterations. However, it is unclear whether typical convergence properties of the EM are preserved in this procedure. In recent work (Städler and Bühlmann, 2012; Städler et al., 2012), sparse penalties have been employed in the EM framework to handle high-dimensional multivariate data when missingness is MAR, with the goal of controlling the sparsity of the inverse covariance matrix. For non-ignorable missing data, Little and Rubin (2002, Chapter 15.2) have outlined a framework for the EM algorithm to incorporate general NMAR mechanisms. In comparison, the major innovation of this paper is that we introduce a specific NMAR mechanism into the general penalized EM framework. We propose PEMM, a penalized EM algorithm incorporating missing data mechanism, for data with non-ignorable missingness. We implement the proposed PEMM algorithm for parameter estimation of multivariate Gaussian data with abundance-dependent missingness.

The remainder of this paper is organized as follows. In Section 2, we present the penalized joint likelihood model that incorporates missing data mechanism. In Section 3, we outline the PEMM algorithm and implement it in detail for the models proposed in Section 2. We compare the proposed method with competing ones on simulated data and a proteomics dataset in Section 4 and 5, respectively. We then provide a brief summary in Section 6.

2. Model

Let $\mathbf{X} = (x_{ij})$ denote the complete Gaussian data without missing values, where x_{ij} represents the j^{th} feature for the i^{th} subject. Let $\mathbf{O}_i$ denote the index set of features being observed in the i^{th} sample; and let $\mathbf{X}_{i,obs} = \{x_{ij} : j \in \mathbf{O}_i\}$ and $\mathbf{X}_{i,mis} = \{x_{ij} : j \notin \mathbf{O}_i\}$ represent the observed and the missing component in the i^{th} sample respectively. Let $\mathbf{M} = (m_{ij})$ be the missingness indicator matrix, such that $m_{ij} = 1$ if x_{ij} is missing, and $m_{ij} = 0$ if x_{ij} is observed. The missing data mechanism is characterized by $P(\mathbf{M}|\mathbf{X})$. If $P(\mathbf{M}|\mathbf{X})$ depends on the missing values in $\mathbf{X}$, the mechanism is NMAR, and ignoring the missing data mechanism leads to invalid parameter inference. Therefore, in this paper, we will explicitly model $P(\mathbf{M}|\mathbf{X})$ and estimate the mean and variance of **X**, $\boldsymbol{\mu}$ and Σ, by seeking the MLEs of the full likelihood, which is the joint likelihood of the observed data $\mathbf{X}_{obs}$ and the missing-indicator matrix **M**:

$$(\hat{\boldsymbol{\mu}}, \hat{\sum}) = \arg\max_{\boldsymbol{\mu}, \sum} \sum_{i=1}^{n} \log f(\mathbf{X}_{i,obs}, \mathbf{M}_i; \boldsymbol{\mu}, \sum).$$

2.1 Missing data mechanism in proteomics data

It is typically reasonable to assume that, given the complete abundance data **X** and the pertinent corresponding covariate data **C**, missingness of different features are independent of each other:

$$P(\mathbf{M}|\mathbf{X}, \mathbf{C}) = \prod_{i,j} P(m_{ij}=1|\mathbf{X}, \mathbf{C}).$$

We also assume that for a given feature, its missingness does not depend on abundances or covariates of other features:

$$P(m_{ij}=1|\mathbf{X}, \mathbf{C}) = P(m_{ij}=1|x_{ij}, \mathbf{c}_{ij}),$$

where $\mathbf{c}_{ij}=\{c_{ij}^{k}\}_{k}$ is a vector of covariates associated with the j^{th} feature for the i^{th} subject. For example, c_{ij}^{k} can be the size (# peptides) of the j^{th} protein, which will take the same value across different subjects; or it can be the total detected ion abundance at the corresponding eluting time of the j^{th} protein (output from the MS1 data) in the i^{th} experiment. We then propose to model $P(m_{ij} = 1|x_{ij}, \mathbf{c}_{ij})$ with a bounded exponential function:

$$g(x_{ij}; \mathbf{c}_{ij}, \boldsymbol{\Gamma}) = \min(\exp\{-\gamma_1 - \gamma_2 x_{ij} - \boldsymbol{\gamma}_3^T \mathbf{c}_{ij}\}, 1), \quad (1)$$

where $\Gamma = \{\gamma_1, \gamma_2, \gamma_3\}$ is the parameter of the missing data mechanism and is distinct from the parameter of interest ($\boldsymbol{\mu}$, Σ). For positive γ_2, this probability function monotonically decreases with the abundance, x_{ij}, to be measured and is consistent with the abundance-dependent missing data mechanism discussed above.

Note, for some platforms utilizing labelling strategies, the outputs from the experiments are the log-ratios of abundances in the test samples versus the reference samples, and log-ratios with smaller absolute values are more likely to be missing. This missing data mechanism can be modelled using a bounded quadratic exponential form:

$$g(x_{ij}; \mathbf{c}_{ij}, \boldsymbol{\Gamma}) = \min(\exp\{-\gamma_1 - \gamma_2 x_{ij}^2 - \boldsymbol{\gamma}_3^T \mathbf{c}_{ij}\}, 1). \quad (2)$$

Estimation procedure development for log-ratio-abundance data using the above missing data mechanism is provided in Web Appendix F. Here, we choose to use x_{ij}^2 instead of $|x_{ij}|$ in (2) for computational convenience.

2.2 Penalized joint likelihood

Denote the joint log-likelihood of the observed data and missing-indicator matrix as $L(\mathbf{X}_{obs}, \mathbf{M}; \boldsymbol{\mu}, \Sigma, \Gamma, \mathbf{C})$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance parameters of interest; $\mathbf{C}$ is the observed covariate data; and Γ is the nuisance missing data mechanism parameter. Our goal is to obtain the MLE:

$$(\hat{\boldsymbol{\mu}}, \hat{\sum}) = \arg\max_{\boldsymbol{\mu}, \sum} L(\mathbf{X}_{obs}, \mathbf{M}; \boldsymbol{\mu}, \sum, \boldsymbol{\Gamma}, \mathbf{C}). \quad (3)$$

Temporarily we assume that Γ is completely known. We will elaborate estimation with Γ parameters known and unknown in Sections 3.2.1 and 3.2.2, respectively.

When $p > n$ or $p \approx n$, the MLE in equation (3) is typically unsatisfactory or even not meaningfully defined. For example, if there is no missingness, the MLE for $(\boldsymbol{\mu}, \Sigma)$ of a multivariate Gaussian distribution is the sample mean $\bar{\mathbf{X}}$ and the sample covariance matrix $\mathbf{S}_{\mathbf{X}}$. When $p \geq n$, $\mathbf{S}_{\mathbf{X}}$ becomes singular and is no longer a proper estimator for Σ.

