Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2009 *Paper* 61

Composite Likelihood EM Algorithm with Applications to Multivariate Hidden Markov Model

Xin Gao* Peter Xuekun Song†

*University of Michigan, xingao@umich.edu

†University of Michigan, Ann Arbor, pxsong@umich.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/cobra/art61

Copyright ©2009 by the authors.

Composite Likelihood EM Algorithm with Applications to Multivariate Hidden Markov Model

Xin Gao and Peter Xuekun Song

Abstract

The method of composite likelihood is useful to deal with estimation and inference in parametric models with high-dimensional data, where the full likelihood approach renders to intractable computational complexity. We develop an extension of the EM algorithm in the framework of composite likelihood estimation in the presence of missing data or latent variables. We establish three key theoretical properties of the composite likelihood EM (CLEM) algorithm, including the ascent property, the algorithmic convergence and the convergence rate. The proposed method is applied to estimate the transition probabilities in multivariate hidden Markov model. Simulation studies are presented to demonstrate the empirical performance of the method. A time-course microarray data is analyzed using the proposed CLEM method to dissect the underlying gene regulatory network.

Composite Likelihood EM Algorithm with Applications to Multivariate Hidden Markov Model *

XIN GAO and PETER X.-K. SONG

ABSTRACT

The method of composite likelihood is useful to deal with estimation and inference in parametric models with high-dimensional data, where the full likelihood approach renders to intractable computational complexity. We develop an extension of the EM algorithm in the framework of composite likelihood estimation in the presence of missing data or latent variables. We establish three key theoretical properties of the composite likelihood EM (CLEM) algorithm, including the ascent property, the algorithmic convergence and the convergence rate. The proposed method is applied to estimate the transition probabilities in multivariate hidden Markov model. Simulation studies are presented to demonstrate the empirical performance of the method. A time-course microarray data is analyzed using the proposed CLEM method to dissect the underlying gene regulatory network.

KEY WORDS: Composite likelihood; EM algorithm; Hidden Markov model; Latent variables; Time-course microarray data

1 INTRODUCTION

This paper focuses on the development of statistical theory and method of the EM algorithm in the context of composite likelihood (CL) for analyzing incomplete high-dimensional correlated data. The CL paradigm (*e.g.* Lindsay, 1988) helps to make statistical estimation and inference via dimension reduction, in the sense that a pseudo likelihood is constructed with the utility of low dimensional likelihood objects. It is particularly appealing in dealing with data with high-dimensional response variables. High-dimensionality in the response variables appear in many practical studies, such as a genetic pathway analysis involving gene regulatory networks and longitudinal cohort studies involving space-time mea-

*X. Gao is Associate Professor, Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada M3J 1P3. (Email: *xingao@mathstat.yorku.ca*). P. Song is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, (Email: *pxsong@umich.edu*).

Hosted by The Berkeley Electronic Press

surements. A significant difficulty in parameter estimation with high-dimensional data via Fisher's full likelihood approach is computational feasibility. At many occasions, the likelihood function is too complex to be numerically manageable. The CL method pertains to a compromise between the estimation efficiency and computational ease. That is, a high-dimensional full likelihood is simplified to several low dimensional pseudo-likelihoods for the benefit of computing. On the other hand, this simplification is at the cost of some efficiency loss.

1.1 Composite Likelihood Methodology

The history of the CL method is relatively short, and it has drawn much attention in recent years. This method has been successfully applied in many areas, including generalized linear mixed models (Renard et al., 2004), statistical genetics (Fearnhead and Donnely, 2002), spatial statistics (Hjort and Omre, 1994; Heagerty and Lele, 1998; Varin et al., 2005), multivariate survival analysis (Parner, 2001), and high-dimensional data (Fieuws and Verbeke, 2006; Faes, et al., 2008) among others. It has demonstrated to possess good theoretical properties, such as consistency for the parameter estimation, and can be utilized to establish hypothesis testing procedures.

This general formulation of composite likelihood comprises two main types. The first type is the omission method, which forms the composite likelihood by removing some terms in the full likelihood to simplify the evaluation. This includes Besag pseudolikelihood (Besag, 1974, 1977), the m-order likelihood for stationary processes (Azzalini, 1983), and the approximate likelihood (Stein, 2004), among others. The removed terms are chosen so that they contain little information about the parameter of interest, and the loss of efficiency compared to the full likelihood method is tolerable. The other type includes pseudolikelihood constructed from lower dimensional densities (Cox and Reid, 2004), which is the focus of this paper.

We begin the discussion of the second type with some necessary notation. Let $\mathbf{z} = (z_1, \ldots, z_n)^T$ be the vector of n variables observed from a single unit. Let $\{f(\mathbf{z}; \boldsymbol{\psi}), \mathbf{z} \in \mathcal{Z}, \boldsymbol{\psi} \in \Psi\}$ be a class of parametric models, with $\mathcal{Z} \subseteq \mathcal{R}^n$, $\Psi \subseteq \mathcal{R}^q$, $n \geq 1$, and $q \geq 1$. For a subset of $\{1, \ldots, n\}$, say a, $\mathbf{z}_a$ denotes a subvector of $\mathbf{z}$ with components indexed by the elements in set a. For instance, given a set $a = \{1, 2\}$, $\mathbf{z}_a = (z_1, z_2)^T$. Let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^p$, $p \leq q$, is the parameter of interest, and $\boldsymbol{\eta}$ is the nuisance parameter. According to Lindsay (1988), the CL of a single vector-valued observation is $L_c(\boldsymbol{\theta}; \mathbf{z}) = \prod_{a \in A} L_a(\boldsymbol{\theta}; \mathbf{z}_a)^{w_a}$, where A is a collection of index subsets called the composite sets, $L_a(\boldsymbol{\theta}; \mathbf{z}_a) = f_a(\mathbf{z}_a; \boldsymbol{\theta}_a)$, and $\{w_a, a \in A\}$ is a set of positive weights. Here f_a denotes all the different marginal densities and θ_a indicates the parameters that are identifiable in the marginal density f_a. Later in this article, the subscripts of f_a and

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

θ_a are omitted for notational simplicity. The weights w_a are positive to ensure the ascent property of the proposed CLEM algorithm discussed later.

For example, the independence CL can be formulated as a product of one-dimensional marginal likelihood objectives, namely $L_c = \prod_{a\in A}^{n} f(\mathbf{z}_a;\boldsymbol{\theta})^{w_a}$, with $A = \{1,\ldots,n\}$, and $\mathbf{z}_a, a \in A$, denotes a single variable indexed by the element in a. Likewise, the pairwise CL takes the production of all possible two-dimensional marginal likelihoods, where $A = \{\{1,2\},\{1,3\},\ldots,\{n-1,n\}\}$ is apparently the collection of all indices for pairs. Both independence CL and pairwise CL can be combined in some optimal way to ensure the satisfactory asymptotic properties of the resulting estimator (Cox and Reid, 2004). It is important to note that the choice of margins to use depends on the identifiability of the model parameters. If some parameters of interest cannot be identified from bivariate margins, then the composite sets in the CL formulation may have to contain three- or even higher dimensional margins.

