Estimation of the diffusion coefficient in a mixture model with diffusing and fixed particles

Mats Kvarnström
Chalmers University of Technology and Göteborg University

October 22, 2002

Abstract

Particle positions have been observed and estimated in a series of images. The particles are assumed to perform a Brownian motion, however some of them seem to be fixed. A model is introduced with two kinds of particles, diffusing and fixed. To each particle position estimate we assume an additive normal measurement error. The parameter of the model consists of the diffusion variance, the measurement error variance, and the proportion of diffusing particles. The problem can be considered as an incomplete data problem since we do not know *a priori* which particles are really diffusing. The complete data is of curved exponential type and the observed data is a mixture of two normal components. The maximum likelihood estimator is computed via the EM algorithm. The estimator is shown to be strongly consistent and asymptotically normal, as the number of particles approaches infinity, under a reasonable restriction on the parameter space.

Key words: asymptotic normality, curved exponential family, discretely observed diffusion, EM algorithm, measurement error, mixture distribution, strong consistency

1 Introduction

This article deals with the estimation of the diffusion variance (or equivalently, the diffusion coefficient) of colloidal particles. Particle positions have been observed and estimated in a series of images (frames) recorded on a video microscope using more or less standard image processing algorithms and tools. The position estimates of the particles are then linked so that we get a trajectory for each particle in the sequence.

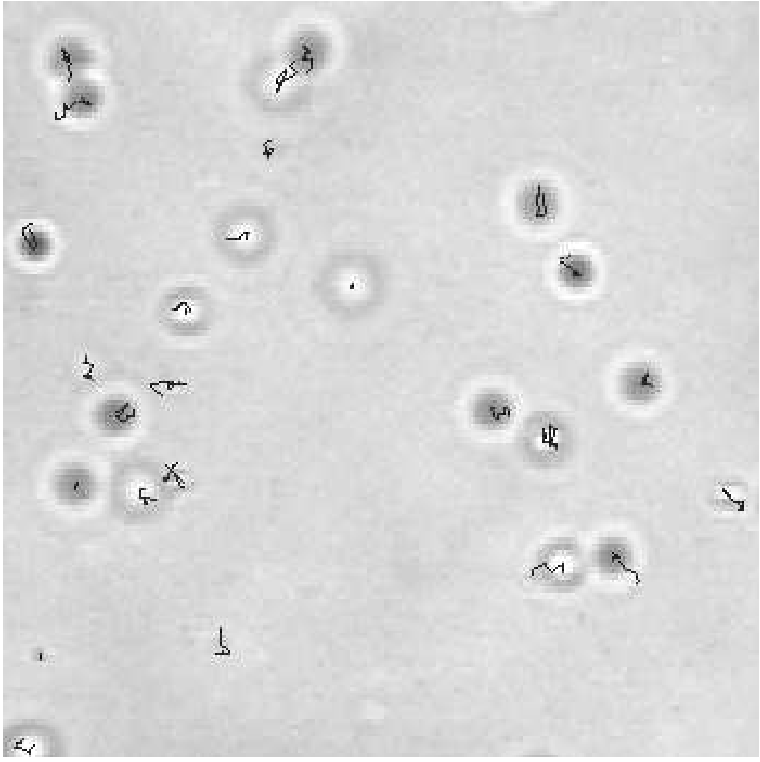

Figure 1: The 26 trajectories estimated in a sequence of 12 images together with the first image in the sequence. Notice that three of the particles seem to be fixed.

The particles are performing a Brownian motion in three dimensions and they move independently of each other. The naive estimate of the diffusion variance is the mean of the squared increments. If we, however, assume that the position estimates of the particles are imperfect, i.e. if we assume a measurement error, this estimate of the diffusion variance will be biased.

Another complicating fact is that some of the observed particles are not moving but are instead either particles adsorbed on the objective or cover glass of the specimen, or "false" particles which are due to for example defects in the optics of the microscope. One solution to this problem is of course to remove these false particles manually. This is not satisfactory from a statistical point of view, first because we should be able to do it using statistical methods, second because these particles actually gives us information on the measurement error.

The proposed method takes care of both of these problems by assuming a mixture distribution of diffusing and fixed particles and then estimate the relevant parameters using maximum likelihood estimation. For fixed number of observed increments the estimator is shown to be strongly consistent and asymptotically normally distributed, as we let the number of particles go to infinity.

An example of what the situation may look like, can be seen in Figure 1. The figure shows the initial image in a sequence of 12 images, together with the position estimates of a major part of the particles in the subsequent 11 images, thereby forming the estimated trajectories of the particles. By manual inspection, we made sure that no change of the identities of the particles occurred in the process of converting the position estimates in the images into trajectories. The time interval between two images is 40 milliseconds. The particles are spherical, made of polystyrene and are all equal in size, 494 nm in diameter. The apparent difference in size and brightness are due to an out-of-focus effect and depend on their placement in depth relative to the focal plane. Particles above the focal plane are bright in the middle and dark on the circumference and vice versa for the particles below the focal plane. The depicted size of a particle is also increased the further away from the focus plane it is. Here three of particles seem to be fixed; one adsorbed on the cover glass, one on the objective glass, and one which probably correspond to a defect in the optics.

1.1 Outline of the paper

The article is organized as follows.

In Section 2 we introduce a model with two kinds of particles, diffusing and fixed, both observed with additive measurement error on the position estimates. The observation length is $N+1$ frames. We have three parameters in the model, σ^2 is the diffusion variance, σ_e^2 is the variance of the measurement error on the position estimates, and p is the proportion of diffusing particles. The problem can now be considered as a missing data problem since the only way to infer whether a particle is diffusing or not is by the observed movement of the particle. In this section we also look at the structure of the covariance matrices for the two kinds of particles.

In Section 3 we introduce two concepts of data, observed and complete data. The observed data is the observed increments of each particle and the complete data is the observed data together with the classification variable of each particle (indicating whether it is diffusing or fixed). We also look at the different densities these two kinds of data correspond to. In particular, the observed data is of finite mixture type.

The likelihood is discussed in Section 4 together with the EM algorithm (see Dempster et al. (1977) and McLachlan and Krishnan (1997)) in Section 4.1 and some basic theory regarding this method of finding the maximum likelihood estimate.

In Section 5 we study the asymptotic properties of the estimator when we keep the observation length and and let the number of particles go to infinity. We show that the estimator of the triple $(\sigma^2, \sigma_e^2, p)$ using only the observed increments is strongly consistent and asymptotically normally distributed under a small but reasonable restriction on the parameter space.

Finally, in Section 6, we use the model assumption and estimate the diffusion variance for the data corresponding to the trajectories seen in Figure 1.