To circumvent these difficulties, regularization is quite valuable. A natural strategy is to consider the penalized likelihood and seek the *maximum penalized likelihood estimate* (MPLE):

$$(\hat{\boldsymbol{\mu}}, \hat{\sum}) = \arg\max_{\boldsymbol{\mu}, \sum} \{L(\mathbf{X}_{obs}, \mathbf{M}; \boldsymbol{\mu}, \sum, \boldsymbol{\Gamma}, \mathbf{C}) - P(\sum)\}. \quad (4)$$

where $P(\cdot)$ is a penalty function on Σ. The performance of the MPLE heavily depends on the choice of the penalty term $P(\cdot)$. It is important to take both the interpretability and computational feasibility into account when specifying $P(\cdot)$. In recent literature on high dimensional Gaussian graphical modeling, including Yuan and Lin (2006), Friedman et al. (2008), Rothman et al. (2008), and Städler and Bühlmann (2012), the l_1 norm of the concentration matrix ($\|\Sigma^{-1}\|_{l_1}$) has typically been penalized, thereby helping to control the number of non-zero entries in the MPLE of the concentration matrix. In other applications involving covariance estimation, such as pathway analysis in proteomics studies (Chen et al., 2011), it may not be necessary or reasonable to assume a sparse Σ^{-1}. In this paper, we consider an alternative approach and propose to use a penalty amounting to an Inverse-Wishart prior with penalty parameters λ and K:

$$P(\sum) = \lambda Tr(\sum^{-1}) + K \log|\sum|. \quad (5)$$

Denote the eigenvalues of Σ as $\{d_\ell\}_{\ell=1}^{p}$. The above penalty term can be rewritten as: $P(\Sigma) = \lambda$ &Sum;$_\ell$ $1/d_\ell + K \log(\Pi_\ell d_\ell)$. Thus positive (λ, K) values constrain the parameter space of Σ and bounds each eigenvalue from above and below. This helps to ensure that both covariance matrix and concentration matrix are non-singular, an important property that is not readily achieved if $p > n$. In addition, since the Inverse-Wishart distribution is the conjugate prior for the covariance matrix of multivariate Gaussian, the PEMM algorithm for solving (4), which is outlined in the next section, enjoys computational efficiency. More discussion on the interpretation of λ and K is provided in later sections.

The Inverse-Wishart penalty equally penalizes the variance of each of the p-variables. If some of the variables have substantially larger variances than the others, one could standardize each variable first, or incorporate different weights in the penalty term for different variables.

3. Algorithm

In Section 3.1, we outline a PEMM algorithm for calculating MPLEs of general distributions and investigate its convergence properties. Then, in Section 3.2, we implement the PEMM algorithm for obtaining the MPLEs of multivariate Gaussian parameters with Inverse-Wishart penalty on the covariance matrix and abundance-dependent (non-ignorable) missing data mechanism.

3.1 A PEMM algorithm

Consider the general problem of seeking MPLE for data with non-ignorable missingness:

$$\hat{\boldsymbol{\Omega}}=\arg\max_{\boldsymbol{\Omega}} \tilde{L}(\mathbf{X}_{obs},\mathbf{M};\boldsymbol{\Omega})=\arg\max_{\boldsymbol{\Omega}}\{L(\mathbf{X}_{obs},\mathbf{M},\boldsymbol{\Omega})-P(\boldsymbol{\Omega})\},$$

where Ω is the parameter of interest. Note that covariate **C** is omitted to simplify the presentation in this section. To solve the above optimization problem, we outline PEMM, a modified version of the EM algorithm, in Algorithm 1 below.

Algorithm 1

A PEMM algorithm.

```
1  Obtain the initial estimate $\Omega^{(0)}$.
2  E-step: calculate $L^{(\hat{t}-1)}(\mathbf{X}, \mathbf{M}; \Omega) = E_{\mathbf{X}_{mis}|\mathbf{X}_{obs}\mathbf{M};\Omega^{(t-1)}} \{L(\mathbf{X}, \mathbf{M}; \Omega)\}$;
3  M-step: calculate $\boldsymbol{\Omega}^{(t)}=\arg\max_{\boldsymbol{\Omega}} \left\{\hat{L}^{(t-1)}(\mathbf{X},\mathbf{M};\boldsymbol{\Omega})-P(\boldsymbol{\Omega})\right\}$.
4  Repeat 2–3 until convergence.
```

Following Beale and Little (1975), we establish the convergence property of the PEMM algorithm in Web Appendix A. Briefly, the targeted penalized likelihood will always increase in successive iterations of the PEMM algorithm, and thus the algorithm will converge to a stationary point of the penalized log-likelihood (though not necessarily to the global maximum, similar to the EM algorithm). Therefore, the PEMM algorithm is an appropriate algorithm for solving the optimization problem in equation (6).

3.2 A PEMM algorithm for estimating multivariate Gaussian parameters

In this section, we implement the PEMM algorithm in detail to calculate the MPLEs of multivariate Gaussian parameters with the missing data mechanism specified in equation (1) and the penalty term specified in equation (5). To better illustrate the impact of the missing data mechanism on parameter estimation, we first review a penalized EM algorithm for data with ignorable missingness, and then present the PEMM algorithm for data with abundance-dependent missingness. When data are MAR, one can ignore the missing data mechanisms in parameter estimation (Rubin, 1976):

$$(\hat{\boldsymbol{\mu}},\hat{\sum})=\arg\max_{\mu,\sum}\{L(\mathbf{X}_{obs},\mathbf{M};\boldsymbol{\mu},\sum,\boldsymbol{\Gamma},\mathbf{C})-P(\sum)\}=\arg\max_{\mu,\sum}\{\mathbf{L}(\mathbf{X_{obs}};\boldsymbol{\mu},\sum)-\mathbf{P}(\sum)\}.$$

Then, to obtain the MPLEs of multivariate Gaussian parameters with Inverse-Wishart penalty (5) on the covariance matrix, we can implement a penalized EM algorithm as outlined in Algorithm 1. Specifically, the E-step is the same as described in Beale and Little (1975):

$$\hat{\mathbf{X}}_{i,obs}^{(t)}=\mathbf{X}_{i,obs}, \quad \hat{\mathbf{X}}_{i,mis}^{(t)}=\boldsymbol{\mu}_{i,mis}^{(t-1)}+\sum\nolimits_{i:mis,obs}^{(t-1)}\left(\sum\nolimits_{i:mis,obs}^{(t-1)}\right)^{-1}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}_{i,obs}^{(t-1)}); \\ \mathbf{A}_i^{(t)}=\mathbf{0}, \quad \mathbf{A}_{i:mis,mis}^{(t)}=\sum\nolimits_{i:mis,mis}^{(t-1)}-\sum\nolimits_{i:mis,obs}^{(t-1)}\left(\sum\nolimits_{i:obs,obs}^{(t-1)}\right)^{-1}\sum\nolimits_{i:obs,mis}^{(t-1)},$$

where $\mathbf{A}_i^{(t)}$ is a $p \times p$ matrix and $\mathbf{A}_{i:mis,mis}^{(t)}$ is the submatrix of $\mathbf{A}_i^{(t)}$ that corresponds to the missing features in the i^{th} sample. Here, $\mathbf{A}_{i:mis,mis}^{(t)}$ represents $\mathrm{cov}(\mathbf{X}_{i,mis}|\mathbf{X}_{i,obs}; \boldsymbol{\mu}^{(t-1)}, \Sigma^{(t-1)})$, which measures the additional covariance caused by the missing data in sample i.

Then, with the Inverse-Wishart penalty in equation (5), the MPLE updates in the M-step become:

$$\boldsymbol{\mu}^{(t)}=n^{-1}\sum_i\hat{\mathbf{X}}_i^{(t)}, \quad \sum\nolimits^{(t)}=(n-K)^{-1}\left(n\mathbf{S}^{(t)}+\lambda\mathbf{I}\right); \tag{6}$$

where $\mathbf{S}^{(t)}=n^{-1}\sum_i\left((\hat{\mathbf{X}}_i^{(t)}-\boldsymbol{\mu}^{(t)})(\hat{\mathbf{X}}_i^{(t)}-\boldsymbol{\mu}^{(t)})^T+\mathbf{A}_i^{(t)}\right)$. The regularization induced by the Inverse-Wishart penalty helps to assure the positive definiteness of the estimated covariance matrix. For the special case of $K = 0$, the corresponding update in the M-step simplifies to $\Sigma^{(t)} = \mathbf{S}^{(t)} + \lambda n^{-1}\mathbf{I}$. This resembles the popular ridge regularization for covariance estimation (Lin and Perlman, 1985; Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005). However, with a positive K, the additional shrinkage factor $n/(n + K)$ imposed on $\mathbf{S}^{(t)}$ helps to stabilize the variability of $\Sigma^{(t)}$ and often improves the performance of parameter estimates, as illustrated in Section 4.1.