The fundamental argument for the CL method to work lies on the theory of estimating functions (Song 2007, Chapter 3). Under the assumption that the true parameter $\boldsymbol{\theta}_0$ belongs to the interior of a compact parameter space, the maximum composite likelihood estimator solves the composite score functions,

$$\sum_{a\in A} w_a \frac{\partial \log f(\mathbf{z}_a;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1}$$

As the composite score function is a linear combination of several valid likelihood score functions, it is unbiased under the usual regularity conditions. Therefore, even though the composite likelihood is not a real likelihood, the maximum composite likelihood estimate is still consistent for the true parameter. The asymptotic covariance matrix of the maximum composite likelihood estimator takes the form of the inverse of the Godambe information (Godambe, 1960):

$$H(\boldsymbol{\theta})^T J(\boldsymbol{\theta})^{-1} H(\boldsymbol{\theta}),$$

where $H(\boldsymbol{\theta}) = E\{-\sum_{a\in A} \partial^2 \log f(\mathbf{z}_a;\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\}$ and $J(\boldsymbol{\theta}) = \mathrm{var}\{\sum_{a\in A} \partial \log f(\mathbf{z}_a;\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}$ are the sensitivity matrix and the variability matrix, respectively. When the composite set takes the only full set of indices, namely $A = \{\{1,\ldots,n\}\}$, the CL is the same as the ordinary full likelihood, and the Godambe information reduces to the usual Fisher information, because $H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ in this case. But when A contains multiple composite sets, the difference between the Fisher and Godambe information is always positive semi-definite, which quantifies the efficiency loss incurred by using the CL instead of the full likelihood.

Concerning the asymptotic behavior of the resulting estimator, we distinguish between two scenarios. One is the usual setting when the sample size tends to infinity, under which the maximum composite likelihood estimator is asymptotically normally distributed under the usual regularity conditions. The other is that there are few replications but the data is of long sequence, such as spatio-temporal data. Under this

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

situation, the maximum CL estimator may or may not be consistent depending on the interdependency structure of the data. For example, for stationary time series and spatial processes with good mixing properties, the autocorrelation function follows an exponential decay and the maximum composite likelihood estimator retains the good properties. Whereas for the stationary time series with long range dependence, the convergence of the maximum composite likelihood estimator may be slow or even fail. Readers are referred to Varin (2008) and Cox and Reid (2004) for a detailed discussion.

1.2 EM Algorithm in Non-standard Settings

In practical applications, missing data further complicates the analysis of high dimensional correlated data. The traditional EM algorithm plays an important role in the full MLE with missing data. The procedure iterates between the E step, in which the expected log likelihood of the complete data is computed conditionally on the observed data, and the M step, in which the expected log likelihood of complete data is maximized to update the parameter estimate. However, to naively apply the EM strategy in high dimensional setting, we will encounter inevitable difficulties of solving the expectation step conditionally on the high-dimensional observed data. This often involves high dimensional integrals that are hard to evaluate. Thus a modified EM algorithm, which is computationally less intensive is desired for the composite likelihood inference in the presence of missing data. We anticipate the composite likelihood EM (CLEM) algorithm developed in this paper will provide a fundamental tool to the analysis of high-dimensional data with missing observations. We intend to provide a thorough investigation on the EM algorithm in the CL framework. Extending the EM algorithm to non-standard likelihood settings has been considered by many researchers, such as McLachlan and Krishnan (2008) and more references therein. Some simple versions of the CLEM algorithm have been scattered in the literature, e.g. Liang and Yu (2003) who named it as the pseudo EM algorithm and Varin et al. (2005) based on the pairwise CL. There is a clear need of developing a general CLEM algorithm methodology based on arbitrary sizes of composite sets, in order to deal with a wide range of high-dimensional data types. For instance, in the analysis of familial data of genetic copy number variations (e.g. Wang et al., 2007), it appears more desirable to form the CL based on nuclear families of trios (i.e., two parents and one offspring), as a trio pertains to a full inheritance core in a pedigree than a pair. Another example is the spatio-temporal data analysis, where in order to model the spatio-temporal interactions, quadruplets seem to be the minimal elementary set in the formulation of the CL.

This article aims to establish several key theoretical properties of the CLEM algorithm, including the ascent property, the algorithmic convergence, and the rate of convergence. We apply the CLEM algorithm

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

in the construction of gene networks with time-course microarray data based on multivariate hidden Markov models, in which the related computational complexity prohibits us from using the full likelihood EM (FLEM) algorithm.

The paper is organized as follows. Section 2 presents the CLEM algorithm and its properties. Section 3 discusses the application of CLEM in the multivariate hidden Markov model. Simulation studies on a three-variate and a 21-variate hidden Markov models are presented. Section 4 is devoted to a data analysis example of gene network construction, and Section 5 gives some concluding remarks. All the technical proofs are included in the appendix.

2 COMPOSITE LIKELIHOOD EM ALGORITHM

In many practical settings, we observe incomplete data. Assume under the composite likelihood framework, for each composite set a, there exists a many-to-one mapping $\mathbf{z}_a \to \mathbf{y}_a(\mathbf{z}_a)$ from $\mathcal{Z}_a$ to $\mathcal{Y}_a$, where $\mathcal{Z}_a$ and $\mathcal{Y}_a$ denote the sample spaces. Instead of observing the complete data $\mathbf{z}_a$, we observe the incomplete data $\mathbf{y}_a$. Let the the full set of the incomplete data be denoted as $\mathbf{y} = (\mathbf{y}_a, a \in A)$. Then, the observed CL is given by $L_c^o(\boldsymbol{\theta}; \mathbf{y}) = \prod_{a \in A} L_a^o(\boldsymbol{\theta}; \mathbf{y}_a)^{w_a}$ with $L_a^o(\boldsymbol{\theta}; \mathbf{y}_a) = \int_{\mathcal{Z}_a(\mathbf{y}_a)} f(\mathbf{z}_a; \boldsymbol{\theta}) d\mathbf{z}_a$, where $\mathcal{Z}_a(\mathbf{y}_a) = \{\mathbf{z}_a : \mathbf{y}_a = y_a(\mathbf{z}_a)\}$, which is the subset of $\mathcal{Z}_a$ determined by the equation $\mathbf{y}_a = y_a(\mathbf{z}_a)$.

Our goal is to develop a CL version EM (CLEM) algorithm that can produce the maximum CL estimation of the model parameter $\boldsymbol{\theta}$ in the presence of missing data. Suppose the CLEM algorithm has completed the $(r-1)$-th iteration and produced an update $\boldsymbol{\theta}^{(r-1)}$. Now at the r-th iteration, the CL E-step for a single vector-valued observation takes the form

$$Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = \sum_{a \in A} w_a \int_{\mathcal{Z}_a(\mathbf{y}_a)} \log L(\mathbf{z}_a; \boldsymbol{\theta}) f(\mathbf{z}_a|\mathbf{y}_a, \boldsymbol{\theta}^{(r-1)}) d\mathbf{z}_a. \tag{2}$$

When applied to data analysis, obviously, the Q_c will take an additional summation over the sample replicates.

It is worth noting that in the calculation of the Q_c function, we propose to replace the full set of observed data $\mathbf{y}$ by a subset-specific observed data $\mathbf{y}_a$ in the conditional part in order to make related computations feasible. This leads to a further dimension reduction, in addition to the previous one taken in the formulation of the CL.

Then, the proposed CLEM algorithm iterates the following E-step and M-step until convergence.

- **CL-E Step:** Given the previous update $\boldsymbol{\theta}^{(r-1)}$, obtain the expected composite likelihood $Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$;
- **CL-M Step:** Maximize $Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$ with respect to $\boldsymbol{\theta}$ to produce an update $\boldsymbol{\theta}^{(r)}$.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

2.1 Main properties

To justify the proposed CLEM algorithm, we investigate the following three key properties, similar to those in the establishment of the full likelihood EM (FLEM) algorithm. They are, (i) the proposed CLEM algorithm retains the ascent property; (ii) it is a fixed point algorithm converging to a stationary point; and (iii) the convergence rate of the CLEM depends on the curvature of the CL function surface.

