Accounting for Spot Matching Uncertainty in the Analysis of Proteomics Data from Two-dimensional Gel Electrophoresis

Volodymyr Melnykov*, Ranjan Maitra† and Dan Nettleton†

Abstract

Two-dimensional gel electrophoresis is a biochemical technique that combines isoelectric focusing and SDS-polyacrylamide gel technology to achieve simultaneous separation of protein mixtures on the basis of isoelectric point and molecular weight. Upon staining, each protein on a gel can be characterized by an intensity measurement that reflects its abundance in the mixture. These can then conceptually be used to determine which proteins are differentially expressed under different experimental conditions. We propose an EM approach to identify differentially expressed proteins using an inferential strategy that accounts for uncertainty in matching spots to proteins across gels. The underlying mixture model has trivariate Gaussian components. The application of the EM is however, not straightforward, with the main difficulty lying in the E-step calculations because of the dependent structure of proteins within each gel. Therefore, the usual model-based clustering approach is inapplicable, and an MCMC approach is employed. Through data-based simulation, we demonstrate that our proposed method effectively accounts for uncertainty in spot matching and more successfully distinguishes differentially and non-differentially expressed proteins than a naïve t-test which ignores uncertainty in spot matching.

Keywords: EM algorithm, Markov chain Monte Carlo, Gaussian mixture model

1 Introduction

Two-dimensional gel electrophoresis (2DGE) is one of the oldest and most commonly used proteomics technologies (Morris et al., 2008). One of the main uses of 2DGE is to identify proteins that differ in abundance across two or more conditions. In this article, we propose a new method for identifying such proteins that incorporates uncertainty in spot matching. Accounting for this uncertainty is especially important because spot mismatches are considered to be one of the major sources of variation in 2DGE data (Almeida et al., 2003).

In 2DGE, proteins in a sample are separated on a gel in one dimension according to their isoelectric points and in a second, perpendicular dimension according to their molecular weights. The separated proteins are dyed or stained and the gel scanned so that specialized 2DGE image analysis software can be used to identify protein spot locations and determine measures of protein abundance at each distinct spot. In studies or experiments that involve multiple gels, alignment algorithms are used to align gel images so that the vertical and horizontal coordinates providing spot locations are comparable from image to image. Dowsey et al. (2003) describe various alignment approaches and provide references to much of the relevant literature. An overview of these problems from a statistical perspective was also provided by Roy et al. (2003). More recently, Green and Mardia (2006) proposed a Bayesian hierarchical approach to aligning electrophoresis gels.

Following alignment, the location coordinates and an abundance measure– usually a reflectance intensity – are available, in principle, for each of several hundred spots on each gel. However, such data are typically not recorded. Despite alignment, spot locations are not identical from gel to gel, so gel analysis software (recently reviewed by Palagi et al. (2006)) is used to match spots across gels. Following spot matching, a data matrix with one row for each protein, one column for each gel, and a normalized intensity as each entry is typically produced and used for further analysis. Such a data matrix typically contains missing values because it is not always possible to find a spot for each protein on each gel. Furthermore, there is no guarantee that all the values in a given row correspond to a single protein. Rather mismatches may occur so that the intensities recorded for a single protein across multiple gels might actually be a mixture of observations from two or more proteins. Despite these problems, no measures of uncertainty

*Department of Statistics, North Dakota State University, Fargo, ND, USA.
†Department of Statistics, Iowa State University, Ames, IA, USA.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

in the spot matching process are carried forward when making inferences about changes in protein abundance across conditions using the standard analysis methods. Instead, it is assumed that spots have been correctly matched across gels, and analysis proceeds with no accounting for the possibility of errors in the spot matching process.

We propose an alternative analysis strategy that explicitly acknowledges uncertainty in spot matching and incorporates that uncertainty into estimation and testing of differential protein quantity across two or more conditions. The location and intensity for a given protein across multiple aligned gels are modeled as trivariate Gaussian with a mean intensity that may depend on condition. Further, we assume that the true spot matching information across aligned gels is contained in unobserved random variables, and the Expectation Maximization (EM) algorithm is used to estimate model parameters and test for differences in intensity across conditions for each protein. The E-step in our EM algorithm is nontrivial in that a novel implementation of Markov chain Monte Carlo (MCMC) is used to estimate the expected value of the unobserved variables given the observed data and previous parameter estimates. Our method is formally described in Section 2. In simulation studies described in Section 3, we demonstrate the effectiveness of our MCMC approach. Data-based simulation in Section 4 illustrates the superiority of our procedure over a naïve analysis that ignores uncertainty in spot matching. Section 5 examines robustness of the procedure to violations of the normality assumption. The paper concludes with some discussion.

2 Methodology

2.1 Modeling

Let $\boldsymbol{X} = \{\boldsymbol{X}_{ijk} : i = 1, 2; j = 1, 2, \ldots, n; k = 1, 2, \ldots, m\}$ denote the complete dataset. Here, i indexes two treatments, n denotes the number of experimental units for each treatment, and m denotes the number of protein spots in each two-dimensional gel. We assume that there is one gel for each experimental unit. Each $\boldsymbol{X}_{ijk}$ represents the trivariate observation vector associated with the corresponding (i, j, k)th protein spot, with the three variates corresponding to its isoelectric point, its molecular weight, and its log-intensity (a quantitative measure of protein abundance), assumed without loss of generality to be in that order. If the identity of the protein associated with the (i, j, k)th spot is known to be protein ℓ, then $\boldsymbol{X}_{ijk}$ is assumed to be trivariate Gaussian with mean and dispersion given by $\boldsymbol{\mu}_{i\ell} = (\mu_{\ell 1}, \mu_{\ell 2}, \mu_{i\ell 3})'$ and $\boldsymbol{\Sigma}_{i\ell} = \mathrm{diag}(\sigma^2_{\ell 1}, \sigma^2_{\ell 2}, \sigma^2_{i\ell 3})$. Note that only the log-intensity mean and variance depend on the treatment i; therefore, we make the standard assumptions that protein isoelectric point, molecular weight, and associated variances are unaffected by treatment.

The identity of the protein is, of course, not usually known. Thus, when jointly modeling the data from any given gel, we must consider all possible mappings from the m spots to the m proteins. Without loss of generality, we assume that the spots are randomly labeled within each gel so that $1/m!$ is the prior probability of any spot-to-protein mapping. This prior probability should not be confused with the conditional probability of a spot-to-protein matching, given observed spot locations and intensities. These conditional probabilities are certainly not all equal to $1/m!$, and we account for this in our estimation procedure described in the next section.

We assume that each observed spot corresponds to exactly one protein in each gel and this correspondence is bijective. As a consequence, the observations within a gel cannot be assumed to be independent. We assume, however, that the observations are conditionally independent given the spot-to-protein matching. Our objective is to identify the proteins for which $H_{0\ell} : \mu_{1\ell 3} = \mu_{2\ell 3}$ is false; *i.e.*, we wish to identify proteins that differ in mean log-intensity across treatments.

2.2 Parameter Estimation

We reformulate the problem in terms of one with missing observations by postulating that the protein identity of each observed spot is the missing part of the dataset. A common approach to parameter estimation in such contexts is the EM algorithm pioneered by Dempster et al. (1977), which we adapt here for our purpose. To do so, we introduce unobserved integer-valued random variables $\{Z_{ijk} : i = 1, 2; j = 1, 2, \ldots, n; k = 1, 2, \ldots, m\}$ that contain the protein identity of each spot. In particular, $Z_{ijk} = \ell$ implies that the (i, j, k)th spot corresponds to protein ℓ. We use $\boldsymbol{Z}_{ij} = \boldsymbol{\ell}$ to denote the event $Z_{ij1} = \ell_1, Z_{ij2} = \ell_2, \ldots, Z_{ijm} = \ell_m$ where $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_m)' \in \rho(m)$, the set of all permutations of $1, 2, \ldots, m$.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