When missingness is non-ignorable, we need to incorporate the mechanism of missingness in the likelihood function in order to obtain reliable parameter estimates. In Section 3.2.1, we consider the situation where the nuisance parameter of the missing data mechanism Γ is known; while in Section 3.2.2, we propose a profile likelihood approach to jointly estimate the missing data model parameters and the multivariate Gaussian parameters.

3.2.1 A PEMM algorithm for abundance-dependent missing data—

With the missing data mechanism incorporated, the *E*- and *M-steps* of PEMM (Algorithm 1) aim to solve:

$$\mathbf{\Omega}^{(t)}=\arg\max_{\mathbf{\Omega}}\left\{E_{\mathbf{X}_{mis}|\mathbf{X}_{obs},\mathbf{M};\mathbf{\Omega}^{(t-1)}}\left[L(\mathbf{X},\mathbf{M};\mathbf{\Omega})\right]-P(\mathbf{\Omega})\right\}. \quad (7)$$

Following the notation in previous sections, we denote the missing data mechanism function as $P(m_{ij} = 1|x_{ij}) = g(x_{ij}; \mathbf{c_{ij}}, \Gamma)$, where $\mathbf{c_{ij}}$ is some known covariate information; Γ is the parameter of the missing data mechanism and is distinct from the parameter of interest $\Omega = \{\boldsymbol{\mu}, \Sigma\}$. Assume Γ is known and denote $g(\cdot) = g(\cdot; \mathbf{c_{ij}}, \Gamma)$. Note, $g(\cdot)$ may depend on feature/sample-specific covariates. We first derive the solution in (7) for a general missing data mechanism and the Inverse-Wishart penalty (5).

Proposition 1: Let $G(\mathbf{X}_{i,mis})=\prod_{j:m_{ij}=1} g(x_{ij})$. For $i \in \{1, \cdots, n\}$ and $j, j' \in \{1, \cdots, p\}$, denote $\alpha_i^{(t)}=E_{\mathbf{X}_{i,mis}|\mathbf{X}_{i,obs};\,\boldsymbol{\mu}^{(t-1)},\sum^{(t-1)}}[G(\mathbf{X}_{i,mis})];\beta_{ij}^{(t)}$ $=E_{\mathbf{X}_{i,mis}|\mathbf{X}_{i,obs};\,\boldsymbol{\mu}^{(t-1)},\sum^{(t-1)}}[x_{ij}G(\mathbf{X}_{i,mis})];\delta_{ijj'}^{(t)}$ $=E_{\mathbf{X}_{i,mis}|\mathbf{X}_{i,obs};\,\boldsymbol{\mu}^{(t-1)},\sum^{(t-1)}}[x_{ij}x_{ij'}G(\mathbf{X}_{i,mis})]$; and $\Theta_{i,jj'}^{(t)}(\boldsymbol{\mu})=\delta_{ijj'}^{(t)}-\beta_{ij'}^{(t)}\mu_j-\beta_{ij}^{(t)}\mu_{j'}+\alpha_i^{(t)}\mu_j\mu_{j'}$. Then for multivariate Gaussian data, (7) can be calculated as

$$\boldsymbol{\mu}^{(t)}=n^{-1}\sum_{i=1}^{n}\boldsymbol{\beta}_i^{(t)}/\alpha_i^{(t)}, \quad \sum^{(t)}=(n+K)^{-1}\left\{\sum_{i=1}^{n}\mathbf{\Theta}_i^{(t)}\left(\boldsymbol{\mu}^{(t)}\right)/\alpha_i^{(t)}+\lambda\boldsymbol{I}\right\}. \quad (8)$$

The proof of Proposition 1 is provided in the Web Appendix B.

Comparing equations (6) and (8), we see that the assumption about the missing data mechanism leads to weighting the sample *i* by $1/\alpha_i^{(t)}$, when updating the mean and covariance estimates. This is similar in spirit to inverse probability weighting, another popular technique for handling missing data, including NMAR data, in other contexts (Robins et al., 1995; Robins and Rotnitzky, 1995).

In Section 2.1, we introduced an abundance dependent missing data mechanism in equation (1), motivated by data characteristics in proteomics studies. Here we assume that with the true parameter values, $(\boldsymbol{\mu}, \Sigma, \Gamma)$, the missing probability is always non-negative such that $P(-\gamma_1-\gamma_2x_{ij}-\boldsymbol{\gamma}_3^T\mathbf{c}_{ij}<0)<\varepsilon$, for some small $\varepsilon > 0$. Then, instead of equation (1), we choose to use the unbounded $g(\cdot)$ function:

$$g(x_{ij};\mathbf{c}_{ij},\mathbf{\Gamma})=\exp\{-\gamma_1-\gamma_2x_{ij}-\boldsymbol{\gamma}_3^T\mathbf{c}_{ij}\}. \quad (9)$$

when solving the optimization problem. Equation (9) is a simplified form of equation (1). It can be easily integrated in the density function of multivariate Gaussian. This strategy has very limited impact on the parameter estimation while simplifying the computation. In application, if a positive value of $-\gamma_1-\gamma_2 x_{ij}-\boldsymbol{\gamma}_3^T\mathbf{c}_{ij}$ is encountered, i.e., the probability of missingness exceeds 1, the x_{ij} value can be reset to be missing, to avoid the possibility of a high leverage observation.

In equation (9), the $\exp\left(-\gamma_1-\gamma_3^T\mathbf{c}_{ij}\right)$ component does not depend on x_{ij}, so it can be treated as a scaling constant and moved to the outside of the conditional expectation when calculating α_i, β_{ij}, and $\delta_{ijj'}$ as defined in Proposition 1. Since the scaler appears in both the numerator and the denominator in equation (8), it cancels out and does not contribute to the calculation of MPLEs. Thus, we can focus on the reduced model:

$$g(x,\boldsymbol{\Gamma})=\exp\{-\gamma_2 x\}. \quad (10)$$

Below, we will derive the detailed updating formula in the PEMM algorithm for the missing data mechanism in (10) with multivariate Gaussian data. The log-ratio abundance dependent missing data mechanism in (2) can be similarly simplified to $g(x, \Gamma) = \exp\{-\gamma_2 x^2\}$. Derivations of the log-ratio incomplete data are provided in the Web Appendix F.