We proceed our justification in the following sequence of steps. All the technical details of the proofs are listed in the appendix. First, for each subset index $a \in A$, we define a conditional density of $\mathbf{z}_a$ on $\mathbf{y}_a$:

$$f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta}) = \frac{f(\mathbf{z}_a;\boldsymbol{\theta})}{\int_{\mathcal{Z}_a(\mathbf{y}_a)} f(\mathbf{z}'_a;\boldsymbol{\theta})d\mathbf{z}'_a}, \tag{3}$$

where the denominator is the likelihood of the observed data $\mathbf{y}_a$, namely L^o_a. Define a CL version H-function as follow:

$$H_c(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \sum_{a\in A} w_a \int \log f(\mathbf{z}_a|\mathbf{y}_a;\tilde{\boldsymbol{\theta}}) f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta})d\mathbf{z}_a.$$

Then, we obtain an inequality stated in Lemma 1 below, which is crucial to establish the ascent property.

Lemma 1 *For any pair of* $(\boldsymbol{\theta}', \boldsymbol{\theta})$ *in* $\Theta \times \Theta$, $H_c(\boldsymbol{\theta}'|\boldsymbol{\theta}) \leq H_c(\boldsymbol{\theta}|\boldsymbol{\theta})$.

Moreover, Theorem 1 states that the CLEM algorithm satisfies the ascent property.

Theorem 1 *The composite log-likelihood of the observed data* $\mathbf{y}$, $l^o_c(\boldsymbol{\theta};\mathbf{y}) = \log L^o_c(\boldsymbol{\theta};\mathbf{y})$, *is nondecreasing over the sequence of updated estimates* $\boldsymbol{\theta}^{(r)}, r = 1, \ldots$; *that is,* $l^o_c(\boldsymbol{\theta}^{(r)};\mathbf{y}) \geq l^o_c(\boldsymbol{\theta}^{(r-1)};\mathbf{y})$.

Second, we present the sufficient conditions under which any limit points of any instance of the CLEM updates $\boldsymbol{\theta}^{(r)}$ are stationary points, and $\log L^o_c(\boldsymbol{\theta}^{(r)};\mathbf{y})$ converges monotonically to $\log L^o_c(\boldsymbol{\theta}^*;\mathbf{y})$ for some stationary point $\boldsymbol{\theta}^*$. For a bivariate function $f(u,v)$, let $\nabla^{(ij)}f(u,v)$ denote the i-th and j-th derivatives with respective to u and v.

Lemma 2 *Under the regularity condition that the order of differentiation and expectation can be exchanged, for all* $\boldsymbol{\theta} \in \Theta$,

(a) $\nabla^{(10)}H_c(\boldsymbol{\theta}|\boldsymbol{\theta}) = 0$, *and*

(b) $\nabla^{(11)}H_c(\boldsymbol{\theta}|\boldsymbol{\theta}) = \sum_{a\in A} w_a var\left\{\frac{\partial \log f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}|\mathbf{y}_a;\boldsymbol{\theta}\right\}$, *where* $f(\cdot)$ *is given in (3).*

Theorem 2 states that the CLEM algorithm is a fixed point algorithm converging to a stationary point of the observed composite likelihood surface.

Theorem 2 *Assume the following regularity conditions:*

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

(i) $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : L_c^o(\boldsymbol{\theta}; \mathbf{y}) \geq L_c^o(\boldsymbol{\theta}_0; \mathbf{y})\}$ *is compact for any* $\boldsymbol{\theta}_0$ *satisfying* $L_c^o(\boldsymbol{\theta}_0; \mathbf{y}) > -\infty$*;*

(ii) $L_c^o(\boldsymbol{\theta}; \cdot)$ *is continuous in* Θ *and differentiable in the interior of* Θ*; and*

(iii) the function $Q_c(\boldsymbol{\theta}'|\boldsymbol{\theta})$ *in (2) is smooth in both* $\boldsymbol{\theta}'$ *and* $\boldsymbol{\theta}$*.*

Then, all the limit points of any instance of the CLEM algorithm $\{\boldsymbol{\theta}^{(r)}\}$ *are stationary points, and* $L_c^o(\boldsymbol{\theta}^{(r)}; \mathbf{y})$ *converges monotonically to* $L_c^o(\boldsymbol{\theta}^*; \mathbf{y})$ *for some stationary point* $\boldsymbol{\theta}^*$*.*

Third, we investigate which factors affect the convergence rate of the CLEM algorithm. This would provide useful insights to improve the algorithmic speed. Theorem 3 presents our findings.

Theorem 3 *Suppose the regularity conditions (i)-(iii) stated in Theorem 2 hold. In addition, assume that*

(i) an instance of the CLEM algorithm $\boldsymbol{\theta}^{(r)}$*,* $r = 0, 1, 2, \ldots$*, converges to* $\boldsymbol{\theta}^*$ *in the closure of* Θ*, and*

(ii) $\nabla^{(20)} Q_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)})$ *is negative definite with eigenvalues bounded away from zero.*

Then the $\boldsymbol{\theta}^*$ *is a stationary point. Moreover, let* $\mathbf{M}(\boldsymbol{\theta}^*) = -\left\{\nabla^{(11)} H_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)\right\}\left\{\nabla^{(20)} Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)\right\}^{-1}$*. Then, the convergence rate of the CLEM equals to the* $\mathbf{M}$ *for a scalar parameter or equals to the largest eigenvalue of the* $\mathbf{M}$ *for a parameter vector.*

When $\boldsymbol{\theta}$ is a scalar, it is easy to see that the convergence rate is proportional to the information due to the missing data, $I_{a,mis}(\theta^*)$ and anti-proportional to the information due to the complete data, $I_{a,com}(\theta^*)$, in the form of

$$\mathbf{M}(\theta^*) = \left\{\sum_{a \in A} w_a I_{a,mis}(\theta^*)\right\}\left\{\sum_{a \in A} w_a I_{a,com}(\theta^*)\right\}^{-1}, \tag{4}$$

where

$$\begin{aligned} I_{a,mis}(\theta) &= \mathrm{E}\left\{-\frac{\partial^2 \log f(\mathbf{z}_a|\mathbf{y}_a;\theta)}{\partial\theta^2}\,|\mathbf{y}_a;\theta\right\}, \\ I_{a,com}(\theta) &= \mathrm{E}\left\{-\frac{\partial^2 \log f(\mathbf{z}_a)}{\partial\theta^2}\,|\mathbf{y}_a;\theta\right\}. \end{aligned}$$

The CLEM convergence rate in (4) may be slower than that of the FLEM algorithm, depending on how the current choice of term $\mathbf{y}_a$ in the CLEM is chosen. Obviously, the size of composite set a plays a key role in the trade-off between the convergence rate and computational convenience.

In order to estimate the standard error of the CLEM estimates, we need to estimate the Godambe information matrix $H(\boldsymbol{\theta})^T J(\boldsymbol{\theta})^{-1} H(\boldsymbol{\theta})$. For $H(\boldsymbol{\theta})$, under standard regularity conditions, a consistent estimator is the negative Hessian matrix evaluated at the maximum composite likelihood estimator. Given

Hosted by The Berkeley Electronic Press

$\mathbf{y}^1, \ldots, \mathbf{y}^m$ independent samples of the observed data, the estimate takes the following form:

$$\hat{H} = -\sum_{m=1}^{M} \frac{\partial^2 \log L_c^o(\theta; \mathbf{y}^m)}{\partial\theta\partial\theta^T}|_{\theta^*}.$$

If the Hessian is difficult to compute,

$$\hat{H} = \sum_{m=1}^{M} \sum_{a\in A} w_a \big(\frac{\partial \log L_a^o(\theta, \mathbf{y}_a^m)}{\partial\theta}|_{\theta^*}\big)\big(\frac{\partial \log L_a^o(\theta, \mathbf{y}_a^m)}{\partial\theta}|_{\theta^*}\big)^T,$$

as the second Bartlett identity remains true for each subset.