Note that the log likelihood function for the complete data is given by

$$l(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \sum_{i=1}^{2} \sum_{j=1}^{n} \sum_{\boldsymbol{\ell} \in \rho(m)} I(\boldsymbol{Z}_{ij} = \boldsymbol{\ell}) \sum_{k=1}^{m} \log \phi(\boldsymbol{X}_{ijk} | \boldsymbol{\mu}_{i\ell_k}, \boldsymbol{\Sigma}_{i\ell_k}),$$

where $\boldsymbol{Z} = \{Z_{ijk} : i = 1, 2; j = 1, 2, \ldots, n; k = 1, 2, \ldots, m\}$ is the set of missing identifiers, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_{i\ell} : i = 1, 2; \ell = 1, \ldots, m\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_{i\ell} : i = 1, 2; \ell = 1, \ldots, m\}$, and $\phi(\boldsymbol{x} | \boldsymbol{\mu}_{i\ell}, \boldsymbol{\Sigma}_{i\ell})$ denotes the trivariate normal density with mean $\boldsymbol{\mu}_{i\ell}$ and variance $\boldsymbol{\Sigma}_{i\ell}$ evaluated at $\boldsymbol{x}$. The expectation step of the EM algorithm involves constructing the expected value of the complete-data log likelihood function, conditional on the observed data. In this case the conditional expected value is simply the complete-data log likelihood with each $I(\boldsymbol{Z}_{ij} = \boldsymbol{\ell})$ replaced by

$$\mathbb{P}\mathrm{r}(\boldsymbol{Z}_{ij} = \boldsymbol{\ell} | \boldsymbol{X}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \frac{\frac{1}{m!} \prod_{k=1}^{m} \phi(\boldsymbol{X}_{ijk} | \boldsymbol{\mu}^*_{i\ell_k}, \boldsymbol{\Sigma}^*_{i\ell_k})}{\sum_{\boldsymbol{\ell} \in \rho(m)} \frac{1}{m!} \prod_{k=1}^{m} \phi(\boldsymbol{X}_{ijk} | \boldsymbol{\mu}^*_{i\ell_k}, \boldsymbol{\Sigma}^*_{i\ell_k})} = \frac{\prod_{k=1}^{m} \phi(\boldsymbol{X}_{ijk} | \boldsymbol{\mu}^*_{i\ell_k}, \boldsymbol{\Sigma}^*_{i\ell_k})}{\sum_{\boldsymbol{\ell} \in \rho(m)} \prod_{k=1}^{m} \phi(\boldsymbol{X}_{ijk} | \boldsymbol{\mu}^*_{i\ell_k}, \boldsymbol{\Sigma}^*_{i\ell_k})}, \quad (1)$$

where asterisks indicate the initial parameter estimates or estimates obtained from the previous maximization step of the EM algorithm. It is straightforward to show that the maximization step of the EM algorithm depends on (1) only through the weights

$$w_{ijk\ell} = \mathbb{P}\mathrm{r}(Z_{ijk} = \ell | \boldsymbol{X}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \sum_{\{\boldsymbol{\ell} \in \rho(m) : \ell_k = \ell\}} \mathbb{P}\mathrm{r}(\boldsymbol{Z}_{ij} = \boldsymbol{\ell} | \boldsymbol{X}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),$$

which (for $i = 1, 2$; $j = 1, \ldots, n$; $k = 1, \ldots, m$; and $\ell = 1, \ldots, m$) give the conditional probability that spot k on gel j from treatment group i matches protein ℓ. Given these weights, solutions that maximize the conditional expected value of the complete-data log likelihood function are

$$\begin{aligned}
\hat{\mu}_{\ell r} &= \frac{\sum_{i=1}^{2} \sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell} X_{ijkr}}{\sum_{i=1}^{2} \sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell}}, \quad r = 1, 2; \quad \hat{\sigma}^2_{\ell r} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell} (X_{ijkr} - \hat{\mu}_{\ell r})^2}{\sum_{i=1}^{2} \sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell}}, \quad r = 1, 2; \\
\hat{\mu}_{i\ell 3} &= \frac{\sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell} X_{ijk3}}{\sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell}}, \quad i = 1, 2; \quad \hat{\sigma}^2_{i\ell 3} = \frac{\sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell} (X_{ijk3} - \hat{\mu}_{i\ell 3})^2}{\sum_{j=1}^{n} \sum_{k=1}^{m} w_{ijk\ell}}, \quad i = 1, 2.
\end{aligned} \quad (2)$$

Maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained by iterating the E- and M-steps until convergence.

2.2.1 Computational Issues

One critical limitation for the procedure suggested above is that it is practical only for a small number of protein spots as finding the weights $w_{ijk\ell}$'s based on a complete enumeration of all possible permutations is computationally prohibitive. Indeed, in our experience with simulations, it becomes impractical to implement even when m is as low as 20. Thus, it can not be a viable strategy to implement for practical scenarios where two-dimensional gels typically have several hundred protein spots.

Many stochastic approaches to EM have been suggested to address intractability of calculations in the E-step. Celeux and Diebolt (1992) proposed approximating the expected complete log likelihood for the mixture problem, with results on convergence provided by Delyon et al. (1999). Even before that, Wei and Tanner (1990) suggested using Monte Carlo integration to approximate the conditional expectations in the E-step, with an automated version provided by Booth and Hobert (1999). When the Monte Carlo sampling is itself intractable, Levine and Casella (2001) and Levine and Fan (2004) suggested using Markov Chain Monte Carlo (MCMC). We adopt this idea here, but note that development of an MCMC approach is not straightforward. So we propose a MCMC algorithm that borrows ideas from the literature on conditional point process simulation (Baddeley and Møller, 1989). The algorithm that we now describe must be implemented separately for each gel. However, to simplify notation, we have suppressed the subscripts i and j throughout the remainder of this subsection. Let $\boldsymbol{Z}^* = (Z_1^*, Z_2^*, \ldots, Z_m^*)$ denote the set of current protein assignments for the gel under consideration. Pick a ν with probability m^{-1} from $\{1, 2, \ldots, m\}$. Compute

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

the distance from the Z_ν^*th protein spot to each of the other protein spots, where the distance between spots Z_ν^* and Z_u^* for $u \in \Omega_{-\nu} = \{1, ..., \nu-1, \nu+1, ..., m\}$ is denoted $d_{\nu,u}$ and is defined as the Euclidean distance between the current estimated mean locations associated with the Z_ν^*th and Z_u^*th proteins. Letting $H_\nu = \sum_u d_{\nu,u}^{-1}$, pick an ω with probability $d_{\nu,\omega}^{-1}/H_\nu$. Let $\tilde{\boldsymbol{Z}}$ be the same as $\boldsymbol{Z}^*$ except in the νth and ωth positions, which are reversed, *i.e.*, if $\boldsymbol{Z}^* = (Z_1^*, ..., Z_\nu^*, ..., Z_\omega^*, ..., Z_m^*)$, then $\tilde{\boldsymbol{Z}} = (Z_1^*, ..., Z_\omega^*, ..., Z_\nu^*, ..., Z_m^*)$. The latter is our proposal with Markov transition probability given by $\mathbb{P}\mathrm{r}(\boldsymbol{Z}^* \to \tilde{\boldsymbol{Z}}) = m^{-1} d_{\nu,\omega}^{-1} H_\nu^{-1}$.

Similar arguments mean that the reverse transition probability is given by $\mathbb{P}\mathrm{r}(\tilde{\boldsymbol{Z}} \to \boldsymbol{Z}^*) = m^{-1} d_{\omega,\nu}^{-1} H_\omega^{-1}$. Thus from the current state $\boldsymbol{Z}^*$, we accept $\tilde{\boldsymbol{Z}}$ with probability

$$\pi(\boldsymbol{Z}^* \to \tilde{\boldsymbol{Z}}) = \min\left\{1, \frac{\pi(\tilde{\boldsymbol{Z}}|\boldsymbol{X}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)}{\pi(\boldsymbol{Z}^*|\boldsymbol{X}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)} \frac{\mathbb{P}\mathrm{r}(\tilde{\boldsymbol{Z}} \to \boldsymbol{Z}^*)}{\mathbb{P}\mathrm{r}(\boldsymbol{Z}^* \to \tilde{\boldsymbol{Z}})}\right\} = \min\left\{1, \frac{H_\nu}{H_\omega} \prod_{k=1}^{m} \frac{\phi(\boldsymbol{X}_k|\boldsymbol{\mu}_{\tilde{Z}_k}^*, \boldsymbol{\Sigma}_{\tilde{Z}_k}^*)}{\phi(\boldsymbol{X}_k|\boldsymbol{\mu}_{Z_k^*}^*, \boldsymbol{\Sigma}_{Z_k^*}^*)}\right\},$$

otherwise we stay at the current $\boldsymbol{Z}^*$. We use this to collect 100,000 MCMC realizations, after a burn-in period of 10,000 iterations, and estimate our $w_{ijk\ell}$ values.