2 The model

Denote the true and the observed position of a particle at time $k = 0, \dots, N$ by R_k and S_k, respectively, where S_k is the true position with measurement noise added to it. We arrive at the following state-space model

$$\begin{array}{rcccc} R_k & = & R_{k-1} & + & w_k \\ S_k & = & R_k & + & e_k \end{array} \qquad (1)$$

where w is the position increment of the motion of the particle and e the measurement error of the position.

The particle is performing a Brownian motion so the w_k:s are i.i.d. zero mean normally distributed random variables with variance σ^2. For a fixed particle, w_k is zero for all k (an alternative is to think about this as σ^2 being equal to zero for fixed particles). The errors e_k are also assumed to be i.i.d. zero mean normally distributed random variables with variance σ_e^2, independent of the true position of the particle, of other particles, and of the increments w_k.

The initial value, R_0 is assumed to be a constant.

Let n be the number of observed particles and introduce an indicator Z_i to each particle to be one if the i:th particle is diffusing (performing a Brownian motion, $\sigma^2 > 0$) and zero if fixed ($\sigma^2 = 0$). Let $\{Z_i\}_{i=1}^n$ be i.i.d. and introduce a third parameter p, defined as

$$p = \mathbf{P}(Z_i = 1).$$

The model can easily be extended to noisy observations of a Brownian motion in d dimensions if we assume the measurement error in each dimension to be distributed as w_k above and independent of each other. Then a particle follows same the state-space model (1) in each dimension independently of each other.

For ease of notation, we will assume that d is one. The only exception from this is in Section 6, which deals with the analysis on the trajectories plotted in Figure 1.

2.1 Covariance matrix of the observed increment vector

We define the observed increments for a particle as $Y_k = S_k - S_{k-1}$, $k = 1, \ldots, N$. The covariance matrix of the increment vector, $Y = [Y_1, \ldots, Y_N]^T$, is

$$\Sigma_1 = \sigma^2 I + \sigma_e^2 T \tag{2}$$

for a diffusing particle and

$$\Sigma_0 = \sigma_e^2 T$$

for a fixed particle, where I is the N times N identity matrix and T is the tridiagonal matrix defined as

$$T = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix}.$$

We see from the covariance matrix above that the measurement noise in the position induces a dependence between the increments, which originally, by definition of a Brownian motion, were independent.

2.2 Transformation of the increment vector

To make our formulas look cleaner in the subsequent sections, we use some basic linear algebra to transform the increment vector so that the transformed vectors become uncorrelated.

In (2), Σ_1 has the same eigenvectors as T since every vector is an eigenvector to I. If we denote the eigenvalues of T as λ_k, $k = 1, \ldots, N$, the eigenvalues of Σ_1 are

$$\gamma_k = \sigma^2 + \sigma_e^2 \lambda_k, \; k = 1, \ldots, N.$$

Let U have the eigenvectors of T as columns. Then we can write, by the spectral decomposition theorem, $T = U \Lambda U^T$, where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_N\}$. So if

$$\tilde{Y} = U^T Y \tag{3}$$

is the transformed increment vector, its covariance matrix will be diagonal:

$$\begin{aligned} \mathbf{Var}\{\tilde{Y}\} &= U^T \mathbf{Var}\{Y\} U \\ &= U^T(\sigma^2 I + \sigma_e^2 U \Lambda U^T) U \\ &= \sigma^2 I + \sigma_e^2 \Lambda \\ &= \text{diag}\{\gamma_1, \ldots, \gamma_N\} \end{aligned} \tag{4}$$

for a diffusing particle and likewise, with $\sigma^2 = 0$, for a fixed particle. The dependence between the increments is now "hidden" in U and Λ, which do not depend on σ^2 or σ_e^2, but only on the length of the increment vector N, which of course is known.

3 Data and densities

The observed data consists of the vectors of noise corrupted increments Y_i while the classification variables Z_i are unobserved. Together, they make up the complete data, denoted $X_i = (Y_i, Z_i)$, $i = 1, \dots, n$.

The probability density function of the complete data X is

$$g_c(x\,;\sigma^2,\sigma_e^2,p) = [pf_1(y\,;\sigma^2,\sigma_e^2)]^z\,[(1-p)f_0(y\,;\sigma_e^2)]^{1-z} \tag{5}$$

where f_1 and f_0 are the pdf of a zeros mean N-variate normally distributed random variable with covariance matrices Σ_1 and Σ_0, respectively. Using the transformed increment vector $\tilde{Y}$, we write

$$f_1(y\,;\sigma^2,\sigma_e^2) = \frac{1}{(2\pi)^{N/2}\prod_{k=1}^{N}(\sigma^2+\lambda_k\sigma_e^2)^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\sigma^2+\lambda_k\sigma_e^2}\right\} \tag{6}$$

and

$$f_0(y\,;\sigma_e^2) = \frac{1}{(2\pi)^{N/2}\prod_{k=1}^{N}(\lambda_k\sigma_e^2)^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\lambda_k\sigma_e^2}\right\} \tag{7}$$

$$= \frac{1}{(2\pi\sigma_e^2)^{N/2}(N+1)^{1/2}}\exp\left\{-\frac{1}{2\sigma_e^2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\lambda_k}\right\} \tag{8}$$

where $\tilde{y} = U^T y$ from (3). The second equality in the last expression, comes from the fact that $\prod \lambda_k = |\Lambda| = |T| = N+1$.

In the d dimensional case, f_i will be a dN-variate normal density with d independent parts, one in each in each dimension, since each coordinate process is independent of the others.

The probability density of the observed data, Y, we get by integrating (5) over Z

$$g(y\,;\sigma^2,\sigma_e^2,p) = pf_1(y\,;\sigma^2,\sigma_e^2) + (1-p)f_0(y\,;\sigma_e^2). \tag{9}$$

Our observed data is a finite mixture of two normal components. For a thorough account on finite mixture models and their applications, we refer to McLachlan and Peel (2000).

4 Likelihood

Denote our parameter $\theta = [\sigma^2, \sigma_e^2, p]^T$.

The complete likelihood L_c induced by the complete data (increments and classification variables) from n observed particles is

$$L_c(\theta) = \prod_{i=1}^{n} [pf_1(y_i\,; \sigma^2, \sigma_e^2)]^{z_i} [(1-p) f_0(y_i\,; \sigma_e^2)]^{1-z_i} \tag{10}$$

However, our observed data consists of only the increments so the observed likelihood becomes

$$L(\theta) = \prod_{i=1}^{n} pf_1(y_i\,; \sigma^2, \sigma_e^2) + (1-p) f_0(y_i\,; \sigma_e^2) \tag{11}$$

4.1 The EM algorithm

A intuitive method to get the maximum likelihood estimate from our observed data is to use a method whose name, the EM algorithm, comes from the article by Dempster et al. (1977), but whose essence actually was introduced and used, for the special case of finite mixtures of distributions from the exponential family, by Hasselblad (1969). Further examples of its use, before it was actually called the EM algorithm, can be found in Day (1969), Behboodian (1970) and Sundberg (1976). For an overview of the theory and applications of the method we refer to McLachlan and Krishnan (1997).