The estimation of $J(\boldsymbol{\theta})$ poses more difficulties , since the corresponding naive estimator

$$\hat{J} = \big(\sum_{m=1}^{M} \sum_{a\in A} w_a \frac{\partial \log L_a^o(\theta, \mathbf{y}_a^m)}{\partial\theta}|_{\theta^*}\big)\big(\sum_{m=1}^{M} \sum_{a\in A} w_a \frac{\partial \log L_a^o(\theta, \mathbf{y}_a^m)}{\partial\theta}|_{\theta^*}\big)^T$$

vanishes when evaluated at the maximum composite likelihood estimator. Instead, J can be estimated by the sample variances of the individual contributions to the composite score function. An interesting alternative is to perform jackknife (Zhao and Joe, 2005) for the evaluation of the variance matrix. For non-independent samples, one might partition the sample Y so that the corresponding contributions to the composite score function are approximately uncorrelated. Then the empirical and jackknife estimation can be derived based on these contributions. A more detailed discussion on the estimation of J especially for time series and spatial data, may be found in Varin (2008).

Such estimation of $J(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ involves the calculation of the derivatives of the log-likelihood, which may not be computationally convenient in some situations. Another alternative approach is to perform nonparametric bootstrap. The asymptotic covariances among the CLEM estimates from different bootstrap samples can be used to estimate the standard errors of the CLEM estimates.

3 Application: Multivariate Hidden Markov Models

In this section, we focus on the application of the proposed CLEM algorithm in the estimation of transition probabilities in multivariate hidden Markov model, which has direct applications in the analysis of time-course microarray data. Recent technological advances have allowed biologists nowadays to collect gene expression data at multiple times (Rangel et al., 2004; Kobayashi et al., 2005; Spellman et al., 1998). Time course expression data are essential to understand individual cellular behaviors such as mobility, division and differentiation, and gene regulatory networks are important knowledge of biological pathways. As pointed by Somogyi and Kitano (1999), the ultimate goal researchers may dream of pursuing is to infer, from the data obtained from microarray experiments, the genetic regulatory networks that lie in at their basis.

COBRA
Collection of Biostatistics
Research Archive

Let $\mathbf{Y} = \{Y^m_{g,t}, m = 1,\ldots,M, g = 1,\ldots,N, t = 1,\ldots,T\}$ be a time-course microarray data set that collects M replicates of time-series expression trajectories from a collection of N genes over T time points. Suppose the data $\mathbf{Y}$ are generated from an HMM with the set of binary hidden variables, $\mathbf{X} = \{X^m_{g,t}, m = 1,\ldots,M, g = 1,\ldots,N, t = 1,\ldots,T\}$, under the conditional density functions f_0 and f_1 on states 0 and 1, respectively. The unobserved $X^m_{g,t}$, $g \in G$, $t = 1, 2, \ldots$, are a stationary Markov order-one process. At a fixed time point t, the cross-sectional set of hidden variables, which is a subset of $\mathbf{X}$, is denoted as $\mathbf{X}^m_{\cdot t} = (X^m_{1t}, \ldots, X^m_{Nt})$. Given a collection of N genes, the joint analysis requires to estimate a $2^N \times 2^N$ transition matrix, and the related computational burden presents a serious challenge.

The pairwise CL method concerns only submatrices of the Λ, including 4×4 transition matrices $\Lambda^{gg'}$ of all gene pairs (g, g') and 2×2 transition matrices Λ^g of one gene g. Precisely, for a pair of genes (g, g'), the joint transition matrix $\Lambda^{gg'}$ constitutes the transition probabilities of the form:

$$P[(X_{g,t+1}, X_{g',t+1}) = (s_g, s_{g'})|(X_{g,t}, X_{g',t}) = (\tilde{s}_g, \tilde{s}_{g'})],\ (s_g, s_{g'}) \text{ or } (\tilde{s}_g, \tilde{s}_{g'}) \in \mathcal{S}_2 = \{\{0,0\},\{1,0\},\{0,1\},\{1,1\}\}.$$

Likewise, the marginal transition matrix Λ^g comprises of the transition probabilities:

$$P(X_{g,t+1} = s_g | X_{g,t} = \tilde{s}_g),\ s_g \text{ or } \tilde{s}_g \in \mathcal{S}_1 = \{0,1\};$$

As a result, the dimensionality of the parameter space is reduced by the CL method now only to be of order N^2, which is considerably smaller than that of the full parameter space, 2^{2N}, and hence computations in the estimation and inference become feasible.

To implement the CLEM algorithm, we need to identify distinct parameters and their constraints among the model parameters. In the HMM, the network parameters are involved in the following limiting distributions: (i) The joint limiting distribution of bivariate vectors of hidden variables for pairs of genes (g, g') at two time points $(t, t+1)$,

$$p^{gg'}_{jj'} = \lim_{t\to\infty} P[(X_{g,t}, X_{g',t}) = (s_{g,j}, s_{g',j}), (X_{g,t+1}, X_{g',t+1}) = (s_{g,j'}, s_{g',j'})],$$

where $(s_{g,j}, s_{g',j})$ and $(s_{g,j'}, s_{g',j'})$ are, respectively, the j-th and j'-th elements in $\mathcal{S}_2$; (ii) the cross-sectional pairwise limiting distribution for pairs of genes (g, g'),

$$\pi^{gg'}_j = \sum_{j'=1}^{4} p^{gg'}_{jj'} = \lim_{t\to\infty} P[(X_{g,t}, X_{g',t}) = (s_{g,j}, s_{g',j})],\ (s_{g,j}, s_{g',j}) \in \mathcal{S}_2;$$

(iii) the cross-time pairwise limiting distribution for one gene g,

$$q^g_{jj'} = \sum_{s_{g',j}=0}^{1} \sum_{s_{g',j'}=0}^{1} p^{gg'}_{jj'} = \lim_{t\to\infty} P(X_{g,t} = s_{g,j}, X_{g,t+1} = s_{g,j'}),\ (s_{g,j}, s_{g,j'}) \in \mathcal{S}_2.$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Under these limiting distributions, the transition probabilities of interest are given by

$$\Lambda_{jj'}^{gg'} = P[(X_{g,t+1}, X_{g',t+1}) = (s_{g,j'}, s_{g',j'})|(X_{g,t}, X_{g',t}) = (s_{g,j}, s_{g',j})] = p_{jj'}^{gg'} / \pi_j^{gg'}.$$

Under this re-parametrization, it is sufficient to estimate all the distinct parameters of marginal probabilities $q_{jj'}^{g}$ and pairwise probabilities $p_{jj'}^{gg'}$.

Therefore, for an HMM the expected composite likelihood can be expressed through the parameter vector $\boldsymbol{\theta}$ that includes all the distinct marginal and pairwise probabilities. Given the current update $\boldsymbol{\theta}^{(r)}$, the CL-E step computes the expected composite likelihood of the form

$$Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{\text{all } (g,g')} \mathrm{E}\left\{\log f\left(\mathbf{Y}_{g\cdot}, \mathbf{Y}_{g'\cdot}, \mathbf{X}_{g\cdot}, \mathbf{X}_{g'\cdot}; \boldsymbol{\theta}\right) |\boldsymbol{\theta}^{(r)}, \mathbf{Y}_{g\cdot}, \mathbf{Y}_{g'\cdot}\right\}.$$

Since all the expectations are restricted within a pair of Markov chains, the calculation is easily carried out using the well-known forward and backward algorithm (Baum et al., 1970).

In the CL-M step, maximizing $Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ is subject to the set of constraints that the marginal transition probabilities should be compatible with all the bivariate probabilities. The maximization under constraints is dealt with using the method of Lagrange multipliers. Iterating between the CL-E step and the CL-M step to convergence gives the maximum CL estimates of all the marginal and pairwise probabilities. The CL-E procedure is benefited from the idea of conducting local expectation which considerably simplifies the computational complexity. The essence of the CL-M step allows the sharing of information across different subsets while conducting the global maximization. Finally, we obtain the standard errors of the CL estimates by the nonparametric bootstrap method. The alternative way is to estimate the asymptotic covariance matrix, which involves the calculation of the derivatives of log-likelihood under constraints for hidden Markov models and is computationally more difficult than the nonparametric bootstrap method.