2.2.2 Initialization and Stopping Criteria

As with most iterative algorithms, the performance of the EM algorithm can be severely affected by the choice of initial values. Maitra (2009) provides detailed examples where choice of starting values can completely degrade performance in the context of estimating parameters of mixtures-of-Gaussians. His multi-stage algorithm for initialization does not extend to our situation easily, thus we chose a gel at random and used the observed isoelectric point and molecular weight values for that gel as the initializing means. As each particular gel contains only one set of intensities, we initialized the mean intensities under both treatment and control conditions to be the same. In order to obtain the initial values for variances, we considered isoelectric point values and molecular weight separately. For each of these variables, we estimated variances over all gels according to their order. For instance, we estimated the variance in molecular weight for the protein with the smallest molecular weight by calculating the variance of the proteins with the smallest molecular weights in each gel. The same approach was applied to the protein with the second smallest molecular weight in each gel, and so on. The variances for the isoelectric points were handled similarly. As intensity values were originally taken on a log-scale, the variability of the intensity values was small compared to the variances in isoelectric points and molecular weights. Thus, we set the initial values for intensity variances to be some small value (in our experiments, we set it equal to unity) relative to the estimated variances of isoelectric points and molecular weight. Although a more sophisticated approach may be possible to develop, our heuristic suggestion has been found to perform very well in our experiments.

Finding a stopping criterion for the EM algorithm within the framework of MCMC-based computation of probabilities $w_{ijk\ell}$ is challenging because the log likelihood of the observations can not be calculated at any step, for even a moderate number of proteins. Thus, a criterion based on relative changes in observed log likelihood is inapplicable. On the other hand, a criterion based on relative changes in parameter estimates as recommended by Altman et al. (2003) cannot reasonably be implemented because of simulation variability induced by the stochastic estimation of $w_{ijk\ell}$s, which will need to be accounted for in the calculation of the estimates. Therefore, we suggest stopping the EM algorithm after some reasonably large number of iterations that assures convergence. In all our experiments, fifty iterations (which we recommend) have been found to be more than adequate. Indeed, we note that all our experiments reported in this paper showed evidence of convergence in as few as ten iterations.

2.3 Variance Estimation

As mentioned by McLachlan and Peel (2000) and McLachlan and Krishnan (2008), it is usually complicated to obtain the variance-covariance matrix for maximum likelihood estimates produced by the EM algorithm. First, the observed information matrix has to be obtained and then inverted. One convenient way to approximate the observed information in the case of independent identically distributed random variables (as per the complete dataset) is to use the empirical information (Louis, 1982; McLachlan and Krishnan, 2008; McLachlan and Peel, 2000). This approach relies on obtaining all the derivatives of the expected complete data log likelihood function based on one observation. In our case, we have n independent gels under each of two treatment conditions. The observations themselves, however, are not independent, which violates an important assumption that underlies the simplification

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

in McLachlan and Peel (2000) and McLachlan and Krishnan (2008). Note, however, that although the hidden variables are dependent as discussed earlier, the dependence between any two observations is very weak when the number of proteins per gel is high. Since we typically have several hundred proteins in each gel, this approximation is reasonable: more importantly, it makes calculation of the information matrix practical because of its relatively simple form. Thus, under these assumptions, there are $2nm$ observations in the dataset and the total number of estimated parameters is $8m$ as there are 3 means and 3 variances for every protein spot and intensities have different means and variances under control and treatment conditions. Now, let $\boldsymbol{\vartheta}$ be the vector of all estimated parameters: $\big(\mu_{\ell 1}, \mu_{\ell 2}, \mu_{1\ell 3}, \mu_{2\ell 3}, \sigma^2_{\ell 1}, \sigma^2_{\ell 2}, \sigma^2_{1\ell 3}, \sigma^2_{2\ell 3} : \ell = 1, \ldots, m\big)$. To find the variance-covariance matrix of the estimated parameters, we invert the observed information matrix estimated by $\mathcal{I}_y = \sum_{i=1}^{2}\sum_{j=1}^{n}\sum_{k=1}^{m}\left(\frac{\partial h_{ijk}}{\partial \boldsymbol{\vartheta}}\right)\left(\frac{\partial h_{ijk}}{\partial \boldsymbol{\vartheta}}\right)'$, where h_{ijk} is the portion of the expected complete-data log likelihood involving the (i,j,k)th observation and $\frac{\partial h_{ijk}}{\partial \boldsymbol{\vartheta}}$ is the corresponding gradient vector with elements $\frac{\partial h_{ijk}}{\partial \mu_{\ell 1}}, \frac{\partial h_{ijk}}{\partial \mu_{\ell 2}}, \frac{\partial h_{ijk}}{\partial \mu_{1\ell 3}}, \frac{\partial h_{ijk}}{\partial \mu_{2\ell 3}}, \frac{\partial h_{ijk}}{\partial \sigma^2_{\ell 1}}, \frac{\partial h_{ijk}}{\partial \sigma^2_{\ell 2}}, \frac{\partial h_{ijk}}{\partial \sigma^2_{1\ell 3}}, \frac{\partial h_{ijk}}{\partial \sigma^2_{2\ell 3}}; \ell = 1, \ldots, m$. Then $\mathcal{I}_y$ is a $8m \times 8m$ matrix. The corresponding partial derivatives can be obtained as follows:

$$\frac{\partial h_{ijk}}{\partial \mu_{\ell r}} = w_{ijk\ell}\frac{x_{ijkr} - \mu_{\ell r}}{\sigma^2_{\ell r}},\ r = 1, 2,$$

$$\frac{\partial h_{ijk}}{\partial \mu_{i\ell 3}} = w_{ijk\ell}\frac{x_{ijk3} - \mu_{i\ell 3}}{\sigma^2_{i\ell 3}},$$

$$\frac{\partial h_{ijk}}{\partial \sigma^2_{\ell r}} = -\frac{1}{2}\frac{w_{ijk\ell}}{\sigma^2_{\ell r}}\left[1 - \frac{(x_{ijkr} - \mu_{\ell r})^2}{\sigma^2_{\ell r}}\right],\ r = 1, 2.$$

$$\frac{\partial h_{ijk}}{\partial \sigma^2_{i\ell 3}} = -\frac{1}{2}\frac{w_{ijk\ell}}{\sigma^2_{i\ell 3}}\left[1 - \frac{(x_{ijk3} - \mu_{i\ell 3})^2}{\sigma^2_{i\ell 3}}\right],$$

In order to obtain the variance-covariance matrix, we need to invert the observed information matrix $\mathcal{I}_y$. In general, this is not a major computational issue when the number of spots per gel vary from a few hundred to several thousand. For larger datasets however, this inversion is impractical to implement, so we propose to use the variance estimation approach of Meng and Rubin (1991).

3 Validation of the EM/MCMC Approach

The performance of the suggested procedure was evaluated via a set of simulation experiments designed to mimic realistic settings. Our proposed methodology has two aspects that need to be evaluated: (1) the use of the EM itself in the context of parameter estimation in the mixtures-of-Gaussians modeling scenario and (2) the use of MCMC to perform the E-step in the practical scenario of several hundreds of proteins. We evaluated these two aspects separately via two carefully-designed simulation suites. Our first suite involved a small-scale simulation experiment with two treatments and only six proteins per gel. For six proteins, we needed to consider exactly 720 possible permutations in the calculation of the $w_{ijk\ell}$s; thus E-step calculations are then exact and practical. Our second experimental suite performed simulations on a more realistic framework of 300 protein spots. In this case, exact calculations for the $w_{ijk\ell}$s are no longer possible, so were replaced by the MCMC-based stochastic computations. In both suites, we performed experiments for a range of scenarios reflecting varying estimation difficulty.

3.1 Simulation experiments with six protein spots

Estimation difficulty was specified in this suite via different levels of separation between the Gaussian mixture components. We used the notion of exact-c-separation introduced by Dasgupta (1999) and modified by Maitra (2009) to specify the degree of separation between these clusters. According to this definition, a mixture of p-variate Gaussian densities is c-separated if for every distinct pair (i, j) of p-variate mixture component densities, $c \leq \frac{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|}{\sqrt{p\max(\lambda_{\max}(\boldsymbol{\Sigma}_i), \lambda_{\max}(\boldsymbol{\Sigma}_j))}}$ with equality holding for at least some pair (i, j). From Dasgupta (1999), it follows that higher values of c correspond to better-separated mixture densities and hence relate to greater ease of estimation and class identification. We performed experiments for values of c corresponding to 0.5, 1.0 and 1.5. Thus,

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

Table 1: Comparison of all permutation (top line) and MCMC (bottom line) based methods