The method uses the simple structure of the complete likelihood together with estimates of the unobserved data in a iterative scheme.

4.1.1 Notation

Let k be the conditional density of the unobserved data Z, given the observed Y. Then

$$k(z|\,y; \theta) = \frac{g_c(x; \theta)}{g(y; \theta)}$$

Taking the logarithm and re-arranging, we get

$$\log g(y; \theta) = \log g_c(x; \theta) - \log k(z|\,y; \theta) \tag{12}$$

Denote by $L(\theta)$ and $L_c(\theta)$, the observed and the complete data likelihoods

$$L(\theta) = g(y; \theta)$$

$$L_c(\theta) = g_c(x;\theta)$$

and take the conditional expectation of (12) given Y, at the parameter θ'

$$\log L(\theta) = \mathbf{E}_{\theta'}\{\log g_c(x;\theta)|y\} - \mathbf{E}_{\theta'}\{\log k(z|\, y;\theta)|y\}$$

and denote the first term $Q(\theta|\theta')$ and the second $H(\theta|\theta')$.

Let furthermore

$$S(y;\theta) = \partial \log L(\theta)/\partial\theta$$

and

$$S_c(x;\theta) = \partial \log L_c(\theta)/\partial\theta$$

be the score functions.

4.1.2 Method

The EM algorithm consists of two steps at each iteration. Assume $\theta^{(k)}$ is the estimate of θ from the k:th iteration step. Then we do:

- E-step: Compute $Q(\theta|\theta^{(k)})$
- M-step: Choose $\theta^{(k+1)} \in \text{argmax}\, Q(\theta|\theta^{(k)})$

Since $H(\theta|\theta^{(k)}) \leq H(\theta^{(k)}|\theta^{(k)})$ for all θ by Jensen's inequality, the rule of choosing $\theta^{(k+1)}$ as a maximizer of $Q(\theta|\theta^{(k)})$ gives us that

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)})$$

guaranteeing that we approach a local maximum of the likelihood. In practice, we iterate until some sort of convergence criterion is met.

Notice that there is no guarantee that we converge to the *global* maximum of the likelihood function, and thus at the actual maximum likelihood estimate.

The EM algorithm should simply be thought of a numerical method for maximizing the likelihood. Often, it suffers from painstakingly slow convergence, and then a Newton-Raphson approach usually does better. However, when the data is considered to have missing values, it is very appealing to use it since we also get estimates of the missing values. We write "is considered" because the missing values may be a theoretical construction only. In our problem, though, it is natural to think of the classification variables as being missing data.

4.1.3 Finite mixtures

When the data comes from a mixture, the E-step consists of estimating the unobserved data, i.e. the classification variables. In the M-step we maximize the complete likelihood (10) using the estimated classification variables, $\hat{Z}_i$, from the E-step together with our data Y_i:

- E-step: For each $i = 1, \ldots, n$, compute

$$\hat{Z}_i = \mathbf{E}_{\theta^{(k)}}\{Z_i|Y_i\} = \frac{p^{(k)} f_1(y_i; \Sigma_1^{(k)})}{p^{(k)} f_1(y_i; \Sigma_1^{(k)}) + (1-p^{(k)}) f_0(y_i; \Sigma_0^{(k)})}$$

- M-step: Maximize $\mathbf{E}_{\theta^{(k)}}\{\log L_c(\theta)|y\} =$

$$= \sum_{i=1}^{n} \hat{Z}_i \log\{p f_1(y_i\,; \sigma^2, \sigma_e^2)\} + (1-\hat{Z}_i) \log\{(1-p) f_0(y_i\,; \sigma_e^2)\}$$

  with respect to $\theta = (\sigma^2, \sigma_e^2, p)$.

In this application of the EM algorithm, each of the two steps has a probabilistic meaning; in the E-step we classify each particle using a quadratic discriminant rule, and in the M-step we use these classifications as if we had the complete data. Note however that the classifications are not just zero or one, but any number in between.

A fast, Newton-Raphson based, computational method for the M-step can be found in (30) of Appendix C.

4.1.4 Information matrix

Now we are going to explore how the information matrix from our observed data relates to the information matrix from the complete data. This will also give us a computationally efficient way of calculating the observed information when using the EM algorithm. Appropriate regularity conditions allowing us to differentiate under the integral are assumed in the following. In our application this is true since we are dealing with exponential families, see for example van der Vaart (1999).

Let

$$I(\theta; y) = -\partial^2 \log L(\theta)/\partial\theta\partial\theta^T$$

and

$$I_c(\theta; x) = -\partial^2 \log L_c(\theta)/\partial\theta\partial\theta^T$$

Using the following version of (12)

$$\log L(\theta) = \log L_c(\theta) - \log k(z|y; \theta),$$

and differentiating twice and taking conditional expectation of z given y, we get

$$I(\theta;y) = \mathcal{I}_c(\theta;y) - \mathcal{I}_m(\theta;y) \tag{13}$$

where

$$\mathcal{I}_c(\theta;y) = \mathbf{E}_\theta\{I_c(\theta;x)|y\}$$

and

$$\mathcal{I}_m(\theta;y) = -\mathbf{E}_\theta\{\partial^2 \log k(z|y;\theta)/\partial\theta\partial\theta^T|y\}$$

corresponding to the conditional expectation of the information matrix of the complete data given y, and the missing information, respectively.

In Louis (1982), it is shown that $\mathcal{I}_m$ can be expressed as

$$\mathcal{I}_m(\theta;y) = \mathbf{E}_\theta\{S_c(X;\theta)S_c^T(X;\theta)|y)\} - S(y;\theta)S^T(y;\theta). \tag{14}$$

This is nice, first since $S(y;\theta) = 0$ at the MLE $\hat{\theta}$ and secondly because now the observed information matrix at $\hat{\theta}$ is

$$I(\hat{\theta};y) = \mathcal{I}_c(\hat{\theta};y) - [\mathbf{E}_\theta\{S_c(X;\theta)S_c^T(X;\theta)|y)\}]_{\theta=\hat{\theta}} \tag{15}$$

where both terms easily can be computed in the last M-step in the EM algorithm since the first term is actually the negative of the Hessian of the function to maximize in the M-step, and this is often used in the actual maximization.