Proposition 2: Define $\hat{\mathbf{X}}_i^{(t)}$ such that $\hat{\mathbf{X}}_{i,obs}^{(t)}=\mathbf{X}_{i,obs}$, and

$$\hat{\mathbf{X}}_{i,mis}^{(t)}=\boldsymbol{\mu}_{i,mis}^{(t-1)}+\Sigma_{i:mis,obs}^{(t-1)}\left(\Sigma_{i:obs,obs}^{(t-1)}\right)^{-1}\left(\mathbf{X}_{i,obs}-\boldsymbol{\mu}_{i,obs}^{(t-1)}\right)-\gamma_2\mathbf{A}_{i:mis,mis}^{(t)}\cdot\mathbf{1},$$

where $\mathbf{A}_{i:mis,mis}^{(t)}=\Sigma_{i:mis,mis}^{(t-1)}-\Sigma_{i:mis,obs}^{(t-1)}\left(\Sigma_{i:obs,obs}^{(t-1)}\right)^{-1}\Sigma_{i:obs,mis}^{(t-1)}$. Then for the missing data mechanism specified in equation (10), the solutions in the *M-step* are given by

$$\boldsymbol{\mu}^{(t)}=n^{-1}\sum_i\hat{\mathbf{X}}_i^{(t)},\ \Sigma^{(t)}=(n+K)^{-1}\left\{\sum_i\left(\left(\hat{\mathbf{X}}_i^{(t)}-\boldsymbol{\mu}^{(t)}\right)\left(\hat{\mathbf{X}}_i^{(t)}-\boldsymbol{\mu}^{(t)}\right)^T+\mathbf{A}_i^{(t)}\right)+\lambda\mathbf{I}\right\},$$

where $\mathbf{A}_i^{(t)}$ is a $p \times p$ matrix with all elements being zero except the submatrix $\mathbf{A}_{i:mis,mis}^{(t)}$.

The proof of Proposition 2 is provided in Web Appendix D.

In the PEMM algorithm, there are two penalty parameters, λ and K, to be specified by the user. Non-zero λ and K help to smooth the penalized likelihood function and speed up the convergence rate (Green, 1990). Particularly, $\Sigma^{(t)}$ is made positive-definite by sufficiently large λ at each EM iteration. Therefore, we propose to use a descending sequence of $\{\lambda^{(t)}\}_t$ when implementing PEMM. This strategy helps to further stabilize and speed up the algorithm. Specifically, we begin with a sufficiently large $\lambda^{(0)}$ to assure the positive definiteness of the covariance matrix estimate. As iteration proceeds, we allow $\lambda^{(t)}$ to change with the minimum eigenvalues of $\Sigma^{(t)}$. After a few iterations, as the Σ estimate becomes positive-definite, $\lambda^{(t)}$ is set to the user-specified λ and stays unchanged for the remaining iterations. Details of this procedure are summarized in Algorithm 2. In the next section, we show via simulations that the performance of the PEMM algorithm is relatively robust to the choice of penalty parameters.

Algorithm 2

A PEMM algorithm for the abundance-dependent missing data mechanism in (9).

```
1  Specify positive $\lambda$ and $K$.
2  Based on available cases, obtain the sample mean $\bar{\mathbf{X}}$ and sample covariance $\mathbf{S}_\mathbf{X}$. Then find the smallest positive $\lambda^{(0)}$ such that $\lambda^{(0)} \geq \lambda$ and the minimum eigenvalue of matrix $n\mathbf{S}_\mathbf{X} + \lambda^{(0)}\mathbf{I}$ is positive. Set $\boldsymbol{\mu}^{(0)} = \bar{\mathbf{X}}$ and $\Sigma^{(0)} = (n + K)^{-1}(n\mathbf{S}_\mathbf{X} + \lambda^{(0)}\mathbf{I})$.
3  Calculate the conditional expectation of the sufficient statistics given the current parameter estimate $(\boldsymbol{\mu}^{(t-1)},\Sigma^{(t-1)})$:
       $\mathbf{A}^{t)}_{i:mis,mis} = \Sigma^{(t-1)}_{i:mis,mis} - \Sigma^{(t-1)}_{i:mis,obs}(\Sigma^{(t-1)}_{i:obs,obs})^{-1}\Sigma^{(t-1)}_{i:obs,mis},$
       $\hat{\mathbf{X}}^{(t)}_{i,obs} = X_{i,obs}, \hat{\mathbf{X}}^{(t)}_{i,mis} = \mu^{(t-1)}_{i,mis} + \Sigma^{(t-1)}_{i:mis,obs}(\Sigma^{(t-1)}_{i:obs,obs})^{-1}(\mathbf{X}_{i,obs} - \mu^{(t-1)}_{i,obs}) - \gamma_2\mathbf{A}^{(t)}_{i:mis,mis} \cdot 1.$
4  Calculate the maximum penalized likelihood estimates:
       $\mu^{(t)} = n^{-1}\sum_i \hat{\mathbf{X}}^{(t)}_i, \quad \Sigma^{(t)} = (n + K)^{-1}\left(\sum_i\left((\hat{\mathbf{X}}^{(t)}_i - \mu^{(t)})(\hat{\mathbf{X}}^{(t)}_i - \mu^{(t)})^T + \mathbf{A}^{(t)}_i\right) + \lambda^{(t)}\mathbf{I}\right),$
   where $\lambda^{(t)}$ is chosen to be the smallest value which makes $\Sigma^{(t)}$ positive-definite and is greater than or equal to $\lambda$.
5  Repeat 3–4 until convergence.
```

3.2.2 A profile likelihood approach to jointly estimate Γ and (μ,Σ)—

In Section 3.2.1, we implemented the PEMM algorithm for multivariate Gaussian parameter estimation when the missing data mechanism parameter Γ in (9) is known. Here we propose a profile likelihood approach to jointly search for the MPLEs of $(\Gamma,\boldsymbol{\mu},\Sigma)$. For a given Γ, we can rewrite the penalized log-likelihood as

$$\tilde{L}_\Gamma(\boldsymbol{\mu},\sum)=\tilde{L}(\mathbf{X}_{obs},\mathbf{M};\boldsymbol{\mu},\sum,\boldsymbol{\Gamma})=L(\mathbf{X}_{obs},\mathbf{M};\boldsymbol{\mu},\sum,\boldsymbol{\Gamma})-P(\sum),$$

and we can use the PEMM algorithm (Algorithm 2) to calculate

$$(\hat{\boldsymbol{\mu}}_\boldsymbol{\Gamma},\hat{\sum}_\boldsymbol{\Gamma})=\arg\max_{\boldsymbol{\mu},\sum}\tilde{L}_\boldsymbol{\Gamma}(\boldsymbol{\mu},\sum). \quad (11)$$

Note that we again omit the covariates in the above equations to simplify the presentation. Then, to obtain the MPLE of Γ, we can evaluate $\tilde{L_\Gamma}(\hat{\boldsymbol{\mu}_\Gamma}, \hat{\Sigma_\Gamma})$ at different Γ values and choose the Γ that gives the maximum over the likelihood profile:

$$\hat{\mathbf{\Gamma}}=\arg\max_{\mathbf{\Gamma}}\tilde{L}_{\mathbf{\Gamma}}(\boldsymbol{\mu},\sum)=\arg\max_{\mathbf{\Gamma}}\tilde{L}_{\Gamma}(\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}},\hat{\sum}_{\mathbf{\Gamma}}). \quad (12)$$