3.1 Simulation Experiments

Simulation studies were conducted to evaluate the performance of the CLEM algorithm to estimate the transition probabilities. In the first simulation, we considered a three-gene network with all pairwise dependencies. The three genes are denoted as a, b and c, and the corresponding bivariate transition matrices are Λ^{ab}, Λ^{bc} and Λ^{ac}, respectively. The true joint transition matrix was set by first generating the null matrix under independence and then deviating it by $+0.5$ in the odd-number columns and -0.5 in the even-number columns. In the marginal transition matrix for a single gene, Λ^a, Λ^b or Λ^c, was specified by randomly generating the cell probabilities randomly from a uniform distribution. In addition,

http://biostats.bepress.com/cobra/art61

the conditional densities were set as $f_0 \sim N(0,1)$ and $f_1 \sim N(4,1)$ to generate the observed time series. One thousand simulation rounds were performed. The number of replicates was set $M = 30$, and the number of time points was set as $T = 40$. Table 1 presents the summary of the CLEM estimates of the pairwise transition probabilities. It is easy to see that the CLEM method produced consistent estimates of the transition probabilities, as all the estimates appear very close to the true parameter values with small standard deviations.

In the second simulation, we consider a more complicated situation. We constructed a tree structure containing 21 nodes. The first hub node is at the top of the tree structure. We simulated its hidden states according to its marginal transition distribution. Conditional on the first node's hidden state, we independently simulate four offspring nodes according to a bivariate transition matrix Λ^{12}. Further conditional on each of the four offspring's hidden states, we independently simulate four offsprings for each of them according to another bivariate transition matrix Λ^{23}. Overall it is a tree structure of three layers, with one node on the top, and four nodes in the second layer and 16 nodes at the bottom. All the edges between the first and second layer share the same transition matrix Λ^{12} and all the edges between the second and third layer share the other transition matrix Λ^{23}. In total, we have 21 nodes and 20 edges in the tree structure. Based on each node's hidden states to be 0 or 1, we simulate the observed state according to a normal distribution $N(0,1)$ or $N(4,1)$. Overall we have a 21-variate hidden Markov model. Such kind of structure may be found in the analysis of genetic regulation pathways, whereas the top node serve as the gene regulates the four genes down the path through a same mechanism resulting to a same bivariate transition matrix. Further down the pathway, each of the second layer gene can regulates its own targets through similar mechanisms leading to another bivariate transition matrix. In order to understand the two mechanisms, we need to estimate the two transition matrices. The full likelihood method renders to infeasible calculation as the complicated dependency relationship among the 21 hidden Markov chains. We applied the composite EM method where the composite sets are all the pairs of genes linked by direct edges in the tree structure. The number of replicates was set $M = 40$, and the number of time points was set as $T = 10$. We generated 100 data sets according to the same parameters. In Table 2, the means of the estimates of all the transition probabilities are given. The true values are provided aside for comparison purposes. The standard deviation of the estimates across the 100 data sets are provides in the parenthesis. Based on one of the simulated data set, we also performed 50 times nonparametric bootstrap, so that we obtained the estimated standard deviation of the CLEM estimates. It can be noted that the CLEM method produced consistent estimates of the transition probabilities. The nonparametric bootstrap procedure yields the standard error estimates of the CLEM estimates, which are very close to the empirical standard deviation across the 100 simulations. From Table 2, we can see that the estimators

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

for Λ^{23} are more accurate than that of Λ^{12} with less standard error. This is because the estimation of the Λ^{12} relies on the likelihood compounded from four edges, whereas the the estimation of the Λ^{23} relies on the likelihood compounded from 16 edges.

4 DATA ANALYSIS

We re-analyzed the T-cell data (Rangel et al., 2004) to study the genetic dependency network in the activation process of T-cells. To generate an immune response, the T-cells become activated and then proliferate and produce cytokines involved in the regulation of B cells and macrophages, which are the most important mediators for the immune response. It is known that T-cell activation is initiated by the interaction between the T-cell receptor complex and the antigens. This stimulates a network of signaling molecules, including kinases, phosphatases and adaptor proteins that parallel the stimulatory signals received by the nucleus to control the gene transcription events. In the lab experiment, the calcium ionophore ionomycin and the PKC activator phorbol ester PMA were used to activate signaling transduction pathways leading to T-cell activation. Microarray measurements of 58 genes relevant to the immune response were taken at 10 consecutive time points. In our analysis, to satisfy the assumption of homogeneous Markov process, we used only the first five equally spaced time points after the treatment: 0, 2, 4, 6, 8 hour. At each time point, there were 44 replicated measurements for each gene. This data set is a one-sample scenario with only one experimental condition. We used a mixture of two Gaussian distributions corresponding respectively to the down-regulated and up-regulated states to model the emission distribution of the expression level for each gene. Three genes showed little variation across the time points and were considered as not involving with the response process, and thus they were excluded from the analysis. We employed the CLEM method detailed in Section 3 to simultaneously estimate the marginal transition matrices, λ^g, for all the 55 genes, and the bivariate transition matrices, $\Lambda^{gg'}$, from all the 1485 pairs of genes. To assess the significance of the dependency for each pair of genes, we formed a Pearson's chi-square test statistic for independence based on the estimated expected numbers of transitions between all the bivariate states. We then generated bootstrap samples of the whole 55-gene network by first simulating the hidden paths according to the marginal transition matrices under the null hypothesis of independency, and then simulating the expression values using the estimated Gaussian mixture distributions. In total we sampled 100 bootstrap data sets which gave 148,500 null statistics. Pooling all the null statistics together enabled us to form the empirical null distribution of the chi-square statistic. By comparing the observed statistics with the empirical null distribution, among the 1485 pairs, there existed 17 edges having p-values less than the chosen significant level of 10^{-4}.

COBRA
Collection of Biostatistics
Research Archive

Figure 1 demonstrates a core dependency network of 16 genes found by the CLEM method. Among the 17 edges, nine edges can be verified by the existing literature, which are marked by the pathway names. The edges that appear in certain known pathways, such as FAS pathway, Androgen-receptor NetPath 2, T cell receptor Netpath 11, IL-5 Netpath 17, are labelled by the pathway names. For more information regarding the labelled edges, readers are referred to `http://www.wikipathways.org` and `http://www.netpath.org`. For the other red edges, the supporting literature includes Gudi et al. (2006), Salon et al. (2006), Zheng et al. (2003), and Shin et al. (2006). By examining the network architecture, it is easy to see that CASP8 and JUND emerge as two major hubs that play important roles in the early period (0-8 hr) of the T cell activation.

Figure 1 is inserted here.

For comparison, we employed the dynamical correlation method proposed by Opgen-Rhein and Strimmer (2006) to analyze the same data set. This competing method treats the observed gene expression time series as realizations of random curves. Under the assumption of network sparsity, they proposed a shrinkage estimator of dynamical pairwise correlation matrix that takes account of the functional nature of the observed data. The dependency network was then determined according to the inverse matrix of the dynamical correlation matrix. Using static or dynamic correlation with or without shrinkage, we applied their method that produced four resulted network structures while controlling local false discovery (FDR) rate at 0.20 (Benjamini and Hochberg, 1995). See Figure 2. Each of the four identified networks found merely two edges. Only one edge is verified by the existing literature to be involved in Apoptosis pathway. The edge with biological evidence is marked with the pathway name.

Figure 2 is inserted here.