Denote the expected information matrix by $\mathcal{I}(\theta)$ which can be expressed as

$$\mathcal{I}(\theta) = \mathcal{I}_c(\theta) - \mathbf{E}_\theta\{\mathcal{I}_m(\theta;Y)\} \tag{16}$$

by taking expectation of (13) over the distribution of Y.

5 Asymptotics

Is this section we are going to study the asymptotic properties of the estimator as the number of particles n grows large. As it turns out, our maximum likelihood estimator is both strongly consistent and asymptotically normal. First we address some important issues regarding the data and the parameter space.

The complete data comes from the exponential family of distributions, see for example Lindsey (1996). If $N \neq 1$ however, it is non-regular or curved, since the parameter space is 3-dimensional and the dimension of the sufficient statistics is $N + 2$ (see the Appendix for a derivation of this). The case $N = 1$ is not very interesting though since we think of our problem as studying a video sequence of images of particles.

Let Ω be the parameter space consisting of those θ defining valid finite mixture densities (9). $\Omega = \{\theta = [\sigma^2, \sigma_e^2, p]^T : p \in [0,1], \sigma^2 > 0, \sigma_e^2 > 0\}$. The true parameter point θ_0 is assumed to lie in the interior of Ω, denoted int(Ω).

Often when one deals with finite mixtures, there is a problem of identifiability, i.e. that a permutation of the parameters in the model yields the same distribution. In our model, and as long as the true parameter θ_0 lies in the interior of Ω, we do not have this problem since the two distributions in the mixture are not interchangable.

The asymptotics when using complete data is covered in Appendix D.

5.1 Existence of a maximum likelihood estimator

To guarantee that the likelihood has a global maximizer of for each n, we restrict the parameter space Ω, by using an idea from Hathaway (1985). For fixed $c \in (0,1)$, define Ω_c to be the subset of Ω such that

$$0 < c \leq \frac{\sigma^2}{\sigma_e^2} \leq c^{-1} < \infty \tag{17}$$

This restriction means that we do not allow the "signal-to-noise" ratio to be too small, neither too big.

Lemma 1. *Let $\{Y_1, \ldots, Y_n\}$ be a set of observations from the finite mixture specified by the density* (9)*. Then, with probability one, there exists a global constrained maximizer of $L(\theta)$ in Ω_c.*

Proof. The idea is to show that

$$\sup_{\theta \in \Omega_c} L(\theta) = \sup_{\theta \in K} L(\theta)$$

for some appropriate, compact $K \subset \Omega$.

With probability one, the increment vectors will all be different from zero. Therefore all the terms in the likelihood will stay bounded. Also, it will go to zero if both σ^2 and σ_e^2 either go to zero or to infinity. By condition (17) above however, it is enough that one of the two variances goes to zero or infinity; the other variance "will follow".

So, there exists constants a_i and b_i such that $K = \{\theta \in \Omega_c : a_1 \leq \sigma_e^2 \leq a_2,\ b_1 \leq \sigma^2 \leq b_2\}$ gives the desired result. □

Remark: Without the condition (17), our trouble spots are

- $L \to \prod_{i=1}^n f_0(y_i\,; \sigma_e^2)$ as $\sigma^2 \to 0$
- $L \to p^n \prod_{i=1}^n f_1(y_i\,|\,\sigma^2, 0)$ as $\sigma_e^2 \to 0$
- $L \to (1-p)^n \prod_{i=1}^n f_0(y_i\,; \sigma_e^2)$ as $\sigma^2 \to \infty$

A maximum hence exists, but it does not necessarily have to be unique for finite n: If $\hat{p} = 0$, we see that σ^2 is "free". Likewise, if $\hat{p} = 1$ and $N = 1$, all values of σ^2 and σ_e^2 satisfying $\sigma^2 + 2\sigma_e^2 = c$ for some constant c, are maximum likelihood estimators.

5.2 Consistency

5.2.1 Special case, $N = 1$

When $N = 1$, the complete data is of regular exponential type. Sundberg (1974) gives the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\theta}_n$, under the single condition that the information matrix $\mathcal{I}(\theta)$ is positive definite at the true parameter point θ_0. Since Lemma 2 below says that this is true for all $\theta_0 \in \text{int}(\Omega)$, we are actually done for $N = 1$, both with the consistency and the asymptotic normality.

5.2.2 Generally, $N \geq 1$

To prove consistency of the maximum likelihood estimator for general N, we verify that Wald's classical conditions for the mixture density g in (9) are satisfied when the true parameter is in Ω_c. In the process, we use results from Redner (1981).

Theorem 1. *Let the true parameter point θ_0 be in Ω_c and let $\hat{\theta}_n$ be the global maximizer of $L(\theta)$ over Ω_c, for each n. Then*

$$\mathbf{P}\{\hat{\theta}_n \to \theta_0 \ \textit{as}\ n \to \infty\} = 1$$

Proof. Wald's conditions are enumerated as in Redner (1981) to 1 through 6. We refer the reader to that article.

Conditions 1,2,4' and 5 are satisfied for Ω and the mixture component densities f_1 and f_0. The proof of Redner's Theorem 5 shows that Conditions 2 and 4 are satisfied for the mixture density $g = pf_1 + (1-p)f_0$. If we restrict Ω to Ω_c as defined above (17), then also Conditions 3 and 6 are satisfied, giving us the result by applying Theorems 1 and 2 from Wald (1949). □

Remark 1: The extra condition (17) helps us in the process of first to prove that an maximum likelihood estimator exists for all n and second, to prove that Condition 3 of Redner (1981), $L(\theta_i) \to 0$ if $d(\theta_0, \theta_i) \to \infty$, where d means Euclidean distance.

Remark 2: The restriction (17) of the parameter space also gives us consistency under an expanded model with a drift term in the diffusion together with systematic position measurement errors, that is, if the mixture components have non-zero expected value and we need to estimate these as well. Also, the conclusion of Lemma 1 holds if the number of observations n is larger than three (one more than the number of mixture components).

5.3 Asymptotic normality

Sufficient conditions for the asymptotic normality of the maximum likelihood estimate $\hat{\theta}_n$ can be found in for example Theorem 5.23 in van der Vaart (1999). Since we have consistency and that $\log g(y;\theta)$ is smooth, what remains is to be proven is that the map $\theta \mapsto \mathbf{E}_{\theta_0} \log g(Y;\theta)$ admits a second order Taylor expansion around $\theta_0 \in \text{int}(\Omega)$ with non-singular second derivative matrix. In other words, what we have to prove is that the expected information matrix $\mathcal{I}(\theta_0)$ is positive definite.

Theorem 2. *Let $\theta_0 \in int(\Omega_c)$ be the true parameter point. Then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal, i.e.*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \mathcal{I}(\theta_0)^{-1}) \tag{18}$$

as $n \to \infty$.