Evaluating $\tilde{L}_{\Gamma}(\hat{\boldsymbol{\mu}}_{\Gamma}, \hat{\Sigma}_{\Gamma})$ involves integrating out $\{\mathbf{X}_{i,mis}\}$ in the joint penalized log-likelihood function of the complete data based on $(\Gamma, \hat{\boldsymbol{\mu}}_{\Gamma}, \hat{\Sigma}_{\Gamma})$. For the abundance-dependent missing data mechanism in equation (9), we have

$$\tilde{L}_{\mathbf{\Gamma}}(\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}},\hat{\sum}_{\mathbf{\Gamma}})=\sum_i\left\{\log\Phi_{\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}},\hat{\sum}_{\mathbf{\Gamma}}}(\mathbf{X}_{i,obs})+\sum_{j\in\mathbf{O}_i}\log(1-g_{\mathbf{\Gamma}}(x_{ij}))-\gamma_1(p-|\mathbf{O}_i|)\right.$$
$$\left.-\boldsymbol{\gamma}_3{}^T\sum_{j\notin\mathbf{O}_i}\mathbf{c}_{ij}+\log E_{\mathbf{X}_{i,mis}|\mathbf{X}_{i,obs};\hat{\boldsymbol{\mu}}_{\mathbf{\Gamma}},\hat{\sum}_{\mathbf{\Gamma}}}\left[\exp(-\gamma_2\sum_{j\notin\mathbf{O}_i}x_{ij})\right]\right\}-P(\hat{\sum}_{\mathbf{\Gamma}}),$$

where Φ(·) is the density function of multivariate Gaussian distribution; and $|\mathbf{O}_i|$ denotes the number of elements in set $\mathbf{O}_i$. The conditional expectation in the above equation can be calculated in the same way as outlined in the proof of Proposition 2. Since $g(\cdot)$ is not bounded, it can be greater than one, and then $\log(1 - g_{\Gamma}(x_{ij}))$ cannot be computed. In practise, if a negative value or a value close to zero for $1 - g_{\Gamma}(x_{ij})$ is encountered, we suggest ignoring the corresponding data point by resetting $m_{ij} = 1$ (*i.e.* pretending x_{ij} is missing). This will help to avoid the possibility of a high leverage observation. However, we expect this to be a very rare occurrence for reasonable Γ values. Indeed, in the simulation and the real data application, when the estimated $g(\cdot)$ function is evaluated at the observed data points, its value rarely exceeded one.

As mentioned in the previous section, for the abundance-dependent missing data mechanism in equation (9), the MPLEs of Gaussian parameters in equation (11) only depend on γ_2, while not depending on γ_1 nor $\boldsymbol{\gamma}_3^T\mathbf{c}_{ij}$. However, incorporating the latter two terms in the missing data mechanism would enable one to handle either NMAR ($\gamma_2 > 0$) or MAR data ($\gamma_2 = 0$, $\gamma_1 > 0$ or $\gamma_3 > 0$), and improves the overall performance of the proposed method by enhancing the accuracy of γ_2 estimation in the profile-likelihood approach.

In application, solving the optimization problem in equation (12) is not easy. When there are substantial amounts of missing data, the log-likelihood surface is often not convex. Thus, general purpose optimization algorithms do not apply, as they may easily converge to local extreme or saddle points. On the other hand, performing a thorough grid search for $\Gamma = (\gamma_1, \gamma_2, \gamma_3)$ can be computationally intensive. To circumvent this difficulty, we propose the following strategy: we first obtain a good initial estimate of Γ and then perform a small neighbourhood search of Γ around the initial estimate to find the solution maximizing the profile likelihood. Specifically, we take the available-case mean estimate for each 'protein' as x_j and the missing percentage of that protein as y_j. We then regress $\log(y_j)$ on x_j to obtain $\hat{\mathbf{\Gamma}}^{\mathbf{0}}=(\hat{\gamma}_1^0,\hat{\gamma}_2^0,\hat{\gamma}_3^0)$. Since $\hat{\Gamma}_0$ is based on available-case estimates of mean protein abundance, it might be biased though is likely to be close to the MPLE. Then, we perform a small neighbourhood grid search of γ_2 to find the estimate maximizing the profile likelihood while fixing γ_1 and γ_3 estimates, since the MPLEs of Gaussian parameters in equation (11) only depend on γ_2. In this way, we reduce the potential uncertainty in estimating Γ. Details of this procedure are outlined in Web Appendix E.

4. Simulations

4.1 Penalty parameters

We first investigate the impact of different choices of penalty parameters λ and K. Multivariate Gaussian data $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ were simulated with $p = 10, 20, 30$, where μ_j is randomly sampled from $\{0, -3, 3\}$; $\Sigma_{jj} = 1$; and $\Sigma_{jj'}$, $j \neq j'$ is either set to 0 or sampled from $N(0.5, 0.1^2)$, each with probability 0.5. We then simulate ~ 30% of the data to be MAR. For the i^{th} sample, we first randomly pick two features, j_1 and j_2, and set the missing probability of these two features to be 0 in this sample. For the other features in the sample, we set their missing probabilities to be $P(m_{ij} = 1) \propto \exp(-(|x_{ij_1}| + |x_{ij_2}|))$. Thus, $P(m_{ij} = 1)$ only depend on observed values in each sample, but do not depend on the missing values. This satisfies the definition of MAR.

For each choice of penalty parameters, we calculate the ratios of the mean squared errors of estimates resulting from the penalized EM algorithm based on the incomplete data to the mean squared errors of the MLEs based on the complete data. These ratios are then referred to as *relative mean squared errors* (RMSEs) hereafter. Figure 1 shows the average RMSEs of Σ estimates over 100 simulations at different $\lambda = \{10, 5, 1\}$ and $K = \{0, 1, \ldots, 10\}$. We omit the results on the RMSEs of $\boldsymbol{\mu}$ estimates, as the parameter space of $\boldsymbol{\mu}$ is not regularized in the penalty term and thus different choices of penalty parameters show minor impact on $\boldsymbol{\mu}$ estimates. Evident from Figure 1, for both sample sizes considered (n =100 and 10), the performance of the penalized EM algorithm is relatively robust to different choices of λ and K, especially when sample size is large. When sample size is limited and the dimensionality p is relatively large, non-zero λ and K are necessary to keep the positive definiteness of the covariance and concentration matrix. We observe similar results with other p, n, correlation patterns and missing data patterns (data not shown). Based on these simulations, we choose $\lambda = 5$, $K = 5$ in the following analyses, and this choice seems to give favorable RMSEs throughout all our simulations.

4.2 Comparison with competing methods

We simulate multivariate Gaussian data with $p = 10, 20, 30, \mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. We consider both large and small sample sizes: $n = 100$ ($p < n$) and $n = 10$ ($p \geq n$). We sample μ_j independently from $\{4, 8\}$, while Σ is simulated as before. Missing data are simulated based on the abundance dependent mechanisms for positive data with $\gamma_2 = 0.3$, $\gamma_1 = \gamma_3 = 0$, i.e. $\mathbf{P}(m_{ij} = 1) = \exp(-0.3x_{ij})$. There are a few negative x_{ij} values generated in this scenario for which $\exp(-0.3x_{ij}) > 1$, and they are set to be missing as well. Overall, there are about 40% missing data.

We investigate the performances of six different methods: (1) AC+P: available-case analysis with $\hat{\Sigma}(\lambda) = \mathbf{S} + \lambda/n \cdot \mathbf{I}$; (2)Imp+P: imputing the missing values by the k-nearest neighbors (kNN) algorithm (Troyanskaya et al., 2001) with k chosen by cross-validation, and obtaining $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}_{Imp}$ and $\hat{\Sigma}(\lambda) = \mathbf{S}_{Imp} + \lambda/n \cdot \mathbf{I}$; (3) EM: the EM algorithm; (4) PEM: the penalized EM algorithm; (5) PEMM: the PEMM algorithm using the true missing data parameter Γ (Section 3.2.1); (6) PEMMe: the PEMM algorithm using the profile-likelihood-based estimate of Γ (Section 3.2.2). Note, $\lambda = 5$ is used for AC+P and Imp+P; and $(\lambda, K) = (5, 5)$ is used for PEM, PEMM, and PEMMe.