Par	1	2	3	4	5	6
Intensity I	1.865	4.582	-9.072	0.244	-7.859	0.535
	1.864	4.582	-9.071	0.243	-7.857	0.535
Intensity II	6.096	0.933	-8.716	-13.228	-0.115	-1.004
	6.096	0.933	-8.716	-13.227	-0.115	-1.004
Molecular weight	15.597	-15.082	-1.628	-1.777	8.464	-8.084
	15.596	-15.082	-1.627	-1.777	8.464	-8.083
Isoelectric point	9.867	-4.541	0.232	-3.818	-12.074	-4.456
	9.866	-4.541	0.232	-3.818	-12.074	-4.456

following Dasgupta (1999), our experiments covered a range of cases ranging from poor separation ($c = 0.5$) to moderate ($c = 1.0$) and good separation ($c = 1.5$). We also investigated the performance of our procedure in relation to the number of replications n. To do so, we performed our simulation experiments based on a range of $n = 3$, 10 and 100 gel replications. These sample sizes represent samples that are correspondingly small, moderate and large. By varying these two parameters (c and n), each over three settings, we investigated performance of our methodology over different scenarios. Figure 1 provides a summary display of experimental performance for the 9 different combinations of c and n. In each figure, black unfilled circles reflect the trivariate vectors of true means while red filled circles represent the obtained estimates. The location of each protein on a gel coincides with the corresponding centers of molecular weight and isoelectric point. The third dimension, mean intensity, is represented by the area of the corresponding circle chosen to be proportional to the actual intensities. The figure thus provides a visual assessment of estimation performance relative to the degree of overlap among the filled and unfilled circles. Thus, the ideal case when estimates are equal to estimated parameters would be represented by the plot with exactly overlapping circles (thus red filled circles with perfectly-fitted black borders). As can be seen, estimation performance with three gels was considerably poorer than with larger numbers of samples. Performance was especially degraded for $c = 0.5$ and $c = 1.0$, but more respectable for the case when $c = 1.5$. Performance in all three cases was much improved when $n = 10$, and was at least very good when $n = 100$. When $c = 1.5$ and $n = 100$ (Figure 1(i)), estimated and true mean values were virtually indistinguishable from each other. Thus, we see that there is some evidence of statistical consistency in the estimation procedure. Overall, the value of good separation turned out to be no less important than that of large replication sizes. In this context, we note that the figures can be categorized into groups (b,d), (c,e,g) and (h,f) separately in terms of estimation performance. Thus, the lack of separation, (alternatively increase in estimation difficulty) can be compensated by an increase in sample sizes. If marginal spot distributions are known *a priori* to be well-separated, one can do with a smaller number of gels. Otherwise, a larger number of replications is necessary.

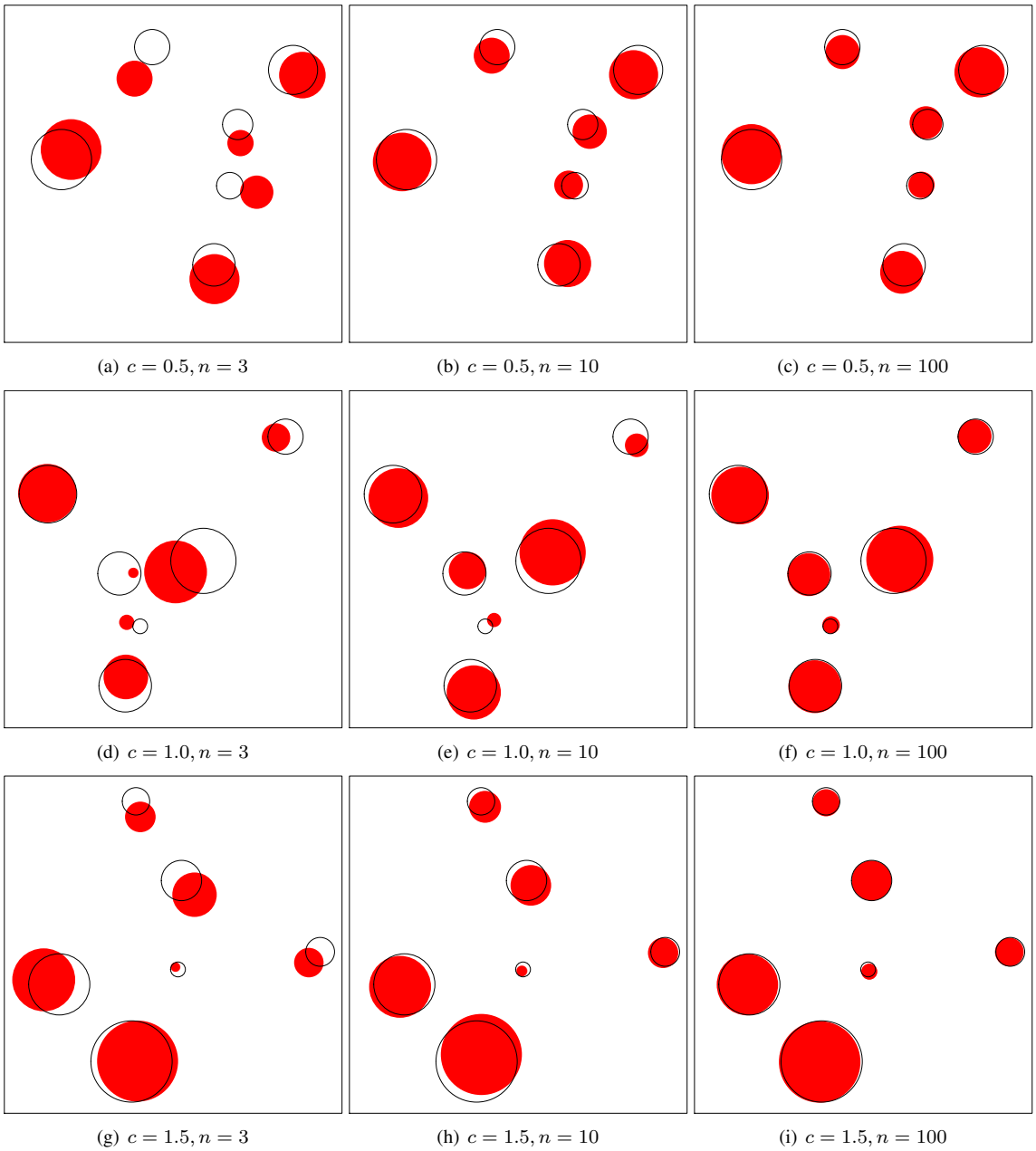
Our next performance evaluation was in the context of comparing the estimates obtained via the exact all-permutations-calculation-based E-step above with those obtained using a stochastic MCMC-based E-step. We report here performance based on moderate separation ($c = 1.0$) and moderate numbers of replications ($n = 10$). Table 1 reports the corresponding parameter estimates for all-permutation-based (top line) and MCMC-based (bottom line) algorithms. We see that both approaches produced nearly indistinguishable estimates. This demonstrates the validity of using MCMC-based stochastic calculations for estimating the probabilities in the E-step. This result is important for proceeding to the next step – testing the procedure on gels with a large number of protein spots, where exact calculations for the E-step are impractical to implement.

4 Evaluation on 2-D Gel Dataset

In this section, we evaluate our methodology on a real-life dataset and demonstrate its utility in analysis and inference relative to more traditional post-hoc ways of analyzing such data. We note, however, that spot locations for each gel are not reported by standard gel analysis software so that datasets with recorded gel-specific spot locations are not easy to come by. Therefore, we also use simulation to evaluate the performance of our methodology on a 2-D gel dataset used in the literature.

We consider a version of the morphine dataset described by Morris et al. (2008). Six adult male rats were assigned to treatment with either morphine or a placebo using a balanced and completely randomized design. Five days after

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.



(a) $c = 0.5, n = 3$ (b) $c = 0.5, n = 10$ (c) $c = 0.5, n = 100$

(d) $c = 1.0, n = 3$ (e) $c = 1.0, n = 10$ (f) $c = 1.0, n = 100$

(g) $c = 1.5, n = 3$ (h) $c = 1.5, n = 10$ (i) $c = 1.5, n = 100$

Figure 1: True (black unfilled circles) and estimated (filled red circles) means for the simulation experiment with six protein spots. In each case, area of the circles is proportional to the true and estimated mean intensities. Performance evaluations are for different numbers of gels and degree of separation, as indicated in the subfigure captions.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

initiation of treatment, proteins were extracted from spinal cord regions and evaluated on six gels with one gel per rat. For the sake of illustration, we focus here on the analysis of 300 protein spots that appear in one corner of each gel. Because the data were preprocessed using the method of Morris et al. (2008), the locations (isoelectric points and molecular weights) are not separately available for each individual gel. Rather one single location estimate is provided for each of the 300 proteins. Each location estimate is accompanied by six normalized intensities – one for each gel. The locations of the 300 spots are plotted as blue points in Figure 2 (a). In this figure, the shading of each blue spot indicates the intensity for the spot averaged across the three control gels. To illustrate the effectiveness of our proposed procedure for estimating mean protein locations and intensities, we simulated 300 trivariate normal observations with means and variances determined from the actual data. Specifically, the components of the mean vector for each protein were set equal to the spot location and average intensity indicated by the blue spots in Figure 2 (a). The variance of intensities for each protein was set equal to the sample variance of the observed intensities pooled over both treatments. Variances among isoelectric points and among molecular weights within a given protein were each set to be equal to 25 in order to provide spot location variation in our simulated data that visually matches the degree of variation seen in actual gels. Figure 2 (a) illustrates variability in such a simulated dataset with respect to the original means. Here, mean values are blue and observations are given in red. Figures 2 (b) and (c) show the estimates of the trivariate means (red) relative to the actual means (blue) using data from 10 and 30 simulated gels, respectively. When true means and estimates overlap, we use a mixture of colors as defined in the color key shown in Figure 2 (d). Here, 0 represents the lowest protein intensity while 1 stands for the highest. Ideally, we would see all spots, red and blue, overlapping and producing mixed colors from the diagonal of the key. Although the plot (b) suggests that 10 gels per treatment produce reasonably good parameter estimates, clearly, 30 gels per treatment match this ideal better, and therefore, estimates produced from 30 gels are substantially more accurate. Thus, our method remains reliable and effective not only for the small datasets investigated in Figure 1 but also in cases when we have hundreds of proteins.

4.1 Method's performance compared to t-test alternatives

In order to evaluate the overall performance of our procedure, we compare it with three t-test procedures. The first version of the t-test, $\mathcal{T}_{gold}$, uses the protein identity of each spot on each gel to perfectly match spots across gels. This method – which cannot be used in practice because spot identities are unknown in real data – serves as the gold-standard approach. The second t-test procedure, $\mathcal{T}_2$, mimics the type of analysis that is usually performed in current practice. First, spots are matched across gels according to their locations, and then inference proceeds with a two-sample t-test on intensities as if the spots have been matched perfectly. We refer to this as the naïve t-test procedure based on matched spot locations because the inferences ignore uncertainty in spot matching. The third and last t-test procedure, $\mathcal{T}_3$, is similar to the described naïve approach with the difference being in the matching procedure. Here, matching is conducted based on trivariate vectors that include spot intensities rather than using bivariate vectors describing locations. This is another naïve procedure which is conceptually more advanced than the naïve approach solely based on locations because it uses the intensity information for matching purposes. The spot-matching algorithm that we used to obtain this naïve t-test is described as follows. For each spot on each gel, our EM procedure produces an estimate of the probability that the spot matches each protein. We assign each spot to the protein with the highest match probability for that spot. This will generate a one-to-one match from many of the spots to many of the proteins. However, within any given gel, some spots may be assigned to the same protein and some proteins may have no spots assigned. Such spots and proteins are then handled with an exhaustive search algorithm that considers all possible assignments of the spots in question to the proteins in question. The assignment that produces the smallest sum of Euclidean distances between the locations of the spots and the estimated location means of the proteins to which the spots are assigned is selected to obtain the best match. The spot-matching algorithm for the naïve t-test based on locations has a similar scheme with the difference that the EM algorithm is based solely on bivariate vectors related to locations.

Note that although the naïve t-test procedure uses results from our EM algorithm to perform spot matching, the naïve approach is fundamentally quite different from our proposed analysis method. If the naïve t-test procedure were to obtain error free matching, it would be equivalent to the gold-standard approach. On the other hand, when spot matching errors occur, the naïve t-test will be based on a mixture of data from multiple proteins. Estimates of mean protein intensity for individual treatments and corresponding standard errors will often be adversely affected. Such errors can lead to incorrect inferences for mismatched proteins. Our proposed procedure avoids such problems by making no explicit attempt to match spots. Rather, our procedure estimates differences across treatments within proteins and assesses the significance of those differences using an inferential procedure that recognizes and accounts

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

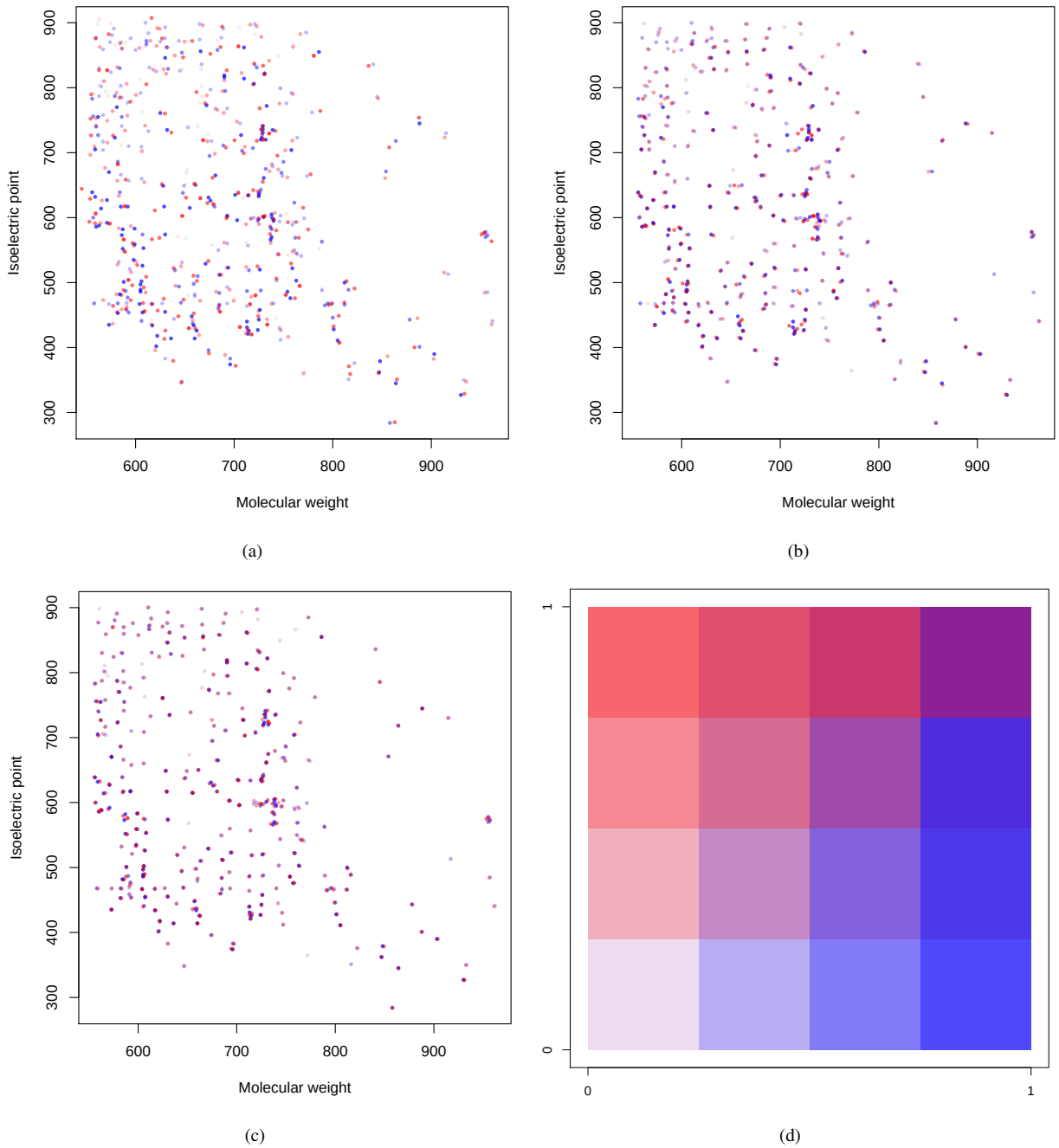


Figure 2: (a) Simulated dataset (in red) with respect to true means (in blue); (b) true means and their estimates based on 10 gels per treatment; (c) true means and their estimates based on 30 gels per treatment; (d) color key for intensities.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

for uncertainty in spot identity. Thus, we expect our procedure to outperform the naïve t-test whenever the probability of spot-matching errors is non-negligible.

To evaluate the performance of our proposed testing procedure relative to the gold-standard and naïve t-tests, we conducted a simulation study as follows. Using the observed intensity measures for the 300 spots from the control and morphine-treated rat gels, we estimated a standardized mean difference for each protein given by $\tau_\ell = (\bar{y}_{1\cdot\ell} - \bar{y}_{2\cdot\ell})/s_\ell$, where $y_{ij\ell}$ denotes the normalized intensity measure for treatment i, gel j, and protein ℓ and $s_\ell^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}(y_{ij\ell} - \bar{y}_{i\cdot\ell})^2/4$ for all $\ell = 1, \ldots, 300$. We considered three cases (P=0.25, 0.50 and 0.75) for the proportion of differentially expressed proteins. For $P = 0.25$, we randomly chose 75 spots from the $K = 75, 150$, or 225 spots with the largest $|\tau_\ell|$ values. For $P = 0.50$, we chose 150 spots out of the $K = 150$ or 225 highest $|\tau_\ell|$ values. In the case of $P = 0.75$, we took the 225 spots with largest $|\tau_\ell|$ values. For the chosen spots, we computed estimates of mean intensities separately for both treatment conditions, while for the other spots, the mean intensity was estimated as the average over both treatments. We used these estimates to simulate realistic measures of protein intensity for the ℓth protein. Protein spot locations were simulated based on the locations of the spots in the morphine data as described previously. By construction, the proportion of proteins whose mean intensity differs between our simulated treatment groups ($\mu_{1\ell 3} \neq \mu_{2\ell 3}$) is given by P.