By the discussion above, the result follows from the next lemma.

Lemma 2. *The information matrix $\mathcal{I}(\theta)$ is positive definite for all $\theta \in int(\Omega)$.*

Proof. Positive definiteness means that $a^T \mathcal{I}(\theta) a > 0$, for all $a \in \mathbb{R}^3 \setminus 0$.

Now, since $\mathcal{I}(\theta)$ is the variance of the score function $\partial \log g(Y;\theta)/\partial\theta$, $a^T \mathcal{I}(\theta) a$ is the variance of the linear combination $a^T \partial \log g(Y;\theta)/\partial\theta$.

So, what we have to prove is that

$$\mathbf{Var}\{a^T \frac{\partial \log g(Y;\theta)}{\partial\theta}\} > 0$$

for all $a \in \mathbb{R}^3 \setminus 0$.

Assume the opposite. Then we have, with probability one, that

$$a^T \frac{\partial \log g(Y;\theta)}{\partial\theta} = 0 \tag{19}$$

for some $a \in \mathbb{R}^3 \setminus 0$ since the mean of the score is zero.

Writing out the components of the score function, we have

$$\begin{aligned}\frac{\partial \log g}{\partial \sigma^2} &= \frac{p\frac{\partial f_1}{\partial \sigma^2}}{pf_1 + (1-p)f_0} \\ \frac{\partial \log g}{\partial \sigma_e^2} &= \frac{p\frac{\partial f_1}{\partial \sigma_e^2} + (1-p)\frac{\partial f_0}{\partial \sigma_e^2}}{pf_1 + (1-p)f_0} \\ \frac{\partial \log g}{\partial p} &= \frac{f_1 - f_0}{pf_1 + (1-p)f_0}\end{aligned}$$

where

$$\begin{aligned}\frac{\partial f_1}{\partial \sigma^2} &= \frac{1}{2}\sum_{k=1}^{N}\left(\frac{\tilde{y}_k^2}{(\sigma^2 + \lambda_k\sigma_e^2)^2} - \frac{1}{\sigma^2 + \lambda_k\sigma_e^2}\right) f_1(y\,;\sigma^2,\sigma_e^2) = k_1(y)f_1(y\,;\sigma^2,\sigma_e^2) \\ \frac{\partial f_1}{\partial \sigma_e^2} &= \frac{1}{2}\sum_{k=1}^{N}\left(\frac{\lambda_k\tilde{y}_k^2}{(\sigma^2 + \lambda_k\sigma_e^2)^2} - \frac{\lambda_k}{\sigma^2 + \lambda_k\sigma_e^2}\right) f_1(y\,;\sigma^2,\sigma_e^2) = k_2(y)f_1(y\,;\sigma^2,\sigma_e^2) \\ \frac{\partial f_0}{\partial \sigma_e^2} &= \left(\frac{1}{2(\sigma_e^2)^2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\lambda_k} - \frac{N}{2}\frac{1}{\sigma_e^2}\right) f_0(y\,;\sigma_e^2) = k_3(y)f_0(y\,;\sigma_e^2)\end{aligned}$$

We write (19) as

$$a_1 p\frac{\partial f_1}{\partial \sigma^2} + a_2\left[p\frac{\partial f_1}{\partial \sigma_e^2} + (1-p)\frac{\partial f_0}{\partial \sigma_e^2}\right] + a_3\left[f_1 - f_0\right] = 0$$

Re-arranging and noticing that $f_1(y) > 0$ and $f_0(y) > 0$ for all y, we see that this is equivalent to saying that

$$\left\{\begin{aligned} a_1 p\, k_1(Y) + a_2 p\, k_2(Y) + a_3 &= 0 \\ a_2(1-p)k_3(Y) - a_3 &= 0 \end{aligned}\right. \tag{20}$$

Since $k_1(Y)$, $k_2(Y)$, and $k_3(Y)$ are non-zero with probability one, we have a contradiction because (20) is satisfied only if a is zero. □

Remark: Notice that (20) is satisfied for non-zero a if $p = 0$. This is also what we would expect since then we have no information on σ^2. Also, if $N = 1$, then $k_2(Y) = \lambda_1 k_1(Y)$, so if $p = 1$, (20) is satisfied as long as $a_1 + \lambda_1 a_2 = 0$ and $a_3 = 0$.

5.4 Note on a further generalization

An interesting article with relevance to our problem, is Kiefer and Wolfowitz (1956). It deals with the consistency of a maximum likelihood estimator when there are infinitely many incidental parameters present. These incidental parameters could be, in a generalization of our problem, the variance of the Brownian motion σ^2 if all diffusing particle have different diffusion coefficients. This corresponds to a so called poly-disperse solution in contrast to our present problem, which is mono-disperse (every particle has the same diffusion coefficient).

Assume that for each $i = 1, \ldots, n$, we have that Y_i is N-variate normally distributed random variable with mean zero and covariance matrix $\Sigma_i = I\sigma_i^2 + T\sigma_e^2$. Then, in the language of Kiefer and Wolfowitz (1956), the σ_i^2:s are the incidental parameters and σ_e^2 the parameter (even though, in our context, these names are misleading since we consider it to be the other way around). Notice that if the σ_i^2:s are constants and different for each i we only observe one increment vector Y_i for each σ_i^2. Obviously the estimates of the σ_i^2:s can not be consistent. It turns out however, that if we consider σ_i^2, $i = 1, \ldots, n$ to independent random variables with common (but unknown) distribution function F, and under certain assumption on F, the maximum likelihood estimator of F converges to F at every point of continuity of F, almost surely. Also, the maximum likelihood estimator of σ_e^2 is strongly consistent.

The case discussed in this section is of course a special case of these σ_i^2 coming from an unknown distribution function with only two values: let

$$F(x) = \begin{cases} 0 & \text{when } x < 0 \\ 1 - p & \text{when } 0 \leq x < \sigma^2 \\ 1 & \text{when } \sigma^2 \leq x \end{cases}$$

6 Application

The data from the example in the introduction were analysed with the EM algorithm. The positions of the 26 particles were estimated in two dimensions in each image using a circle detection algorithm. The total number of frames were 12, so $N = 11$.

By manual inspection, we concluded that three particles in Figure 1 seem to be fixed, and refer to them as particle 1 to 3, where 1 is the big white in the middle, 2 the big black to the left, and 3 the seemingly "false" particle, probably due to an optics defect, in the lower left corner. The remaining 23 are considered as diffusing particles.