For each simulation setting, we generate 1000 independent datasets, and obtain RMSEs of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ by different methods relative to estimates from complete data without missing values. The results are shown in Figures 2 and 3. When sample size is large, the PEMM and PEMMe estimates perform favorably (yield smaller RMSEs) compared to other methods in various scenarios. When sample size is small, the PEMM/PEMMe methods still yield smaller RMSEs for the mean estimates, but the improvement of PEMM/PEMMe over PEM for covariance estimation becomes less obvious. This probably occurs because the variance component is dominant compared to the bias component in the MSE of $\hat{\Sigma}$ when sample size is small. The regularization, which controls the variance of $\hat{\Sigma}$, dramatically improves the MSE of all penalized EM estimates (right column of Figure 3), demonstrating the benefit of proper regularization in high dimensional settings. On the other hand, incorporating the missing data mechanism may help to reduce bias. But since the bias of $\hat{\Sigma}$ is dominated by the variance of $\hat{\Sigma}$, the improvement on the former is not quite visible in this setting.

In Web Appendix F, we also show the simulation results for the two-sided log-ratio abundance-dependent missing data mechanism in equation (2). Similarly, the proposed PEMM and PEMMe methods yield smaller RMSEs for μ and Σ estimates with both large and small sample sizes. The performance of the PEMM and PEMMe estimates also appear to have desirable robustness to departure from normality. (data not shown due to space limitation).

5. Application – estimating the mean abundance for spiked-in human proteins in yeast

We apply the proposed PEMM algorithm to a real data example from the study conducted by the Clinical Proteomic Technologies for Cancer consortium (Paulovich et al., 2010). In the study, the Universal Proteomics Standard Set 1 (UPS1) collection of 45 human source or human sequence recombinant proteins were spiked into yeast protein lysate samples at different concentrations and quantified by mass spectrometry. We focus on the data corresponding to a subset of the spiked-in samples, designated C, D and E, from the study. The compositions of C, D, and E include 60*ng/μL* yeast lysate together with 2.2, 6.7 and 20*fmol/μL* UPS1 respectively. For each of C, D and E experiments, 12 samples were obtained by multiple collaborating labs. Abundance of each protein in each sample was derived using software *Sahale* (Milac et al., 2012). Not surprisingly, the human proteins show different abundances and are subject to different probability of missingness across protein profiles of different samples. Specifically, there are 51.1, 23.7 and 9.8% of the human protein abundance measures missing in the protein profiles of the samples from C, D and E experiments, respectively.

In this dataset, all abundance measures are positive and are roughly normally distributed. We consider the missing data mechanism specified in (9) and adjust for the number of peptides in each protein as a covariate. We use the proposed profile-likelihood-based PEMM algorithm in Section 3.2.2 to estimate the missing data mechanism parameters and the mean and covariance of the abundance of 45 human proteins in the three experiments. The estimated missing probabilities evaluated at observed data range from 0.001 to 0.878. Figure 4 shows the frequency distributions for the differences between the mean estimates by AC +P and those by PEMM. Similar plots for the differences between the mean estimates by PEM and PEMM are shown in Web Figure 3. For data from experiment C, which has more than 50% missingness, the mean estimates by AC+P and PEM are much larger than that by PEMM. In samples from experiment E, the percentage of missingness in the data is smaller. The difference between the estimates by PEMM estimates and that by AC+P and PEM becomes smaller too. These patterns are consistent with what we observed in simulations in the previous subsection: when the data are NMAR, the proposed PEMM algorithm can yield estimates with less bias, compared to estimates based on other methods that ignore the missing data mechanism.

6. Discussion

To estimate the mean and covariance for multivariate Gaussian data with substantial missingness, it is important to characterize the missing data mechanism. If the probability of a value being missing depends on the missing values themselves, one needs to take the missing data mechanism into account in parameter estimation. In this work, we propose a penalized EM algorithm incorporating missing data mechanism (PEMM) for multivariate Gaussian parameter estimation. Specifically, motivated by data characteristics in proteomics studies, we discussed two types of abundance dependent missing data mechanisms, and derived detailed formula for the corresponding PEMM algorithms. Furthermore, in the PEMM algorithm, we introduce penalization into the full log-likelihood to regularize the parameter estimation.

We proposed an Inverse-Wishart penalty, because it yields a positive-definite estimate of the covariance matrix and is computationally efficient with a simple closed form solution in the M-step. There are two tuning parameters, λ and K, in the Inverse-Wishart penalty. In the paper, $\lambda = 5$ and $K = 5$ is used in all of the numerical studies. These values may not be optimal in other general applications. The method for selecting the optimal tuning parameter in the PEMM algorithm warrants future research. In addition to the Inverse-Wishart penalty, other convex penalty functions can be conveniently incorporated in the PEMM framework, because the non-ignorable missing data mechanism only affects the E-step and does not directly interfere with the penalty function in the M-step. Although it is possible to develop a full Bayesian approach within the current framework, it is beyond the scope of this work.

Better estimation of multivariate Gaussian parameters with incomplete data can valuably facilitate high-dimensional ’omics data analysis. For example, the PEMM framework could be used for imputation of missing data under NMAR. In the E-step of PEMM, the conditional expectation of missing data given the observed data and the missing data mechanism could serve as natural “imputation” of the missing data. Future work to investigate the merit of such an approach is warranted. The framework of the PEMM algorithm could also be extended to non-Gaussian distributions and other missing data mechanisms. However, different forms of penalty terms may then be needed. In addition, for general non-ignorable missingness other than the abundance dependent missing data mechanism, the implementation of the PEMM could be complicated. An R package PEMM will be available on CRAN (http://cran.r-project.org/) soon.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH grants R03CA174984, P01CA53996, R01GM082802 and SUB-CA160034. The authors thank Dr. Xiao-Li Meng for insightful suggestions.

References

Afifi AA, Elashoff RM. Missing observations in multivariate statistics: Review of the literature. Journal of the American Statistical Association. 1966; 61I:595–604.

Beale EML, Little RJA. Missing values in multivariate analysis. Journal of the Royal Statistical Society: Series B. 1975; 37(1):129–145.

Chen LS, Paul D, Prentice RL, Wang P. A regularized Hotelling's T^2 test for path-way analysis in proteomic studies. Journal of the American Statistical Association. 2011; 106(496):1345–1360. [PubMed: 23997374]

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B. 1977; 39(1):1–38.

Faca V, Coram M, Phanstiel D, Glukhova V, Zhang Q, Fitzgibbon M, McIntosh M, Hanash S. Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. J Proteome Res. 2006; 5(8):2009–2018. [PubMed: 16889424]

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

Green PJ. On use of the EM algorithm for penalized likelihood estimation. Journal of the Royal Statistical Society: Series B. 1990; 52:443–452.

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004; 88:365–411.

Lin, SP.; Perlman, MD. In: Krishnaiah, P., editor. A Monte Carlo comparison of four estimators of a covariance matrix; Multivariate analysis–VI: Proceedings of the Sixth International Symposium on Multivariate Analysis; 1985. p. 411-429.

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: Wiley; 2002.

McLachlan, GJ.; Krishnan, T. The EM algorithm and extensions. 2. Wiley-Interscience; 1996.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80(2):267–278.

Milac TI, Randolph TW, Wang P. Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. Statistics and Its Interface. 2012; 5(1):75–87. [PubMed: 24163717]

Neal, RM.; Hinton, GE. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, MI., editor. Learning in Graphical Models. 1999. p. 355-368.

Paulovich A, Billheimer D, Ham A, Vega-Montoto L, Rudnick PA, et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. Molecular & Cellular Proteomics. 2010; 9(2):242–254. [PubMed: 19858499]

Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models. Journal of the American Statistical Association. 1995; 90(429):122–129.

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90(429):106–121.

Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.