For each choice of P and K, we simulated $n = 10$ and $n = 30$ gels per treatment and applied the gold-standard t-test, the naïve t-tests, and our proposed method to test the null hypothesis that mean intensities of the ℓth protein are the same under both treatment conditions; namely, $H_{0\ell} : \mu_{1\ell 3} = \mu_{2\ell 3}$ for $\ell = 1, \ldots, 300$. The results of these tests provided a rank ordering of the proteins from the most to the least significant for each method. We then computed the sensitivity and specificity of each method as functions of the number of proteins declared to be differentially expressed between simulated treatment groups.

The results, averaged over 10 independently simulated data sets for each value of P, are displayed as receiver operating characteristic (ROC) curves in Figures 3 and 4. Figure 3 illustrates performance of the methods over the entire range of $1-$specificity values. For practical reasons, we are most interested in the behavior of ROC curves for low values of $1-$specificity. Figure 4 presents ROC curves for type I error rates below 15%. In every plot, we include estimated ROC curves along with bounds formed by each pointwise estimate plus and minus one standard error.

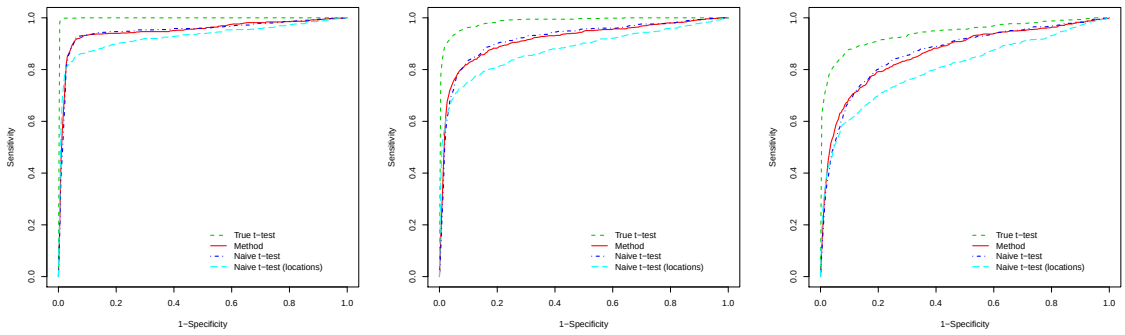
For $n = 10$ (the first two rows of Figures 3 and 4), the naïve t-test $\mathcal{T}_3$ with trivariate vector matching and our proposed method behave similarly. For the very lowest values of $1-$specificity, it appears (Figures 4(a)–(e)) that the t-test $\mathcal{T}_2$ with bivariate vector matching performs marginally better than our method, but in general, we outperform $\mathcal{T}_2$ for values of $1-$specificity greater than around 0.02. Although it may not be clear from the plotted curves, the advantage of our approach over $\mathcal{T}_2$ was very often statistically significant when $n = 10$. For instance, we note that when $1-$specificity is 5%, a pairwise t-test indicates that our proposed approach performed significantly better than $\mathcal{T}_2$ with p-values of less than 0.05 for all but the one case ($P = 0.5, K = 150, n = 10$). Although our approach often appears to hold a slight advantage over $\mathcal{T}_3$ when $n = 10$, the power of the proposed approach was significantly higher than that of $\mathcal{T}_3$ at $1-$specificity of 5% only for two cases ($P = 0.25, K = 150$ and $P = 0.50, K = 225$) with p-values of 0.002 and 0.046, respectively.

When $n = 30$ (the last two rows of Figures 3 and 4), we can see that our proposed method has higher power than both naïve t-tests. Furthermore, the improvement is generally highly significant. This additional sensitivity over both $\mathcal{T}_2$ and $\mathcal{T}_3$ is obtained by accounting for the uncertainty in spot-matching across gels which is an important source of variation that is "swept under the rug" by the naïve approach. Of course, our method falls short of the gold-standard t-test $\mathcal{T}_{gold}$ which utilizes information that is unavailable in practice to match spots perfectly across gels.

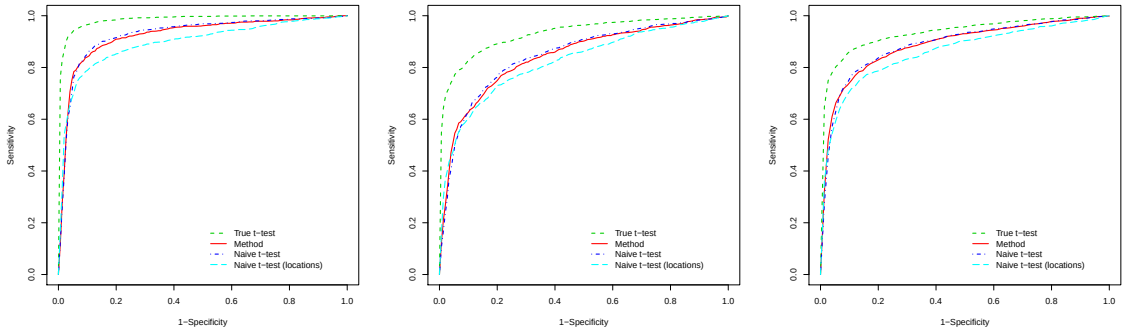
5 Robustness of the method against deviation from normality

The examples studied so far all assumed normal distributions for the spot locations and the log-intensities. A reviewer has asked about the robustness of our procedure to deviations from the normality assumption for the log-intensities. To investigate robustness, we considered several realizations obtained from mixtures with t-distributed components for the log-intensities. Figure 5 illustrates simulation examples with fifteen protein spots. The figures here are analogous to those in Figure 1. The data were simulated from t-mixtures in the following manner: as described in Section 3.1, we first simulated a trivariate Gaussian mixture with a c-separation of 1.0. Then, every normal mixture component for the log-intensity coordinate was replaced with its t_ν-distributed analogue, where ν denotes the degree of freedom of the t-component. Thus, instead of using $N(\mu, \sigma^2)$, we used $\mu + \sigma t_\nu$, where μ and σ are the corresponding mean and variance of the given component. Since a t-distribution has heavier tails than a normal distribution, the obtained t-

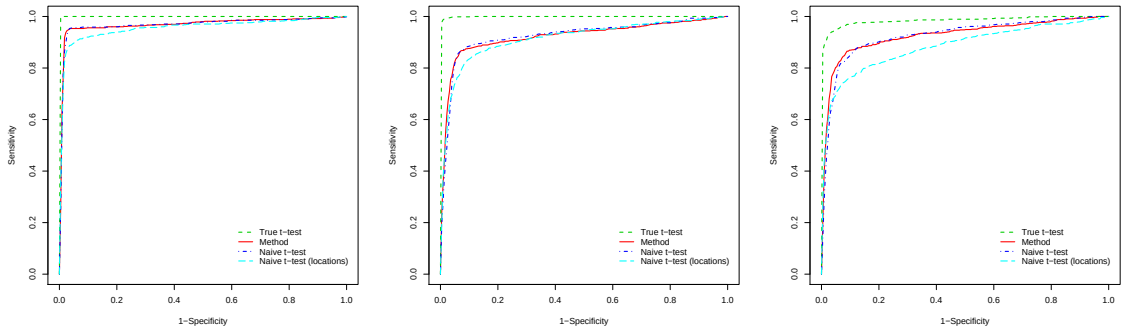
This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.



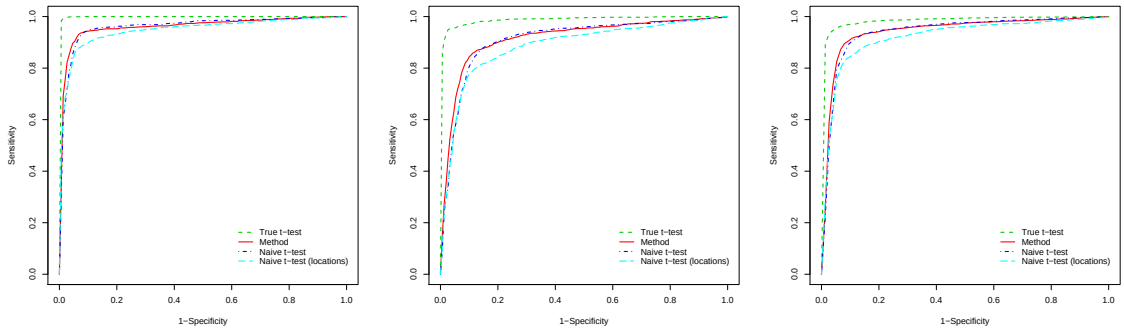
(a) $P = 0.25, K = 75, n = 10$ (b) $P = 0.25, K = 150, n = 10$ (c) $P = 0.25, K = 225, n = 10$



(d) $P = 0.50, K = 150, n = 10$ (e) $P = 0.50, K = 225, n = 10$ (f) $P = 0.75, K = 225, n = 10$



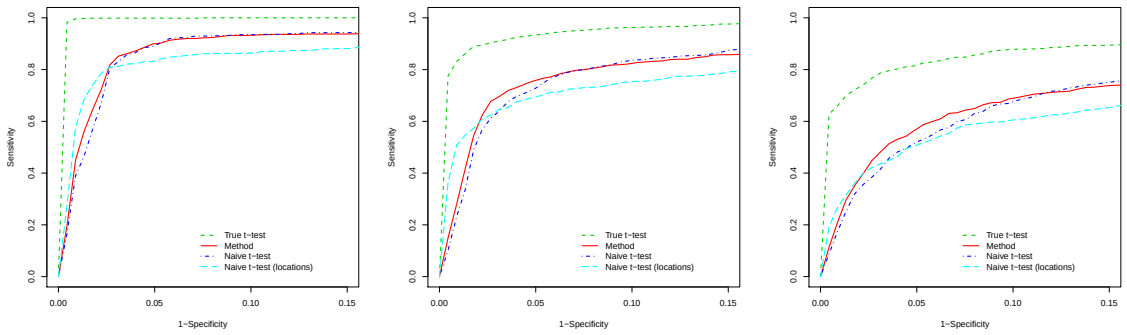
(g) $P = 0.25, K = 75, n = 30$ (h) $P = 0.25, K = 150, n = 30$ (i) $P = 0.25, K = 225, n = 30$



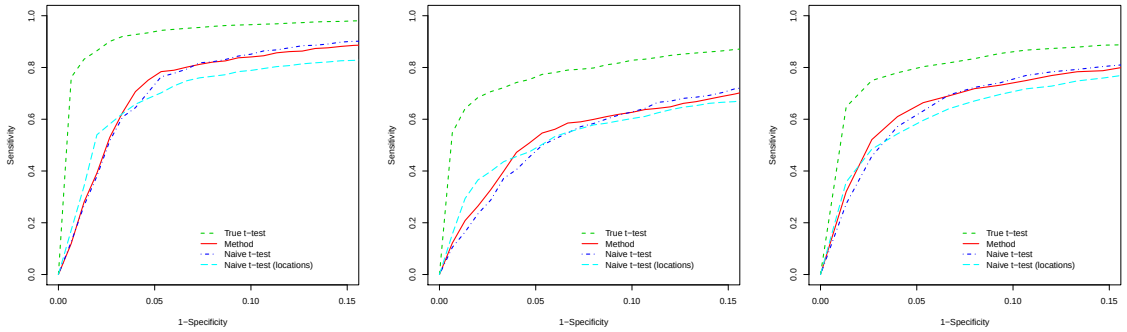
(j) $P = 0.50, K = 150, n = 30$ (k) $P = 0.50, K = 225, n = 30$ (l) $P = 0.75, K = 225, n = 30$

Figure 3: ROC curves for different settings of P, K and n.

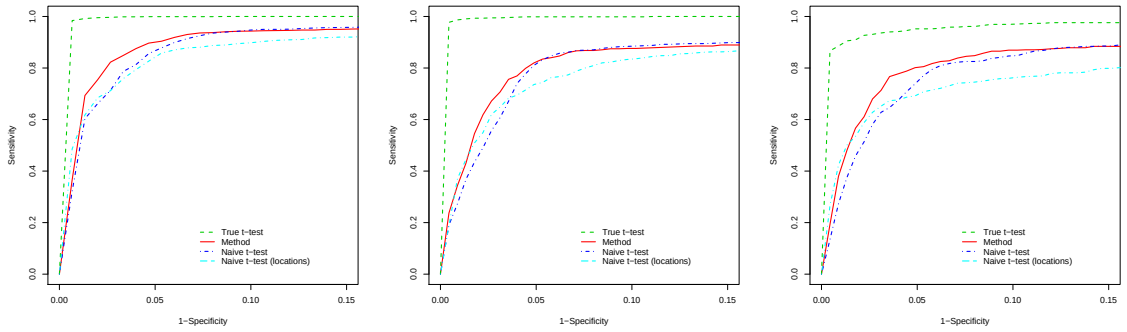
This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.



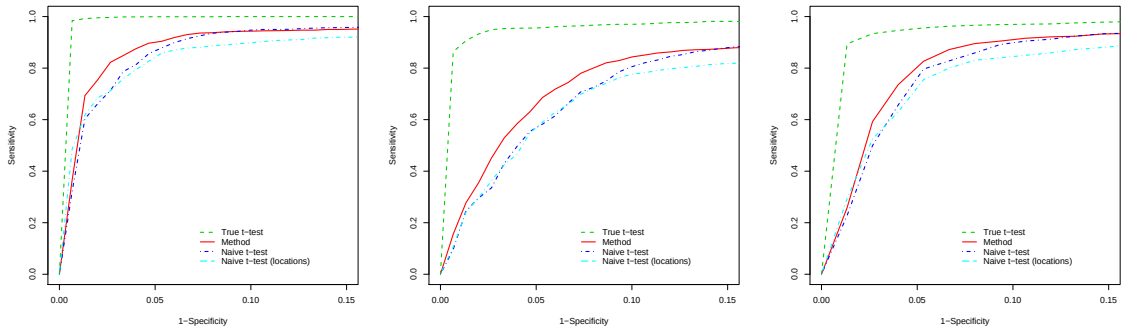
(a) $P = 0.25, K = 75, n = 10$ (b) $P = 0.25, K = 150, n = 10$ (c) $P = 0.25, K = 225, n = 10$



(d) $P = 0.50, K = 150, n = 10$ (e) $P = 0.50, K = 225, n = 10$ (f) $P = 0.75, K = 225, n = 10$



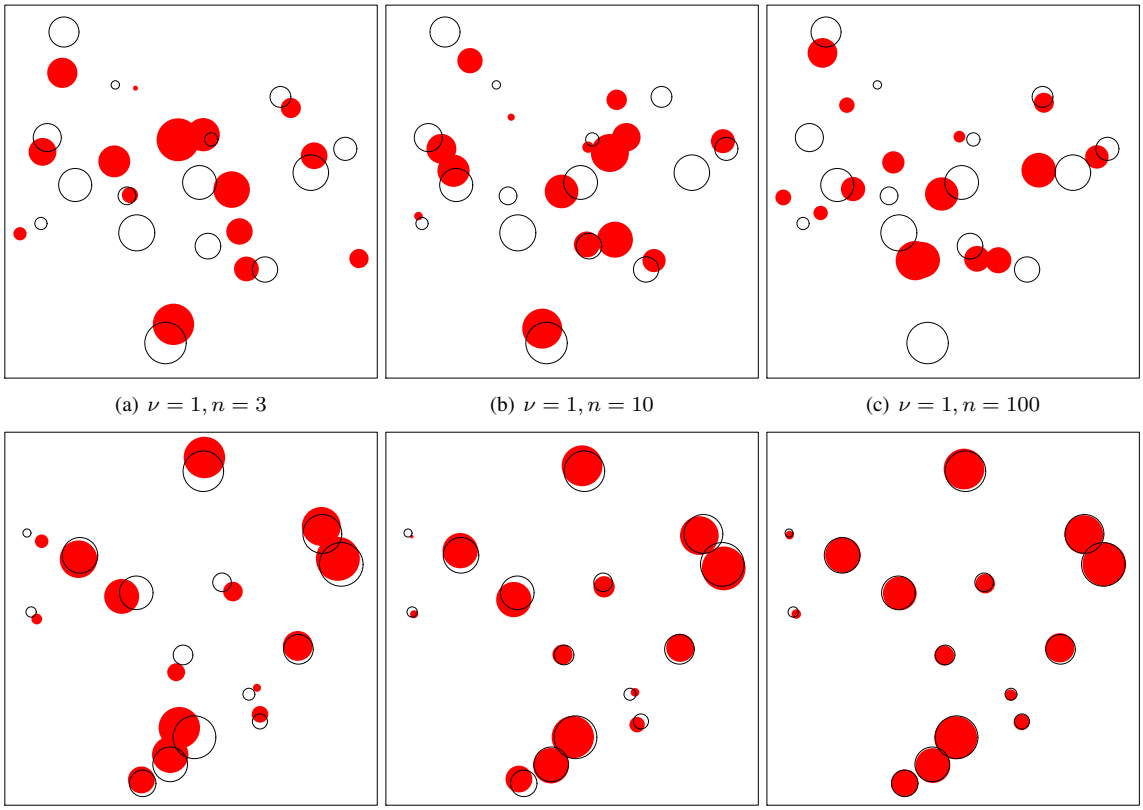
(g) $P = 0.25, K = 75, n = 30$ (h) $P = 0.25, K = 150, n = 30$ (i) $P = 0.25, K = 225, n = 30$



(j) $P = 0.50, K = 150, n = 30$ (k) $P = 0.50, K = 225, n = 30$ (l) $P = 0.75, K = 225, n = 30$

Figure 4: ROC curves for different settings of P, K and n (low values of 1 - specificity).

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.



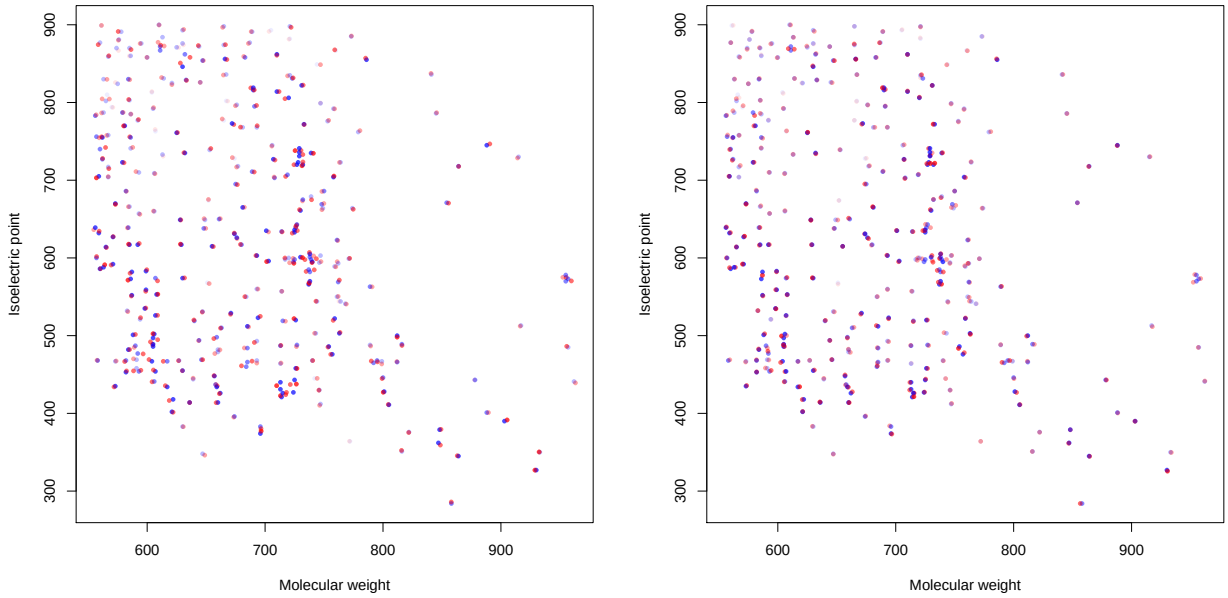
(a) $\nu = 1, n = 3$ (b) $\nu = 1, n = 10$ (c) $\nu = 1, n = 100$

(d) $\nu = 3, n = 3$ (e) $\nu = 3, n = 10$ (f) $\nu = 3, n = 100$

Figure 5: True (black unfilled circles) and estimated (filled red circles) means for the simulation experiment with fifteen protein spots, with different n and ν degrees of freedom for the t-distributions. Plotting characteristics are as in Figure 1.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

mixture components overlap more substantially than the normal mixture, with the degree of increased overlap governed by ν. (For all the experiments here, note that we used the procedure developed in Section 2 – which assumes normality – even though the data were generated from t_ν-mixtures.) When $\nu = 1$, the t-distribution has extremely heavy tails: indeed the mean of the distribution does not exist. Thus, it is not surprising that Figures 5a–c indicate that we are not able to estimate the parameters correctly no matter whether we consider sample sizes of $n = 3, 10$ or 100. However, even a small increase in the number of degrees of freedom improves the convergence results dramatically as indicated in Figures 5d–f with $\nu = 3$ (the smallest ν for which both mean and variance of the distribution exist). As seen, while it is not easy to estimate means based on three gels (Figure 5d), considerably better estimates can be obtained for $n = 10$ (Figure 5e), or for $n = 100$ Figure 5e) for which the results are clearly the best. Our small-scale investigation here is very encouraging because it suggests that our procedure is quite robust, performing well for cases with deviations from normality unless such deviations are very extreme as for $\nu = 1$, for which the mean parameters (to be estimated) do not even exist.



(a) (b)

Figure 6: (a) true means and their estimates based on 10 gels per treatment; (b) true means and their estimates based on 30 gels per treatment. Color intensities are as in Figure 2d.

We also checked robustness using the real life dataset analyzed in Section 4. Once again, we incorporated the simulation ideas described above to obtain realizations from t_3-distributed log-intensities under both treatment conditions. Figure 6 illustrates the obtained estimates with 10 and 30 gels (plots (a) and (b) correspondingly). In both cases, we obtain reasonable solutions. Expectedly, the solution for $n = 30$ is better than that for $n = 10$ as the red and blue circles overlap more producing colors from the main diagonal of the key diagram provided in Figure 2 (d). In summary, our investigations show that our proposed methodology is robust to deviations from the assumption of normal log-intensities unless these deviations are very extreme.

6 Conclusions

In this paper, a new method for analyzing gels from two-dimensional electrophoresis is suggested. We account for one of the major sources of variability – errors that happen during the spot-matching process. The EM algorithm naturally incorporates the uncertainty associated with these errors through posterior probabilities obtained at the E step. An estimated variance-covariance matrix of parameter estimates is then used to make inference about differentially expressed proteins. This approach is conceptually different from the commonly used methodology that assumes that the spot-matching algorithm is able to provide the ideal classification of spots. The proposed technique was tested on several synthetic datasets as well as on the set of real-data-based simulation studies. The results indicate that the

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.

suggested procedure typically outperforms approaches that ignore spot-matching uncertainty. This superior performance is maintained even when there are deviations from the normality of the log-intensities. Thus, we recommend that spot locations be routinely recorded for each gel so that analyses that account for uncertainty in spot matching can be implemented.

Acknowledgements

The authors acknowledge partial support by the National Science Foundation Awards NSF CAREER DMS-0437555, NSF IOS-0236060 and NSF DMS-0502347.

References

J. S. Almeida, R. Stanislaus, E. Krug, and J. M. Arthur. Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. *Proteomics*, 3:1567–1596, 2003.

M. Altman, J. Gill, and M. McDonald. *Numerical Issues in Statistical Computing for the Social Scientist*. Wiley-Interscience, New York, 2003.

A. J. Baddeley and J. Møller. Nearest-neighbour Markov point processes and random sets. *International Statistical Review*, 2:89–121, 1989.

J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, 61:265–285, 1999.

G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastic Reports*, 41:127–146, 1992.

S. Dasgupta. Learning mixtures of Gaussians. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 633–644, New York, 1999.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation of the EM algorithm. *The Annals of Statistics*, 27:94–128, 1999.

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

A. Dowsey, M. J. Dunn, and G. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, 3:1567–1596, 2003.

P. J. Green and K. V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254, 2006.

R. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10:422–439, 2001.

R. Levine and J. Fan. An automated (Markov Chain) Monte Carlo algorithm. *Journal of Statistical Computation and Simulation*, 74:349–359, 2004.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society, B*, 44:226–233, 1982.

R. Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:144–157, 2009. doi: http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70244.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 2008.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons, Inc., New York, 2000.

X. L. Meng and D. B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86:899–909, 1991.

J. S. Morris, B. N. Clark, and H. B. Gutstein. Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. *Bioinformatics*, 24:529–536, 2008.

P. M. Palagi, P. Hernandez, D. Walther, and R. D. Appel. Proteome informatics I: Bioinformatics tools for processing experimental data. *Proteomics*, 6:5435–5444, 2006.

A. Roy, F. Seillier-Moiseiwitsch, K. Lee, Y. Hang, M. R. Marten, and B. Raman. Analyzing two-dimensional gel images. *Chance*, 16:13–18, 2003.

G. C. J. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

This is a manuscript of an article from Sankhya B 73 (2011): 123, doi: 10.1007/s13571-011-0016-x. The final publication is available at Springer via http://dx.doi.org/10.1007/s13571-011-0016-x.