6.1 Results

We applied the EM algorithm to the observed data with initial value $\theta^0 = [1\,,\,1\,,\,0.5]^T$. We stopped when the change of the Z_i:s between two consecutive E-steps was smaller than 10^{-6}. This criterion was satisfied after 3 steps with the resulting estimates

$$\begin{aligned} \hat{\sigma}^2 &= 2.2058 \\ \hat{\sigma}_e^2 &= 0.3172 \\ \hat{p} &= 0.8847 \end{aligned} \tag{21}$$

where the unit for the first two is pixel^2. The estimated classification variables $\hat{Z}_i$ were

$$\begin{aligned} \hat{Z}_1 &= 1.049 \cdot 10^{-5} \\ \hat{Z}_2 &= 1.528 \cdot 10^{-5} \\ \hat{Z}_3 &= 2.473 \cdot 10^{-3} \\ \hat{Z}_i &= 1.000 \quad \text{for} \quad i = 4, \ldots 26 \end{aligned} \tag{22}$$

in good correspondence with our manual classification.

6.2 Observed information matrix

Using the result (15) to compute the observed information matrix at the MLE

$$\begin{aligned} I(\hat{\sigma}^2, \hat{\sigma}_e^2, \hat{p}\,; Y) &= \begin{bmatrix} 33.75 & 52.75 & 0 \\ 52.75 & 476.6 & 0 \\ 0 & 0 & 254.9 \end{bmatrix} - \begin{bmatrix} 0.034 & 0.153 & -0.090 \\ 0.153 & 0.691 & -0.405 \\ -0.090 & -0.405 & 0.240 \end{bmatrix} \\ &= \begin{bmatrix} 33.72 & 52.59 & 0.090 \\ 52.59 & 475.9 & 0.405 \\ 0.090 & 0.405 & 254.7 \end{bmatrix} \end{aligned}$$

were the second term of the upper row corresponds to the missing information due to lack of the unobserved classification variables.

The inverse of this is

$$I^{-1}(\hat{\sigma}^2, \hat{\sigma}_e^2, \hat{p}\,; Y) = \begin{bmatrix} 0.0358 & -0.0040 & 0.0000 \\ -0.0040 & 0.0025 & 0.0000 \\ 0.0000 & 0.0000 & 0.0039 \end{bmatrix} \tag{23}$$

which gives us an approximate variance of the estimate of $\hat{\sigma}^2$ to

$$\mathbf{Var}\{\hat{\sigma}^2\} \simeq 0.0358 \tag{24}$$

6.3 Comparison with the theoretical diffusion coefficient

The estimated $\hat{\sigma}^2$ above corresponds to an estimated diffusion coefficient of

$$\hat{D} = 0.893\,\mu\text{m}^2/s$$

using the relationship between diffusion variance and diffusion coefficient, $\sigma^2 = 2D\tau$ and scaling to μm. Here, τ=0.040 s is the time interval between observations, and each pixel corresponds to a square with side M=180μm.

The asymptotic normality result from Section 5.3 can be used to give an approximate 95%-confidence interval of D:

$$D = \hat{D} \pm 1.96 \cdot \frac{M^2}{2\tau}\sqrt{.0358} = .893 \pm .150\,\mu\text{m}^2/s \tag{25}$$

The theoretical diffusion coefficient is given by Stoke-Einstein's relation (see for example Evans and Wennerström (1999) pages 370-372)

$$D = \frac{k_B T}{6\pi\eta R_H}$$

where k_B is Bolzmann's constant, η the viscosity of the solution, T the temperature and R_H the hydrological radius of the particle.

The appropriate values for the viscosity and temperature are η=0.9 mPa and T=298 K. The geometric radius of the particles are 247 nm and we used this as the hydrological radius, even if the latter is often a bit larger than the former. Plugging this into (6.3) gives us

$$D = 0.982\,\mu\text{m}^2/s$$

Comparing with the confidence interval in (25), we see that the theoretical diffusion coefficient is within this interval.

6.4 Simulation of the approximate distribution of the estimates

We simulated 1000 time series with 26 particles, of which 3 were fixed, over 12 frames in two dimension, using the estimated values of $\sigma^2 = 2.2058$ and $\sigma_e^2 = 0.3172$ from (21) as the true diffusion variance and error variance. For each time series, we used the EM algorithm to estimate σ^2 and σ_e^2.

The histograms of the estimated values are displayed in Figure 2. The mean and empirical covariance of the 1000 estimates of σ^2 and σ_e^2 were

$$\bar{\hat{\sigma}}^2 = 2.2054$$
$$\bar{\hat{\sigma}}_e^2 = 0.3185$$

and

$$\begin{bmatrix} .0348 & -.0040 \\ -.0040 & .0027 \end{bmatrix}$$

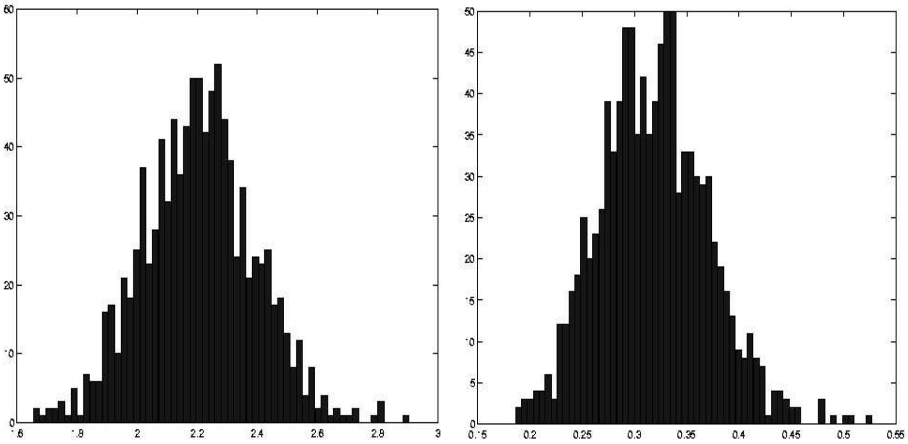

Figure 2: The histograms of the estimated $\hat{\sigma}^2$ and $\hat{\sigma}_e^2$ using the EM algorithm from 1000 simulations using 2.2058 and 0.3172 as true values.

in good agreement with the true values of $\sigma^2 = 2.2058$ and $\sigma_e^2 = 0.3172$ and the inverse of the observed information matrix in (23).

Acknowledgements

This work was financially supported by the Swedish Research Council for Engineering Sciences (TFR) and the Swedish Research Council (VR). The author wishes to thank Mats Rudemo for his supervision and numerous invaluable comments, Ib Skovgaard for interesting and helpful discussions regarding the asymptotics, Lennart Lindfors of AstraZeneca R&D for producing the images and keeping me in touch with the application, and Olle Häggström for reading through the draft.

References

S.-I. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, and C. Rao. *Differential geometry in Statistical Inference.* Lecture Notes - Monograph series. Institute of Mathematical Statistics, first edition, 1987.