Rubin DB. Inference and missing data. Biometrika. 1976; 63(3):581–592.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. J. Wiley and Sons; New York: 1987.

Rubin DB. Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association. 1996; 91(434):473–489.

Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology. 2005; 4(1):Article 32.

Schafer, JL. Analysis of Incomplete Multivariate Data. Chapman and Hall; London: 1997.

Schneider T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate. 2001; 14:853–871.

Städler N, Bühlmann P. Missing values: sparse inverse covariance estimation and an extension to sparse regression. Satistics and Computing. 2012; 22:219–235.

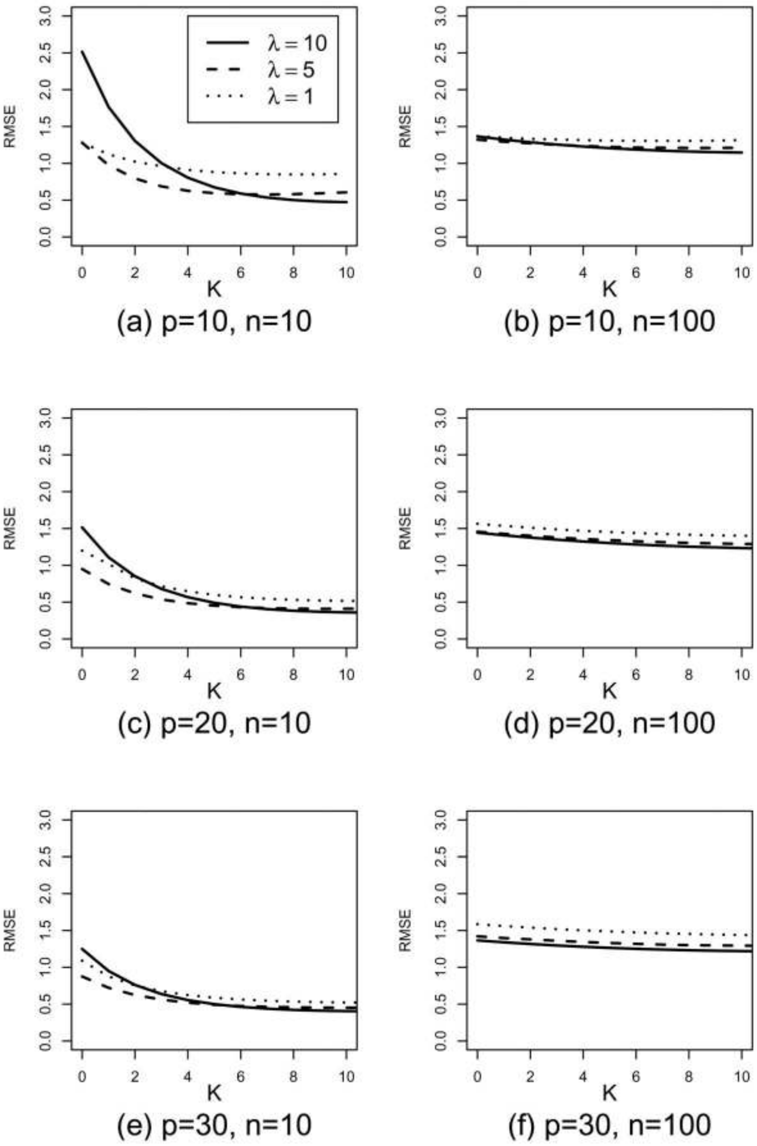
Städler, N.; Stekhoven, DJ.; Bühlmann, P. Pattern alternating maximization algorithm for missing data in large *p*, small *n* problems. 2012.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–525. [PubMed: 11395428]

Vale CD V, Maurelli A. Simulating multivariate nonnormal distributions. Psychometrika. 1983; 48(3): 465–471.

Wu CFJ. On the convergence properties of the EM algorithm. Annals of Statistics. 1983; 11(1):95–103.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Jounral of the Royal Statistical Society: Series B. 2006; 68(1):49–67.



(a) p=10, n=10

(b) p=10, n=100

(c) p=20, n=10

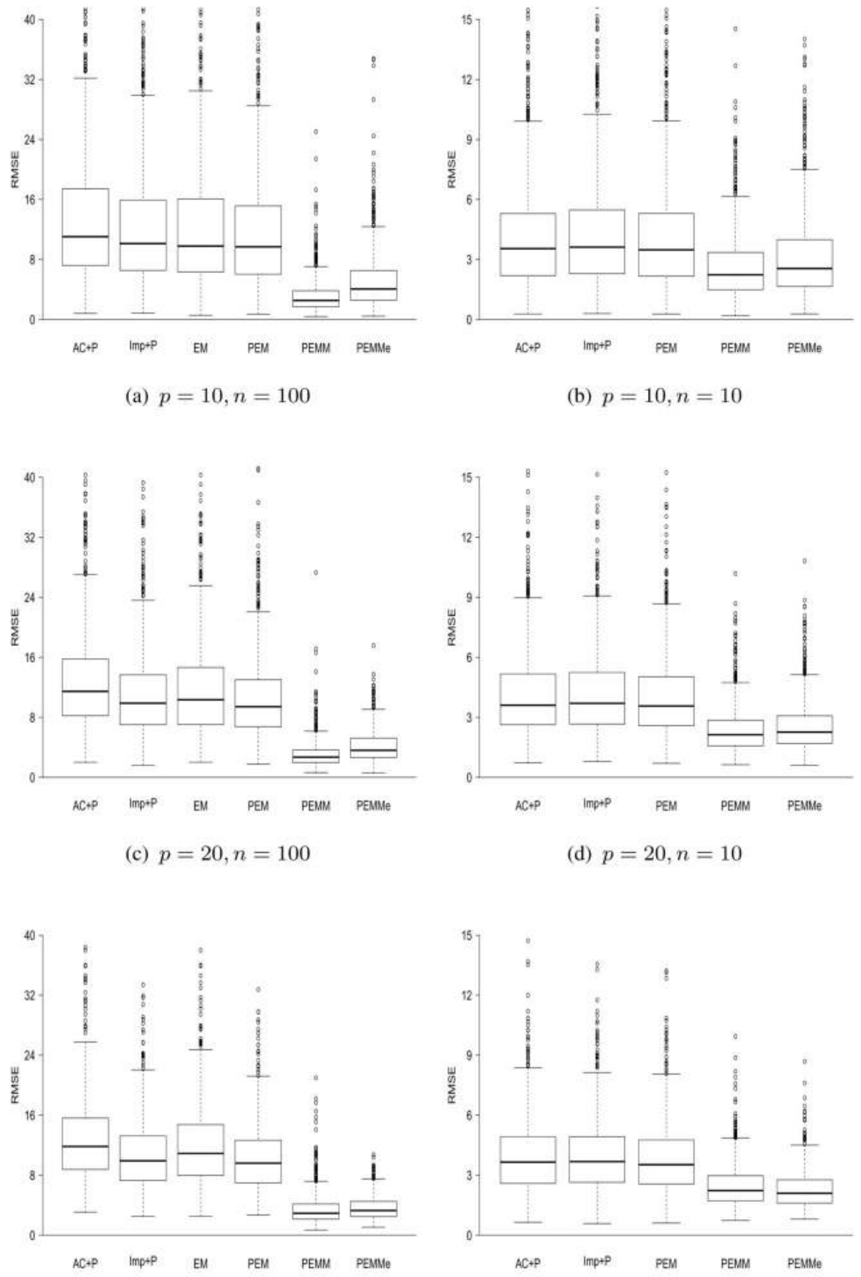
(d) p=20, n=100

(e) p=30, n=10

(f) p=30, n=100

Figure 1.

Impact of different penalty parameters on parameter estimation of the PEM algorithm. Here we compare the average relative mean squared errors (RMSE) of Σ estimates over 100 simulations at different $\lambda = \{10, 5, 1\}$ and $K = \{0, 1, \ldots, 10\}$ for different $p = \{10, 20, 30\}$ and $n = \{10, 100\}$. The RMSEs of Σ estimates are relative robust to the choices of penalty parameters when $K \in [5, 10]$ and $\lambda \in [1, 10]$.



(a) $p = 10, n = 100$

(b) $p = 10, n = 10$

(c) $p = 20, n = 100$

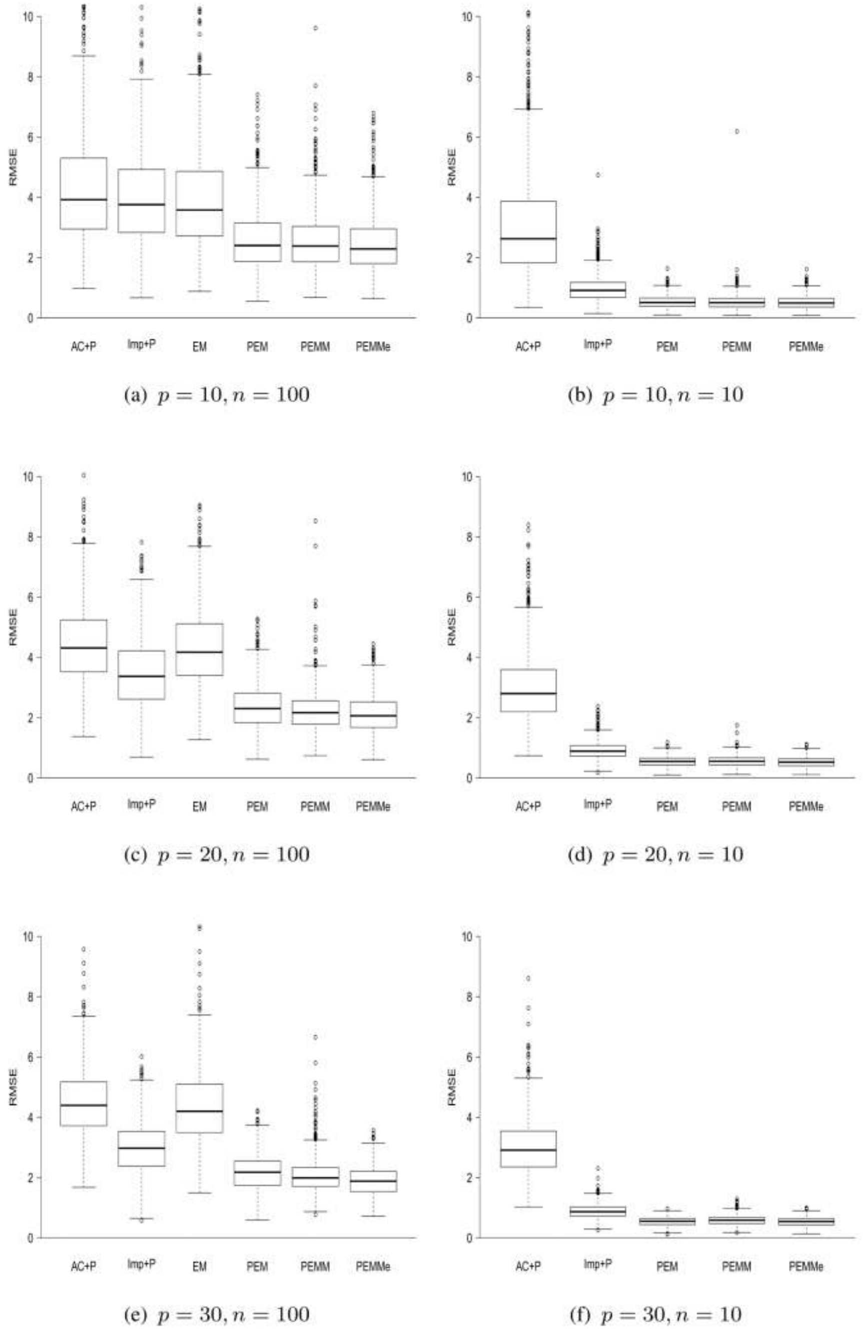
(d) $p = 20, n = 10$

(e) $p = 30, n = 100$

(f) $p = 30, n = 10$

Figure 2.

Boxplots of RMSEs of $\hat{\boldsymbol{\mu}}$ by different methods on positive abundance-dependent missing data. For different combinations of p and n, $p = \{10, 20, 30\}$ and $n = \{100, 10\}$, we compare the RMSEs of $\hat{\boldsymbol{\mu}}$ by six methods: available-case analysis with ridge regularization applied on $\hat{\Sigma}$ (AC+P); imputation with kNN followed by estimating the mean and ridge regularized $\hat{\Sigma}$ on the imputed data (Imp+P); the EM algorithm (only applied for $n > p$); the penalized EM algorithm (PEM); the PEMM algorithm with true Γ (PEMM); and the PEMM algorithm with the profile-likelihood-based estimate of Γ (PEMMe).



(a) $p = 10, n = 100$

(b) $p = 10, n = 10$

(c) $p = 20, n = 100$

(d) $p = 20, n = 10$

(e) $p = 30, n = 100$

(f) $p = 30, n = 10$

Figure 3.

Boxplots of RMSEs of $\hat{\Sigma}$ by different methods on positive abundance-dependent missing data. For different combinations of p and n, $p = \{10, 20, 30\}$ and $n = \{100, 10\}$, we compare the RMSEs of $\hat{\Sigma}$ by six methods: available-case analysis with ridge regularization applied on $\hat{\Sigma}$ (AC+P); imputation with kNN followed by estimating the mean and ridge regularized $\hat{\Sigma}$ on the imputed data (Imp+P); the EM algorithm (only applied for $n > p$); the penalized EM algorithm (PEM); the PEMM algorithm with true Γ (PEMM); and the PEMM algorithm with the profile-likelihood-based estimate of Γ (PEMMe).

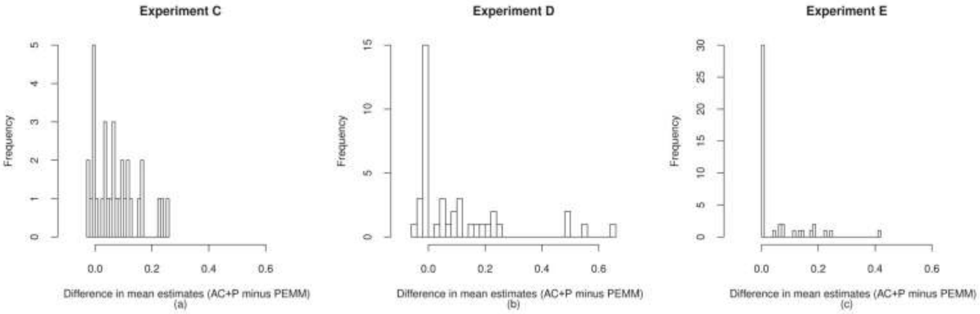


Figure 4.

Histograms of differences in protein abundance estimates based on available-case analysis with ridge penalization on covariance (AC+P) versus the PEMM algorithm, in three different experiments, C, D, and E. There are 51.1, 23.7 and 9.8% of protein abundance measures missing in the samples from Experiments C, D, and E, respectively. For proteins with no missingness, the corresponding differences are 0 and are not plotted in the figure.