In comparison to Opgen-Rhein and Strimmer's approach with the FDR rate control level at 0.20, the CLEM method used a p-value cutoff of 10^{-4}, which corresponds to the FDR control rate less than 0.1485. Nevertheless, the CLEM method identified more biologically meaningful edges than the competing method. such high sensitivity is due to two-fold of reasons: First, the transition probabilities can reveal dependency patterns beyond linear correlation; Secondly, the CLEM-based inference does not make sparse network assumption. This is more appealing for this specific set of genes that were selected through a pre-screening procedure according to their active involvement in the T cell response process. It is naturally anticipated that these genes have high connectivity among them.

The CLEM algorithm also estimated all the pairwise bivariate transition matrices, $\Lambda^{gg'}$. For example, the pair of genes CASP8 and CDC2 in Androgen-receptor NetPath 2 pathway are connected by a significant edge with a p-value less than 6.73e-06. The corresponding estimated bivariate transition matrix

Hosted by The Berkeley Electronic Press

given below can provide interesting biological interpretations.

$$\hat{\Lambda}^{\text{CASP8,CDC2}} = \begin{array}{c|cccc} & (s_{t+1},\tilde{s}_{t+1}) & & & \\ (s_t,\tilde{s}_t) & (0,0) & (0,1) & (1,0) & (1,1) \\ \hline (0,0) & 0.5728 & 0.0691 & 0.0020 & 0.3561 \\ (0,1) & 0.2064 & 0.0139 & 0.0053 & 0.7743 \\ (1,0) & 0.0398 & 0.1529 & 0.0033 & 0.8039 \\ (1,1) & 0.2191 & 0.2056 & 0.0063 & 0.5690 \end{array}$$

where the estimated cell transition probability is

$$\hat{P}(\text{CASP8}_{t+1} = s_{t+1}, \text{CDC2}_{t+1} = \tilde{s}_{t+1} | \text{CASP8}_t = s_t, \text{CDC2}_t = \tilde{s}_t),\ (s_t,\tilde{s}_t) \in \mathcal{S}_2.$$

If both genes are down-regulated, they will have high probability to both remain down-regulated (0.5728) or both change to up-regulated (0.5690); If one of the genes is up-regulated, there is a high probability to stimulate the other one and both become up-regulated (0.7743 or 0.8039); if both of the genes are up-regulated, there is about half of the chance to remain the current state (0.5690), or have a quarter of the chance to down regulate CASP8 only (0.2056), and another quarter of chance to down regulate both genes (0.2191). One interesting finding is that all the probabilities in the third column of the matrix appear very close to zero. This implies that for this pair of genes transition to the states of CASP8's up-regulation and CDC2's down-regulation seldomly happens in the early stage of the T-cell activation. In comparison, the Pearson's product-moment correlation for this pair of genes was estimated as -0.3082, with p-value equal to $3.16e-06$. Such a one-number summary contains much less information to unveil the underlying mechanism of molecular activities than the estimated transition matrix.

5 CONCLUDING REMARKS

In this paper we presented an extension of the full likelihood EM algorithm to the setting of the composition likelihood. We established theoretical properties of the proposed CLEM algorithm. The proposed CLEM is advantageous to deal with high-dimensional data with complex dependence structures. The dimension reduction for the high-dimensional likelihood function invoked by the composite likelihood allows us to gain both computational feasibility and computational efficiency.

We have applied the proposed CL in the multivariate hidden Markov model to dissect regulatory gene network. The multivariate hidden Markov model is capable of detecting nonlinear dependencies and is applicable to describe network structures. Inference on the multivariate HMM is accomplished through the CL, which addresses the problem of high dimensionality and effectively reduces the computational

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

complexity from $O(2^{2N})$ to $O(N^2)$, with N genes involved in the network construction. The proposed methodology is not restricted to the example demonstrated in this paper. It can be viewed as a general approach to analyze high-dimensional incomplete data with complicated dependence structures.

ACKNOWLEDGEMENTS

This research is supported by a grant from Natural Science and Engineering Research Council of Canada to the first author and a grant from National Science Foundation to the second author.

APPENDIX

This appendix is devoted to the detailed proofs of the results stated in Section 2.1.

Proof of Lemma 1: The result holds by a direct application of the Jensen's inequality.

Proof of Theorem 1: By definition, $l_c(\boldsymbol{\theta}^{(r)};\mathbf{y}) = Q_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)}) - H_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)})$. Since $\boldsymbol{\theta}^{(r)}$ maximizes $Q_c(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})$, it implies that $Q_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)}) \geq Q_c(\boldsymbol{\theta}^{(r-1)}|\boldsymbol{\theta}^{(r-1)})$. Combining with the fact in Lemma 1 that $H_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)}) \leq H_c(\boldsymbol{\theta}^{(r-1)}|\boldsymbol{\theta}^{(r-1)})$, we obtain that $l_c(\boldsymbol{\theta}^{(r)}|\mathbf{y}) \geq l_c(\boldsymbol{\theta}^{(r-1)}|\mathbf{y})$.

Proof of Lemma 2: For part (a), note that

$$\begin{aligned}\nabla^{(10)} H_c(\boldsymbol{\theta}|\boldsymbol{\theta}) &= \sum_{a\in A} w_a \mathrm{E}\left\{\frac{\partial \log f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|\mathbf{y}_a;\boldsymbol{\theta}\right\} \\ &= 0.\end{aligned}$$

For part (b), we have

$$\begin{aligned}\nabla^{(11)} H_c(\boldsymbol{\theta}|\boldsymbol{\theta}) &= \sum_{a\in A} w_a \mathrm{E}\left\{\left(\frac{\partial \log f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2|\mathbf{y}_a;\boldsymbol{\theta}\right\} \\ &= \sum_{a\in A} w_a \mathrm{var}\left\{\frac{\partial \log f(\mathbf{z}_a|\mathbf{y}_a;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|\mathbf{y}_a;\boldsymbol{\theta}\right\}.\end{aligned}$$

Proof of Theorem 2: The proof of this theorem is given by a slight modification to that of Theorem 2 in Wu (1983). From the given assumptions, $l_c(\boldsymbol{\theta}^{(r-1)})$ is bounded from above. Define the solution set $\Omega = \{\text{the set of stationary points in the interior of } \Theta\}$. In light of the smoothness assumption of Q function, the point-to-set map ω determined by $\boldsymbol{\theta}^{(r)} = \omega(\boldsymbol{\theta}^{(r-1)})$ is closed under the complement of Ω. Furthermore, for any $\boldsymbol{\theta}^{(r-1)} \notin \Omega$, we have $\nabla^{(10)} H_c(\boldsymbol{\theta}^{(r-1)}|\boldsymbol{\theta}^{(r-1)}) = 0$, and $\nabla^{(10)} Q_c(\boldsymbol{\theta}^{(r-1)}|\boldsymbol{\theta}^{(r-1)}) = \nabla^{(10)} l_c(\boldsymbol{\theta}^{(r-1)}|\boldsymbol{\theta}^{(r-1)}) \neq 0$. Thus, $l_c(\boldsymbol{\theta}^{(r)}) \geq l_c(\boldsymbol{\theta}^{(r-1)})$. According to the Global Convergence Theorem (Wu, 1983), the convergence result of this theorem follows.

COBRA
BIOSTATISTICS REPOSITORY
Collection of Biostatistics
Research Archive

Proof of Theorem 3: The proof utilizes similar arguments to those given in the proof of Theorem 4 by Dempster et al. (1977). By Lemma 2, $\lim_{r\to\infty} \partial l_c(\boldsymbol{\theta}^{(r)})/\partial\boldsymbol{\theta} = \lim_{r\to\infty} \nabla^{(10)} Q_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)}) - \nabla^{(10)} H_c(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r-1)}) =$

Hosted by The Berkeley Electronic Press

0. Thus, $\boldsymbol{\theta}^*$ is a stationary point. Expanding $\nabla^{(10)}Q_c(\boldsymbol{\theta}_2|\theta_1)$ about $\boldsymbol{\theta}^*$, we obtain

$$\nabla^{(10)}Q_c(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) = \nabla^{(10)}Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*) + \nabla^{(20)}Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)(\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*) + \nabla^{(11)}Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*) + \cdots .$$

As $\boldsymbol{\theta}^{(r)} = \omega(\boldsymbol{\theta}^{(r-1)})$, and $\boldsymbol{\theta}^* = \omega(\boldsymbol{\theta}^*)$, we obtain

$$0 = \{\partial M(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}^*\}\nabla^{(20)}Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*) + \nabla^{(11)}Q_c(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*).$$

Since $Q_c(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) = l_c(\boldsymbol{\theta}_2) + H_c(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$, we have $\nabla^{(11)}Q_c(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) = \nabla^{(11)}H_c(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$. The result in Theorem 3 follows.