O. Barndorff-Nielsen and D. Cox. *Inference and Asymptotics.* Chapman & Hall, London, first edition, 1994.

J. Behboodian. On a mixture of normal distributions. *Biometrika*, 57:215–217, 1970.

N. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

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

B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency)(with discussion). *The Annals of Statistics*, 3:1189–1242, 1975.

B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6: 362–376, 1978.

D. Evans and H. Wennerström. *The colloidal domain. Where Physics, Chemistry, Biology, and Technology meet.* Wiley-VCH, New York, second edition, 1999.

V. Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64:1459–1471, 1969.

R. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13:795–800, 1985.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:887–906, 1956.

J. Lindsey. *Parametric Statistical Inference.* Oxford University Press, Oxford, first edition, 1996.

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

P. McCullagh. *Tensor Methods in Statistics.* Chapman & Hall, London, 1987.

G. McLachlan and T. Krishnan. *The EM algorithm and Extensions.* Wiley-Interscience, New York, first edition, 1997.

G. McLachlan and D. Peel. *Finite mixture models.* Wiley-Interscience, New York, first edition, 2000.

R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9:225–228, 1981.

R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1:49–58, 1974.

R. Sundberg. An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics. Part B. Simulation and Computation*, 5:55–64, 1976.

A. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, U.K., first edition, 1999.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601, 1949.

B. Wei. *Exponential Family Nonlinear Models.* Springer, Singapore, first edition, 1998.

Mats Kvarnström, Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University, Göteborg, SE-412 96, Sweden
Email: matskv@math.chalmers.se

Appendix A: Sufficient statistics

Consider the complete data density (5). Take the logarithm and use the transformed increment vectors $\tilde{Y}$ (see section 2.2) for easier notation

$$\begin{aligned}
\log g_c = {} & z\log p - \frac{z}{2}\sum_{k=1}^{N}\log(\sigma^2+\lambda_k\sigma_e^2) - \frac{z}{2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\sigma^2+\lambda_k\sigma_e^2} + \\
& + (1-z)\log(1-p) - \frac{1-z}{2}\sum_{k=1}^{N}\log(\sigma_e^2\lambda_k) - \frac{1-z}{2}\sum_{k=1}^{N}\frac{\tilde{y}_k^2}{\sigma_e^2\lambda_k} \\
& = \sum_{k=1}^{N} z\tilde{y}_k^2\left(-\frac{1}{2}\frac{1}{\sigma^2+\lambda_k\sigma_e^2}\right) - \frac{1}{2\sigma_e^2}\sum_{k=1}^{N}\frac{(1-z)\tilde{y}_k^2}{\lambda_k} \\
& + z\left(\log(\frac{p}{1-p}) - \frac{1}{2}\sum_{k=1}^{N}\log(\frac{\sigma^2+\lambda_k\sigma_e^2}{\lambda_k\sigma_e^2})\right) - \left(\frac{1}{2}\sum_{k=1}^{N}\log(\lambda_k\sigma_e^2) - \log(1-p)\right)
\end{aligned}$$

and we see that a minimal sufficient statistic can be chosen to be

$$\begin{aligned} t_1 &= z\tilde{y}_1^2 \\ &\vdots \\ t_N &= z\tilde{y}_N^2 \\ t_{N+1} &= \sum_{k=1}^{N} \frac{(1-z)\tilde{y}_k^2}{\lambda_k} \\ t_{N+2} &= z \end{aligned}$$

with the corresponding canonical parameter α

$$\begin{aligned} \alpha_1 &= -\frac{1}{2}\frac{1}{\sigma^2+\lambda_1\sigma_e^2} \\ &\vdots \\ \alpha_N &= -\frac{1}{2}\frac{1}{\sigma^2+\lambda_N\sigma_e^2} \\ \alpha_{N+1} &= -\frac{1}{2\sigma_e^2} \\ \alpha_{N+2} &= \log(\frac{p}{1-p}) - \frac{1}{2}\sum_{k=1}^{N}\log(\frac{\sigma^2+\lambda_k\sigma_e^2}{\lambda_k\sigma_e^2}) \end{aligned}$$

which is a function of our parameter θ. Since this is 3-dimensional and the sufficient statistics is $(N+2)$-dimensional, we say that the complete data belongs to a curved exponential family or, with the terminology of Barndorff-Nielsen and Cox (1994), a $(N+2,3)$-exponential model.

Solving for p in the expression for α_{N+2} above, we get

$$p = \frac{e^{\alpha_{N+2}}\prod_{k=1}^{N}\left(\frac{\sigma^2+\lambda_k\sigma_e^2}{\lambda_k\sigma_e^2}\right)^{1/2}}{1+e^{\alpha_{N+2}}\prod_{k=1}^{N}\left(\frac{\sigma^2+\lambda_k\sigma_e^2}{\lambda_k\sigma_e^2}\right)^{1/2}} = \frac{e^{\alpha_{N+2}}\prod_{k=1}^{N}\left(\frac{\alpha_{N+1}}{\lambda_k\alpha_k}\right)^{1/2}}{1+e^{\alpha_{N+2}}\prod_{k=1}^{N}\left(\frac{\alpha_{N+1}}{\lambda_k\alpha_k}\right)^{1/2}}$$

and we can write the complete data density as

$$\log g_c = \alpha^T t - k(\alpha) \tag{26}$$

where $\alpha = \alpha(\theta)$ and k is

$$k(\alpha) = \frac{1}{2}\log(N+1) - \frac{N}{2}\log(-2\alpha_{N+1}) + \log\Big(1 + e^{\alpha_{N+2}}\prod_{k=1}^{N}\Big(\frac{\alpha_{N+1}}{\lambda_k\alpha_k}\Big)^{1/2}\Big) \tag{27}$$

From standard theory of exponential families, we get the cumulants of the sufficient statistics by differentiating $k(\alpha)$. In particular, we have $\mathbf{E}T = \frac{\partial k}{\partial \alpha}$ and $\mathbf{Var}\{T\} = \frac{\partial^2 k}{\partial \alpha \partial \alpha^T}$, which we denote μ and V, respectively.

The expectation of the sufficient statistics

$$\mathbf{E}T = \begin{bmatrix} p(\sigma^2 + \lambda_1 \sigma_e^2) \\ \vdots \\ p(\sigma^2 + \lambda_N \sigma_e^2) \\ (1-p)N\sigma_e^2 \\ p \end{bmatrix}$$

Appendix B: Geometry of the complete data

Differentiating the complete data likelihood once, we get the score function

$$\frac{\partial \log g_c}{\partial \theta} = (\frac{\partial \alpha}{\partial \theta^T})^T T - \frac{\partial k}{\partial \theta} = (\frac{\partial \alpha}{\partial \theta^T})^T (T - \mathbf{E}T)$$

where we used

$$\frac{\partial k}{\partial \theta} = (\frac{\partial \alpha}{\partial \theta^T})^T \frac{\partial k}{\partial \alpha}$$

and $\partial k / \partial \alpha = \mathbf{E}T$.

Hence at the MLE $\hat{\theta}$, the difference vector $T - \mu$ is orthogonal to the derivative of the canonical parameter α with respect to the parameter θ. This relation and other geometrical interpretations of the maximum likelihood estimate in a curved exponential family, were, to the author's knowledge, first made by Efron in two groundbreaking articles, Efron (1975, 1978). Since then a lot of research has been made in this area with fruitful connections between statistics and differential geometry, see for example Amari et al. (1987), McCullagh (1987) and Barndorff-Nielsen and Cox (1994)

Appendix C: Iterative scheme for complete data MLE

Exploring the geometrical relations of the curved exponential further, brings us to a Newton-Raphson style of iterative method of finding the maximum likelihood estimate of θ, given the complete data. Even if the unobserved classification variables Z_i are not available to us, the M-step in the EM algorithm (see section 4.1.3) maximizes the complete data likelihood using the estimated $\hat{Z}_i$:s from the E-step. The idea comes from Wei (1998).

We adopt the notation of Wei:

$$\begin{aligned}\mu(\theta) &= \mathbf{E}_\theta T\\ V(\theta) &= \mathbf{Var}_\theta\{T\}\\ D_\alpha &= \frac{\partial\alpha}{\partial\theta^T}\\ D &= \frac{\partial\mu}{\partial\theta^T}\end{aligned}$$

and derive the following identity

$$D = \frac{\partial\mu}{\partial\theta^T} = \frac{\partial\mu}{\partial\alpha^T}\frac{\partial\alpha}{\partial\theta^T} = VD_\alpha$$

Write $l_c = \log L_c$ and let $\dot{l}_c$ and $\ddot{l}_c$ denote the first and second derivative of l_c with respect to θ.

Expressing the score function $\dot{l}_c$ as

$$\begin{aligned}\dot{l}_c &= \frac{\partial l_c}{\partial\theta} = \left(\frac{\partial\alpha}{\partial\theta}\right)^T\frac{\partial l_c}{\partial\alpha} = \left(\frac{\partial\alpha}{\partial\theta}\right)^T(T-\mu)\\ &= D_\alpha(T-\mu) = D^TV^{-1}(T-\mu)\end{aligned} \tag{28}$$

and the score equation at $\hat{\theta}$ can be written

$$0 = D^TV^{-1}(T-\mu)$$

where D, V, and μ are evaluated at $\hat{\theta}$.

Differentiate $\dot{l}_c$ once again and we get

$$\begin{aligned}\ddot{l}_c &= \left(\frac{\partial l_c}{\partial\alpha}\right)^T\left[\frac{\partial^2\alpha}{\partial\theta\partial\theta^T}\right] - D_\alpha^TVD_\alpha\\ &= (T-\mu)^T\left[\frac{\partial^2\alpha}{\partial\theta\partial\theta^T}\right] - D_\alpha^TVD_\alpha\end{aligned} \tag{29}$$

Now, Newton's classical iterative scheme can be written

$$\theta_{i+1} = \theta_i + [-\ddot{l}_c(\theta_i)]^{-1}\dot{l}_c(\theta_i)$$

Using the expressions in (28) for $\dot{l}_c$ together with $\mathbf{E}_\theta\{-\ddot{l}_c\} = D_\alpha^TVD_\alpha$ instead of $\ddot{l}_c$ we get

$$\begin{aligned}\theta_{i+1} &= \theta_i + \left[D_\alpha^TVD_\alpha\right]^{-1}D_\alpha(T-\mu)\\ &= \theta_i + \left[D^TV^{-1}D\right]^{-1}V^{-1}D(T-\mu)\end{aligned} \tag{30}$$

We find V by differentiating k twice with respect to α. The matrix turns out to be quite complicated with all its elements different from zero. In the iteration scheme (30), it is the inverse V^{-1} we need, and since this is in fact much less complicated, we present it here

$$V^{-1} = \begin{bmatrix} \frac{1}{2p(\sigma^2+\lambda_1\sigma_e^2)^2} & \cdots & 0 & 0 & -\frac{1}{2p(\sigma^2+\lambda_1\sigma_e^2)} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{2p(\sigma^2+\lambda_N\sigma_e^2)^2} & 0 & -\frac{1}{2p(\sigma^2+\lambda_N\sigma_e^2)} \\ 0 & \cdots & 0 & \frac{1}{2(1-p)N(\sigma_e^2)^2} & \frac{1}{2(1-p)\sigma_e^2} \\ -\frac{1}{2p(\sigma^2+\lambda_1\sigma_e^2)} & \cdots & -\frac{1}{2p(\sigma^2+\lambda_N\sigma_e^2)} & \frac{1}{2(1-p)\sigma_e^2} & \frac{2+N}{2p(1-p)} \end{bmatrix}$$

Appendix D: Complete data asymptotics

In applications it may happen that you label the particles manually as diffusing or fixed or defect particles and want to estimate the parameters. Then our problem becomes easier, mainly because the likelihood is composed of a product.

From (29), we get the expected information matrix to the complete data

$$\begin{aligned} \mathcal{I}_c(\theta) &= D^T V^{-1} D = D^T D_\alpha \\ &= \begin{bmatrix} \frac{p}{2}\sum_{k=1}^N \frac{1}{(\sigma^2+\lambda_k\sigma_e^2)^2} & \frac{p}{2}\sum_{k=1}^N \frac{\lambda_k}{(\sigma^2+\lambda_k\sigma_e^2)^2} & 0 \\ \frac{p}{2}\sum_{k=1}^N \frac{\lambda_k}{(\sigma^2+\lambda_k\sigma_e^2)^2} & \frac{p}{2}\sum_{k=1}^N \frac{\lambda_k^2}{(\sigma^2+\lambda_k\sigma_e^2)^2} + \frac{N(1-p)}{2(\sigma_e^2)^2} & 0 \\ 0 & 0 & \frac{1}{p(1-p)} \end{bmatrix} \end{aligned}$$

which is positive definite for all $\theta \in \mathrm{int}(\Omega)$. To see this, apply the Cauchy-Schwarz inequality on the upperleft 2 by 2 matrix elements.

For $\theta_0 \in \mathrm{int}(\Omega_c)$ we get strong consistency from Wald (1949), and since also $\mathcal{I}_c(\theta_0)$ is positive definite, all conditions for asymptotic normality are satisfied.