REFERENCES

Azzalini, A (1983), "Maximum likelihood of order m for stationary stochastic processes," *Biometrika* **70**, 381-397.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Annals of Mathematical Statistics* **41**, 164-171.

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

Besag, J.E. (1977), "Efficiency of pseudolikelihood estimation for simple Gaussian fields", *Biometrika* **64**, 616-618.

Besag, J.E. (1977), "Efficiency of pseudolikelihood estimation for simple Gaussian fields", *Biometrika* **64**, 616-618.

Cox, D. R. and Reid, N. (2004), "A note on pseudolikelihood constructed from marginal densities," *Biometrika* **91**, 729-737.

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

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G. and Bijnens, L. (2008), "A high-dimensional joint model for longitudinal outcomes of different nature," *Statistics in Medicine*, **27**, 4408-4427.

Fearnhead, P. and Donnely, P. (2002), "Approximate likelihood methods for estimating local recombination rates (with di scussion)", *Journal of the Royal Statistical Society Series B* **64**, 657-680.

Fieuws, S. and Verbeke, G. (2002), "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles", *Biometrics,* **62**, 424-431.

Godambe, V. P. (1960), "An optimum property of regular maximum likelihood equation", *Annals of Mathematical Statistics* **31**, 1208-1211.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Gudi, R., Barkinge, J., Hawkins, S., Chu, F., Manicassamy, S., Sun, Z., Duke-Cohan, J. S., Prasad, K. V. (2006), "CASP8 with SIVA, Siva-1 negatively regulates NF-kappaB activity: effect on T-cell receptor-mediated activation-induced cell death (AICD)", *Oncogene* **(25)24**, 3458-62.

Heagerty, P.J. and Lele, S.R. (1998), "A composite likelihood approach to binary spatial data", *Journal of the American Statistical Association* **93**, 1099-1111.

Hjort, N.L. and Omre, H. (1994), "Topics in spatial statistics (with discussion)", *Scandinavian Journal of Statistics* **21**, 289-357.

Kobayashi, S., Voyich, J. M., Whitney, A. R., and Deleo, F. R. (2005), "Spontaneous neutrophil apoptosis and regulation of cell survival by granulocyte macrophage-colony stimulating factor," *Journal of Leukocyte Biology*, **78**, 1408-1418.

Liang, G. and Yu, B. (2003), "Maximum pseudo likelihood estimation in network tomography", *IEEE Transactions on Signal Processing* **51**, 2043-2053.

Lindsay, B. (1988), "Composite likelihood methods," *S*tatistical inference from stochastic processes, Ed. Prabhu, N. U., 221-239. Providence, RI: American Mathematical Society.

McLachlan, G.J. and Krishnan, T. (2008), "*The EM Algorithm and Extensions*", New York: Wiley.

Opgen-Rhein, R. and Strimmer, K. (2006), "Inferring Gene Dependency Networks From Genomic Longitudinal Data: A Functional Data Approach," *Statistical Journal* **4**, 53-65.

Parner, E.T. (2001), "A composite likelihood approach to multivariate survival data", *Scandinavian Journal of Statistics* **28**, 295-302.

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L. and Falciani, F. (2004), "Modeling T-cell activation using gene expression profiling and state space modeling," *Bioinformatics* **20**, 13611372.

Renard, D., Molenberghs, G. and Geys, H. (2004), "A pairwise likelihood approach to estimation in multilevel probit models", *Computational Statistics and Data Analysis* **44**, 649-667.

Salon, C., Eymin, B., Micheau, O., Chaperot, L., Plumas, J., Brambilla, C., Brambilla, E., Gazzeri, S. (2006), "E2F1 induces apoptosis and sensitizes human lung adenocarcinoma cells to death-receptor-mediated apoptosis through specific downregulation of c-FLIP(short)", *Cell Death Differ* **13(2)**, 260-72.

Shin, K. H., Kang, M. K., Kim, R. H., Christensen, R., Park, N. H. (2006), "Heterogeneous nuclear ribonucleoprotein G shows tumor suppressive effect against oral squamous cell carcinoma cells", *Clinical Cancer Research* **12(10)**, 3222-3228.

Somogyi, R. and Kitano, H. (1999), "Gene expression and genetic networks – Session introduction", *Pacific Symposium on Biocomputing*, 3-4.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Song, P.X.-K. (2007), "*Correlated Data Analysis: Modeling, Analytics, and Applications*", New York: Springer.

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

Stein, M. L. (2004), "Approximating likelihoods for large spatial data sets", *Journal of the Royal Statistical Society B*, **66**, 275-296.

Varin, C. (2008) "On composite marginal likelihoods", *AStA Advances in Statistical Analysis*, **92**, 1-28.

Varin, C., Host, G., and Skare, O. (2005), "Pairwise likelihood inference in spatial generalized linear mixed models," *Computational Statistics and Data Analysis* **49** 1173-1191.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., Bucan, M. (2007), "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data", *Genome Research* **17**, 1665-1674.

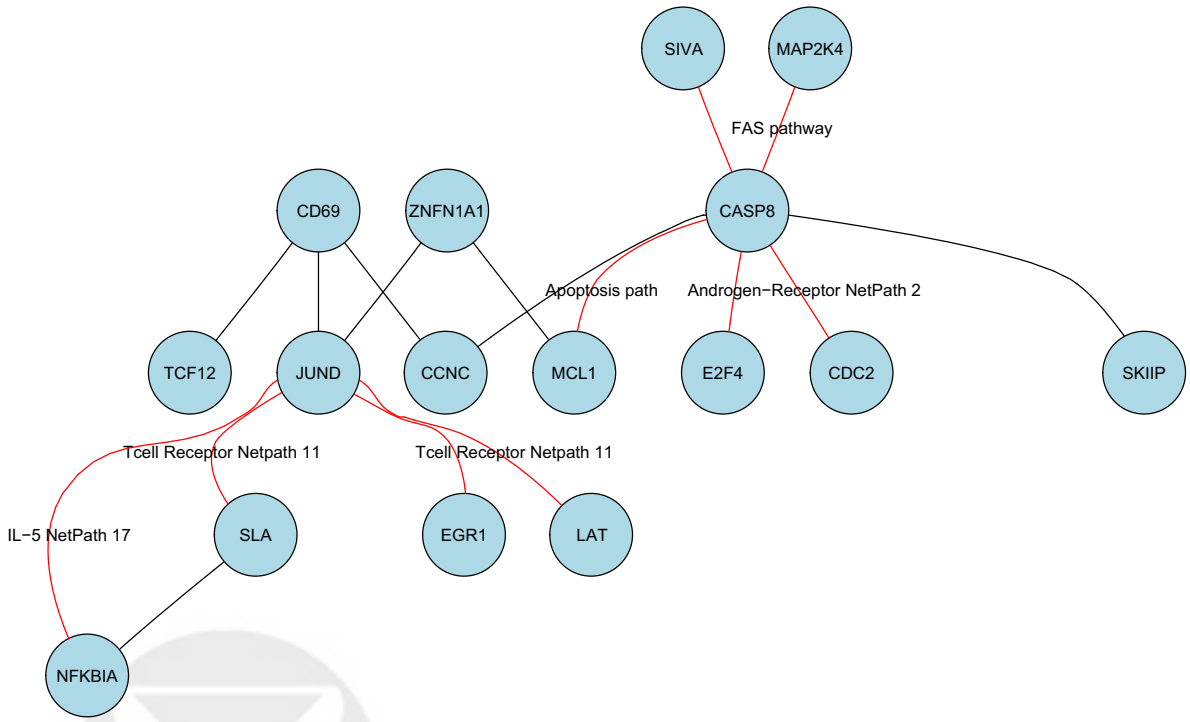
Wu, C.F. J. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, **11**, 95-103.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, **33**, 335-356.

Zheng, B., Fiumara, P., Li, Y. V., Georgakis, G., Snell, V., Younes, M., Vauthey, J. N. , Carbone, A., Younes, A. (2003), "MEK/ERK pathway is aberrantly active in Hodgkin disease: a signaling pathway shared by CD30, CD40, and RANK that regulates cell proliferation and survival", *Blood* **102(3)**, 1019-1027.

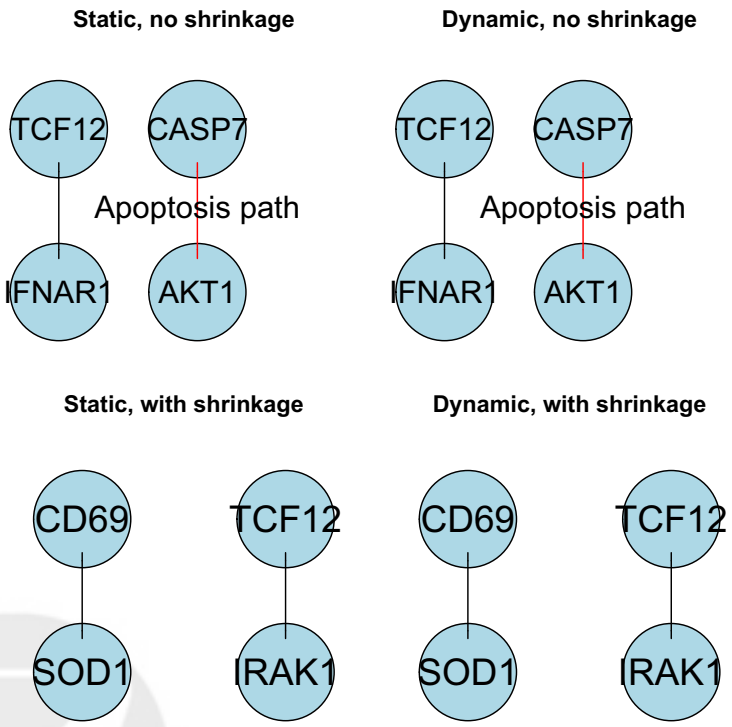
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 1: A core network of 16 genes in the T cell response identified by the CLEM method.



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 2: Networks of gene pairs in the T cell response identified by the dynamic correlation method.



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 1: 3-variate HMM: Average CLEM estimates and empirical standard deviations in the parentheses for the bivariate transition probabilities over 1,000 simulation runs. The true values of probabilities are also listed aside for reference.

Matrix	Estimate	True	Estimate	True	Estimate	True	Estimate	True
Λ^{ab}	0.1560	0.1558	0.1132	0.1131	0.3804	0.3797	0.3504	0.3514
	(0.0225)		(0.0201)		(0.0288)		(0.0288)	
	0.1519	0.1521	0.1485	0.1502	0.3285	0.3273	0.3711	0.3704
	(0.0229)		(0.0233)		(0.0289)		(0.0304)	
	0.3949	0.3953	0.3537	0.3528	0.1331	0.1344	0.1183	0.1174
	(0.0304)		(0.0306)		(0.0222)		(0.0207)	
	0.3270	0.3273	0.3628	0.3629	0.1682	0.1675	0.1421	0.1423
	(0.0301)		(0.0301)		(0.0234)		(0.0221)	
Λ^{bc}	0.2946	0.2962	0.2283	0.2282	0.3054	0.3036	0.1716	0.1719
	(0.0285)		(0.0267)		(0.0296)		(0.0248)	
	0.2082	0.2076	0.3322	0.3327	0.1977	0.1970	0.2620	0.2627
	(0.0256)		(0.0295)		(0.0245)		(0.0266)	
	0.2743	0.2734	0.2285	0.2285	0.3161	0.3166	0.1810	0.1815
	(0.0281)		(0.0266)		(0.0291)		(0.0250)	
	0.2068	0.2063	0.2656	0.2652	0.1896	0.1902	0.3380	0.3384
	(0.0265)		(0.0279)		(0.0250)		(0.0305)	
Λ^{ac}	0.1582	0.1586	0.0934	0.0935	0.4864	0.4846	0.2620	0.2633
	(0.0232)		(0.0198)		(0.0329)		(0.0290)	
	0.1199	0.1204	0.1949	0.1954	0.2836	0.2833	0.4017	0.4010
	(0.0205)		(0.0251)		(0.0275)		(0.0299)	
	0.4253	0.4256	0.3226	0.3217	0.1236	0.1238	0.1285	0.1174
	(0.0306)		(0.0288)		(0.0207)		(0.0208)	
	0.2586	0.2579	0.4310	0.4323	0.1401	0.1397	0.1703	0.1702
	(0.0282)		(0.0305)		(0.0217)		(0.0246)	

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 2: 21-variate HMM: Average CLEM estimates for the bivariate transition probabilities over 100 simulation runs. The true values of probabilities are also listed aside for reference. The empirical standard deviations from 100 data sets are in parenthesis and the estimated standard deviation from bootstrap on one data set are in brackets.

Matrix	Estimate	True	Estimate	True	Estimate	True	Estimate	True
Λ^{12}	0.2799	0.2800	0.1233	0.1200	0.1175	0.1200	0.4793	0.4800
	(0.0288)		(0.0202)		(0.0166)		(0.0354)	
	[0.0314]		[0.0180]		[0.0185]		[0.0353]	
	0.3664	0.3600	0.0420	0.0400	0.2440	0.2400	0.3476	0.3600
	(0.0492)		(0.0187)		(0.0356)		(0.0371)	
	[0.0561]		[0.0231]		[0.0291]		[0.0430]	
	0.3907	0.3200	0.2116	0.1800	0.0664	0.0800	0.3313	0.4200
	(0.0453)		(0.0351)		(0.0186)		(0.0519)	
	[0.0396]		[0.0348]		[0.0190]		[0.0468]	
	0.5017	0.4200	0.0971	0.0800	0.1445	0.1800	0.2567	0.3200
	(0.0361)		(0.0138)		(0.0190)		(0.0336)	
	[0.0393]		[0.0133]		[0.0186]		[0.0390]	
Λ^{23}	0.2185	0.2100	0.1955	0.1900	0.1851	0.1900	0.4008	0.4100
	(0.0150)		(0.0120)		(0.0110)		(0.0203)	
	[0.0138]		[0.0154]		[0.0133]		[0.0196]	
	0.3008	0.2900	0.1140	0.1100	0.3042	0.3100	0.2810	0.2900
	(0.0195)		(0.0098)		(0.0171)		(0.0167)	
	[0.0178]		[0.0113]		[0.0122]		[0.0145]	
	0.3106	0.2900	0.3260	0.3100	0.0988	0.1100	0.2646	0.2900
	(0.0197)		(0.0175)		(0.0109)		(0.0213)	
	[0.0221]		[0.0235]		[0.0147]		[0.0247]	
	0.4332	0.4100	0.2020	0.1900	0.1722	0.1900	0.1925	0.2100
	(0.0172)		(0.0127)		(0.0131)		(0.0144)	
	[0.0229]		[0.0111]		[0.0164]		[0.0128]	

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive