SORT 39 (1) January-June 2015, 109-126

A comparison of computational approaches for maximum likelihood estimation of the Dirichlet parameters on high-dimensional data

Marco Giordan and Ron Wehrens[1]

Abstract

Likelihood estimates of the Dirichlet distribution parameters can be obtained only through numerical algorithms. Such algorithms can provide estimates outside the correct range for the parameters and/or can require a large amount of iterations to reach convergence. These problems can be aggravated if good starting values are not provided. In this paper we discuss several approaches that can partially avoid these problems providing a good trade off between efficiency and stability. The performances of these approaches are compared on high-dimensional real and simulated data.

MSC: 62F10, 65C60

Keywords: Levenberg-Marquardt algorithm, re-parametrization, starting values, metabolomics data.

1. Introduction

The Dirichlet distribution has multiple applications. It is well known for being the conjugate prior of the multinomial distribution and can be therefore used to get Bayesian estimates of the multinomial parameters. It is the basis for complicated models such as Dirichlet Processes and mixture distributions based on the Dirichlet distribution. In addition, it is interesting in its own right. It can be used to analyse positive continuous data that sum up to one, i.e. compositional data. Such kinds of data can arise in many situations. For example, when the data in each unit are represented by an intensity signal it can be of interest to normalize them through the total intensity of that unit. In this way

[1] Biostatistics and Data Management. IASMA Research and Innovation Centre. marco.giordan@fmach.it, ron.wehrens@wur.nl
Received: April 2014
Accepted: February 2015

the data from different units are comparable and the final data can be analysed using the Dirichlet distribution. Another possible application is in the analysis of taxonomic assignements. For each unit the percentages of the microbial composition of the unit are assigned to the specific taxonomies. These data can be easily produced with Next Generation Sequencing (NGS) technologies. Many other examples of the use of the Dirichlet distribution are provided in the paper of Wicker et al. (2008).

The use of the Dirichlet distribution has been criticized due to the strong independence properties associated to this distribution (Aitchison, 1986). In the literature some generalizations have been proposed to overcome these limits, see for example Connor and Mosiman (1969) and Rayens and Srinivasan (1994). In this paper we only marginally discuss this point giving a case where the application of the Dirichlet distribution to high-dimensional compositional data leads to reliable conclusions. The focus of paper is related to the comparison of the computational performances of different methods to get maximum likelihood estimates of the Dirichlet distribution. There is no closed form solution of the maximum likelihood equations, therefore numerical methods must be employed. At the moment, commonly adopted methods are rather unstable. Final estimates can be outside the correct range for the parameters and the algorithms can fail to reach convergence in a reasonable amount of time. Wicker et al. (2008) reported many convergence failures in their simulation studies. Strategies to improve stability have been studied by many authors. Many proposals are focused on the choice of good starting values for the optimization algorithms. Useful references for these problems and the study of Dirichlet maximum likelihood estimation are the papers of Dishon and Weiss (1980), Ronning (1989), Narayanan (1991a), and Narayanan (1991b).

In this work we compare eight different algorithms and four different initialisation methods on real and simulated data. As an application, we consider the analysis of metabolomics data. We consider the Newton-Raphson algorithm as the reference algorithm. A fixed-point algorithm, shown in literature to have very good performance, is taken into account as well. Moreover, a novel and more stable algorithm based on Levenberg-Marquardt ideas (Levenberg, 1944; Marquardt, 1963) will be employed to get the final maximum likelihood estimates. In the appendix we give a proof of the convergence to the optimum for this algorithm. Finally, to avoid the problem of estimates outside the admissible parameter space, a simple re-parametrization of the Dirichlet parameters and an algorithm with box constraints are considered. The re-parametrization will be used together with the Newton-Raphson algorithm and the Levenberg-Marquardt algorithm, but not with the FPI algorithm because this does not suffer from the problem of a constrained parameter space (see Huang, 2005). The re-parametrization and the algorithm with box-constraints are straightforward but have not been considered in the literature before.

2. Dirichlet likelihood

In this section we introduce notation and we summarize useful results from literature (see Minka, 2000). If $\mathbf{y} = (y_1, \ldots, y_K)^\mathsf{T}$ is Dirichlet distributed with vector parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\mathsf{T}$ then its density is

$$\frac{\Gamma(\sum_k \alpha_k)}{\Pi_k \Gamma(\alpha_k)} \Pi_k y_k^{\alpha_k - 1}$$

where $\alpha_k > 0$, $y_k > 0$, $k = 1, \ldots, K$ and $\sum_k y_k = 1$.

The log-likelihood for N independent observations can be written as

$$f(\boldsymbol{\alpha}) = N \ln \Gamma \left(\sum_k \alpha_k \right) - N \sum_k \ln \Gamma(\alpha_k) + N \sum_k (\alpha_k - 1) \frac{1}{N} \sum_i \ln y_{ik}. \tag{1}$$

The gradient of the log-likelihood with respect to one α_k is:

$$[\nabla f(\boldsymbol{\alpha})]_k = N \left(\Psi(\sum_k \alpha_k) - \Psi(\alpha_k) + \frac{1}{N} \sum_i \ln y_{ik} \right) \tag{2}$$

where Ψ denotes the digamma function. In what follows the arguments of a function (e.g. the parameters $\boldsymbol{\alpha}$ for the function $f(\cdot)$ in equation (2)) can be suppressed in the notation when this does not generate confusion. The Hessian can be written in matrix form as

$$\mathbf{H} = \mathbf{Q} + \mathbf{1}\mathbf{1}^\mathsf{T} z \tag{3}$$

$$q_{jk} = -N \Psi'(\alpha_k) \delta(j - k) \tag{4}$$

$$z = N \Psi'(\sum_k \alpha_k) \tag{5}$$

where Ψ' denotes the trigamma function and δ is the Dirac function (zero on the real line, except at the origin where it is one). Let us note that the diagonal form of $\mathbf{Q}$ assures the existence of its inverse when the diagonal elements are different from zero. This is exactly the present case because the trigamma function is positive for positive real arguments.

2.1. Some preliminary considerations

The number of available algorithms to maximize a function is huge and its impossible to summarize all of them in a meaningful way. In this work we have focused our

attention on algorithms suggested by the literature about the Dirichlet distribution and on a modification of the Levenberg-Marquardt algorithm for which we are able to provide a theoretical result related to the properties of the Dirichlet distribution.

In the literature the algorithms studied and suggested are mainly two: the Newton-Raphson algorithm and the Fixed Point Iteration algorithm (see Minka, 2000; Huang, 2005). We include them in our comparison and we describe them briefly in the following sections. Other algorithms, like BFGS (Ronning, 1989) or Gradient Ascent (Huang, 2005), had a poor performance in the literature and are therefore disregarded here. The Levenberg-Marquardt algorithm is an algorithm to find least-squares estimates. In the appendix we give a theoretical result of convergence for a modification of the Levenberg-Marquardt algorithm with a fixed damping parameter when we apply its adaptation to find maximum likelihood estimates. We study its performance in the paper through simulations and real data. Finally, to avoid the problem of estimates outside the allowed space, we consider a re-parametrization of the Dirichlet distribution and an implementation of the BFGS algorithm with bounding box constraints, L-BFGS-B (see Byrd et al., 1995).

In this work the efficiency of the algorithms will be compared essentially by the number of iterations required to reach convergence. This number does not depend on the implementation of the code and in this sense it is objective. We will see that the Levenberg-Marquardt approach can be thought as a penalized version of the Newton-Raphson algorithm and therefore it is expected to be slightly slower. However in general the *iteration time* for different algorithms can vary substantially and the algorithm with fewest iterations for convergence can require the largest amount of time. This can be the case when we want to compare Levenberg-Marquard, FPI and L-BFGS-B, therefore for these algorithms we provide also a comparison on time.

2.2. Newton-Raphson algorithm

The Newton-Raphson (NR) algorithm is used to solve the maximum likelihood equations $[\nabla f(\boldsymbol{\alpha})] = \mathbf{0}$. It can be summarized by the following equations:

$$\boldsymbol{\alpha}^{\text{new}} = \boldsymbol{\alpha}^{\text{old}} - \mathbf{H}^{-1}\nabla f(\boldsymbol{\alpha}^{\text{old}}) \tag{6}$$

$$\mathbf{H}^{-1} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{Q}^{-1}}{\frac{1}{z} + \mathbf{1}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{1}} \tag{7}$$

$$\left[\mathbf{H}^{-1}\nabla f(\boldsymbol{\alpha})\right]_k = \frac{[\nabla f(\boldsymbol{\alpha})]_k - b}{q_{kk}} \tag{8}$$

$$b = \frac{\mathbf{1}^{\mathsf{T}}\mathbf{Q}^{-1}\nabla f(\boldsymbol{\alpha})}{\frac{1}{z} + \mathbf{1}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{1}} = \frac{\sum_j [\nabla f(\boldsymbol{\alpha})]_j / q_{jj}}{1/z + \sum_j 1/q_{jj}}. \tag{9}$$

When $\mathbf{Q}$ is invertible the inversion of the matrix $\mathbf{H}$ is always guaranteed by the use of the Sherman-Morrison formula. However the equations (8) and (9) highlight that the algorithm does not require the storage and the computation of the inverse of $\mathbf{H}$. This is a great advantage, especially when the number of variables is high.

The Newton-Raphson algorithm is expected to converge to the global optimum because the Dirichlet distribution belongs to the exponential family and is therefore concave. However, the convergence can be very slow and the final estimates can be outside the admissible range for the parameters. Good starting values for the algorithm can partially avoid these problems.

Finally let us note that using the relationship (7) is easy to build marginal confidence intervals based on the observed Fisher information at the maximum-likelihood estimate.

2.3. Fixed Point Iteration algorithm

A fixed point iteration (FPI) scheme was considered initially by Minka (2000) and later by Huang (2005) to get the maximum likelihood estimates of the Dirichlet distribution. It is based upon minorize maximize (MM) algorithms (see Lange, 2010) and in our case the minorizing function of the log-likelihood employs an inequality of the gamma function. Specifically the log-likelihood can be bounded as follows

$$\frac{1}{N} f\left(\boldsymbol{\alpha}\right) \geq \left(\sum_k \alpha_k\right) \Psi\left(\sum_k \alpha_k^{\text{old}}\right) - \sum_k \ln\Gamma\left(\alpha_k\right) + \sum_k \left(\alpha_k - 1\right) \frac{1}{N} \sum_i \ln y_{ik} + C$$

where C is a constant. This leads to the following equation that must be solved:

$$\Psi(\alpha_k^{\text{new}}) = \Psi(\sum_k \alpha_k^{\text{old}}) + \frac{1}{N} \sum_i \ln y_{ik}. \tag{10}$$

To get α_k^{new} we need to invert the digamma function and this is done using another iterative algorithm; therefore the whole procedure can be slow.

2.4. Starting values

The Newton-Raphson method is based upon a Taylor approximation and therefore good convergence properties are guaranteed only if the initial starting value is in a neighbourhood of the true parameter. In the literature there are many suggestions to find good starting values. We review four of them that will be used throughout the paper. We use the following notation: $\bar{y}_k = \frac{1}{N}\sum_{i=1}^N y_{ik}$, $\bar{y}_k^{(2)} = \frac{1}{N}\sum_{i=1}^N y_{ik}^2$ and $s_k^2 = \bar{y}_k^2 - (\bar{y}_k)^2$. The four initialisations will be indicated as:

moments. Matching the first two moments of the Dirichlet with the empirical moments provides the useful initialisation:

$$[\boldsymbol{\alpha}^{\text{start}}]_k = \bar{y}_k \frac{\bar{y}_k - \bar{y}_k^{(2)}}{s_k^2}. \tag{11}$$

Such initialisation employs only the marginal distributions and therefore its use of the information can be inefficient.

Ronning. Ronning (1989) proposed an initialisation to guarantee parameters in the correct range after the first iteration of the Newton-Raphson algorithm. There is however no warranty that the final estimates are in the correct range. Such initialisation gives the same value to all the parameters and unlike the moments method uses all the available data for initialising each parameter:

$$[\boldsymbol{\alpha}^{\text{start}}]_k = \min_{i\in 1,\ldots,N,\ k\in 1,\ldots,K} y_{ik}. \tag{12}$$

Dishon. Following a suggestion of Dishon and Weiss (1980), Ronning (1989) proposed a modification to the method of moments using information from all the marginals. Each parameter is estimated through:

$$[\boldsymbol{\alpha}^{\text{start}}]_k = \hat{\alpha}_0 \bar{y}_k \tag{13}$$

$$\hat{\alpha}_0 = \left\{ \Pi_{k=1}^{K-1} \left(\frac{\bar{y}_k(1-\bar{y}_k)}{s_k^2} - 1 \right) \right\}^{1/(K-1)}. \tag{14}$$

This initialisation can give parameters outside the admissible region.

Wicker. Recently Wicker et al. (2008) proposed an initialisation based on an asymptotic approximation of the likelihood. This approximation uses the limiting behaviour of the digamma function when its real argument goes to zero or infinity. Such situations are met when the number of parameters goes to infinity, i.e. for high-dimensional data. However in their simulation study Wicker et al. (2008) considered only a five-dimensional setting. Their initialisation is given by

$$[\boldsymbol{\alpha}^{\text{start}}]_k = \hat{\alpha}_0 \bar{y}_k \tag{15}$$

$$\hat{\alpha}_0 = \frac{N(K-1)\Psi(1)}{N\sum_{k=1}^K \bar{y}_k \ln \bar{y}_k - \sum_{k=1}^K \bar{y}_k \sum_{i=1}^N \ln y_{ik}}. \tag{16}$$

3. A re-parametrization

An algorithm producing estimates outside the correct range is useless. In the case of the Dirichlet distribution this problem can be avoided using a simple re-parametrization. The idea is to see the parameters $\boldsymbol{\alpha}$ as functions of other parameters free to vary on the real line: in the log-likelihood we replace α_k with $\exp(\beta_k)$. In what follows we will indicate with expNR the use of the NR algorithm applied to this re-parametrization. This way, the range of β_k is the real line and $\exp(\beta_k)$ is in the correct range for α_k. With these replacements the log-likelihood $f(\boldsymbol{\beta})$ is $f(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = \exp(\boldsymbol{\beta}) = (\exp(\beta_1), \ldots, \exp(\beta_K))^\mathsf{T}$. The gradient can now be expressed as:

$$[\nabla f(\boldsymbol{\beta})]_k = [\nabla f(\boldsymbol{\alpha})]_k \Big|_{\exp(\boldsymbol{\beta})} \exp(\beta_k).$$

The Hessian has a form similar to the original one:

$$\mathbf{H} = \mathbf{Q} + \exp(\boldsymbol{\beta}) \exp(\boldsymbol{\beta})^\mathsf{T} z \tag{17}$$

$$q_{jk} = \left([\nabla f(\boldsymbol{\beta})]_k - \exp(\beta_k + \beta_j)\Psi'(\exp(\beta_k))\right) \delta(j-k)N \tag{18}$$

$$z = N\Psi'\left(\sum_k \exp(\beta_k)\right) \tag{19}$$

As before, the diagonal form of the square matrix $\mathbf{Q}$ and the Sherman-Morrison formula are sufficient to guarantee that the inverse of $\mathbf{H}$ exists if the diagonal elements of $\mathbf{Q}$ are non zero. However, after the re-parametrization the problem is not necessarily concave in the new parameters. Therefore we cannot say that the inequalities of the previous section hold true also now. For the re-parametrization the elements in the diagonal of $\mathbf{Q}$ are strictly negative in a neighbourhood of the point of maximum. Indeed at the maximum the gradient is zero and therefore $q_{kk} = -\exp(2\beta_k)\Psi'(\exp(\beta_k))N < 0$. Within such neighbourhood we have:

$$\mathbf{H}^{-1} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1} \exp(\boldsymbol{\beta}) \exp(\boldsymbol{\beta})^\mathsf{T} \mathbf{Q}^{-1}}{\frac{1}{z} + \exp(\boldsymbol{\beta})^\mathsf{T} \mathbf{Q}^{-1} \exp(\boldsymbol{\beta})} \tag{20}$$

$$\left[\mathbf{H}^{-1} \nabla f(\boldsymbol{\beta})\right]_k = \frac{1}{q_{kk}} \left([\nabla f(\boldsymbol{\beta})]_k - \exp(\beta_k) b\right) \tag{21}$$

$$b = \frac{\exp(\boldsymbol{\beta})^\mathsf{T} \mathbf{Q}^{-1} \nabla f(\boldsymbol{\beta})}{\frac{1}{z} + \exp(\boldsymbol{\beta})^\mathsf{T} \mathbf{Q}^{-1} \exp(\boldsymbol{\beta})} \tag{22}$$

$$= \frac{\sum_j \exp(\beta_j) [\nabla f(\boldsymbol{\beta})]_j / q_{jj}}{1/z + \sum_j \exp(2\beta_j)/q_{jj}}. \tag{23}$$

Equations (21), (22) and (23) assure an easy way to implement the Newton-Raphson algorithm avoiding explicit matrix inversion. With these new quantities the Newton-Raphson iteration is again:

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \mathbf{H}^{-1}\nabla f(\boldsymbol{\beta}^{\text{old}}).$$

As starting values for the algorithm we can consider the logarithms of the starting values previously described.

4. A stable algorithm

The Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) was originally proposed to solve non-linear least-squares minimization problems. The idea is to use a function of $\mathbf{H}$ instead of $\mathbf{H}$ itself. With an appropriate choice the iterations of the algorithm can take into account the curvature of the function being optimized. For the optimization of the Dirichlet log-likelihood we propose an iteration algorithm similar to the Levenberg-Marquardt one (LM):

$$\boldsymbol{\alpha}^{\text{new}} = \boldsymbol{\alpha}^{\text{old}} - \{\mathbf{H} + \gamma \text{diag}\mathbf{H}\}^{-1}\nabla f(\boldsymbol{\alpha}^{\text{old}}) \tag{24}$$

where γ is a positive constant or a positive function not depending on the parameters. The effect of the damping parameter γ is that of shortening the steps of NR algorithm, providing in this way prudent steps in the iterations. The same algorithm can be applied to the re-parametrization (expLM):

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \{\mathbf{H} + \gamma \text{diag}\mathbf{H}\}^{-1}\nabla f(\boldsymbol{\beta}^{\text{old}}). \tag{25}$$

Let us note that a similar approach is backtracking. In backtracking the matrix approximating the Hessian is multiplied by a damping parameter that is eventually shrunken to assure an ascent step. The damping parameter influences the step length. The rational of this approach is related to the Taylor expansion of the gradient calculated at the new parameter. We prefer instead the Levenberg-Marquardt approach because the damping parameter can influence both the direction and the size of the step (Madsen et al., 2004). Let us denote with $\mathbf{x}$ the parameters of interest ($\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ in the previous cases); working on the iteration map $\mathbf{M}(\mathbf{x})$ defined by

$$\mathbf{M}(\mathbf{x}) = \mathbf{x} - \{\mathbf{H}(\mathbf{x}) + \gamma \text{diag}\mathbf{H}(\mathbf{x})\}^{-1}\nabla f(\mathbf{x}) \tag{26}$$

we show in the appendix that both algorithms (24) and (25) converge to the maximum.

Similarly to what we have seen in the previous sections for the Newton-Raphson algorithm, we show how to rearrange the quantities involved in this version of the Levenberg-Marquardt algorithm using expressions without an explicit use of inverse matrices. Equations (24) and (25) can be rewritten with the following quantities

$$\{\mathbf{H}(\mathbf{x})+\gamma \text{diag}\mathbf{H}(\mathbf{x})\}^{-1} = \mathbf{D}-\mathbf{L} \tag{27}$$

$$\mathbf{D} = \{\mathbf{Q}(\mathbf{x})+\gamma \text{diag}\mathbf{H}(\mathbf{x})\}^{-1} \tag{28}$$

where for the original parametrization

$$\mathbf{L} = \frac{\mathbf{D}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{D}}{\frac{1}{z}+\mathbf{1}^{\mathsf{T}}\mathbf{D}\mathbf{1}} \tag{29}$$

$$[(\mathbf{D}-\mathbf{L})\nabla f(\boldsymbol{\alpha})]_k = \frac{[\nabla f(\boldsymbol{\alpha})]_k - b}{q_{kk}(1+\gamma)+\gamma z} \tag{30}$$

$$b = \frac{\mathbf{1}^{\mathsf{T}}\mathbf{D}\nabla f(\boldsymbol{\alpha})}{\frac{1}{z}+\mathbf{1}^{\mathsf{T}}\mathbf{D}\mathbf{1}} \tag{31}$$

$$= \frac{\sum_k [\nabla f(\boldsymbol{\alpha})]_k/(q_{kk}(1+\gamma)+\gamma z)}{1/z+\sum_k 1/(q_{kk}(1+\gamma)+\gamma z)} \tag{32}$$

while for the re-parametrization

$$\mathbf{L} = \frac{\mathbf{D}\exp(\boldsymbol{\beta})\exp(\boldsymbol{\beta})^{\mathsf{T}}\mathbf{D}}{\frac{1}{z}+\exp(\boldsymbol{\beta})^{\mathsf{T}}\mathbf{D}\exp(\boldsymbol{\beta})} \tag{33}$$

$$[(\mathbf{D}-\mathbf{L})\nabla f(\boldsymbol{\beta})]_k = \frac{[\nabla f(\boldsymbol{\beta})]_k - \exp(\beta_k)b}{q_{kk}(1+\gamma)+\gamma z\exp(2\beta_k)} \tag{34}$$

$$b = \frac{\exp(\boldsymbol{\beta})^{\mathsf{T}}\mathbf{D}\nabla f(\boldsymbol{\beta})}{\frac{1}{z}+\exp(\boldsymbol{\beta})^{\mathsf{T}}\mathbf{D}\exp(\boldsymbol{\beta})} \tag{35}$$

$$= \frac{\sum_k \frac{\exp(\beta_k)[\nabla f(\boldsymbol{\beta})]_k}{(q_{kk}(1+\gamma)+\gamma z\exp(2\beta_k))}}{\frac{1}{z}+\sum_k \frac{\exp(2\beta_k)}{(q_{kk}(1+\gamma)+\gamma z\exp(2\beta_k))}}. \tag{36}$$

All the equalities are valid when we can apply the Sherman-Morrison formula, see Sections 2 and 3. In such cases, as for the described Newton-Raphson algorithms, there is no need to store and invert the Hessian matrix.

4.1. Damping parameter and stopping criteria

The damping parameter γ influences the step size in the LM iterations. A very small value of γ leads to an algorithm that is very close to the NR algorithm. This behaviour is good when we are close to the maximum because we are close to quadratic convergence. However, larger values of γ can be good if the actual value is far from the optimum. In this case, a larger γ produces shorter steps in the iterations. Nielsen (1998) proposed to use values very close to zero or related to the diagonal element of the matrix, approximating the Hessian in the Levenberg-Marquardt original algorithm. Similarly, we propose to use $1/K$ as a value close to zero, or 1 which is the diagonal element of the Hessian when rescaled by its diagonal elements, $[\text{diag}(\mathbf{D})]^{-1}\mathbf{D}$ (this form is similar to that of the Levenberg-Marquardt algorithm). Contrary to the Levenberg-Marquardt algorithm our damping parameter is not adaptive. However, we prove in the appendix a convergence property of our algorithm and we show its performance in simulations and on real data.

We stop the algorithm as soon as one of these three criteria is satisfied: if the norm of the gradient is very close to zero: $\| \nabla f \| < \epsilon_1$; if the relative changes of the parameters are very small: $\| \mathbf{x}_{\text{new}} - \mathbf{x} \| < \epsilon_2(\| \mathbf{x} \| + \epsilon_2)$; if the number of iterations is greater than a pre-established threshold.

5. Simulated data

To compare the different proposals in an high-dimensional setting we have implemented a simulation with 1000 variables and 20 units. With a huge number of parameters there is an high chance that an element of a simulated unity is so close to zero to be considered zero due to machine precision. This problem almost disappears when all the parameters have values far from zero. To investigate the consequences of such choices we consider a range of values for the sum of the parameters $\sum \alpha_k$ from 10000 to 50000 with step size of 2000. Each parameter is drawn from a uniform distribution between $\sum \alpha_k/K - 2$ and $\sum \alpha_k/K + 2$ where $K = 1000$. Let us remark that the final sum of the simulated parameters is not necessarily equal to the pre-established values in the sequence.

We consider that a method has reached convergence when it is stopped before the number of iterations reaches 1000 and the estimated vector is in the correct range. The tolerance parameters ϵ_1 and ϵ_2 are both set to 10^{-8} and for the damping parameter γ we use the values considered in the previous section. The number of simulations is 2000.

The results are reported in Figure 1. In the upper panel we report the convergence rate for each combination of starting values/methods. In the lower panel the mean number of iterations required for convergence is shown. FPI and LM methods with $\gamma = 1$ have a similar performance, reaching convergence very often and for every starting value. L-BFGS-B shows a good range of convergence when coupled with Wicker or the method of moments but not with the other two initialisation methods. The starting values of

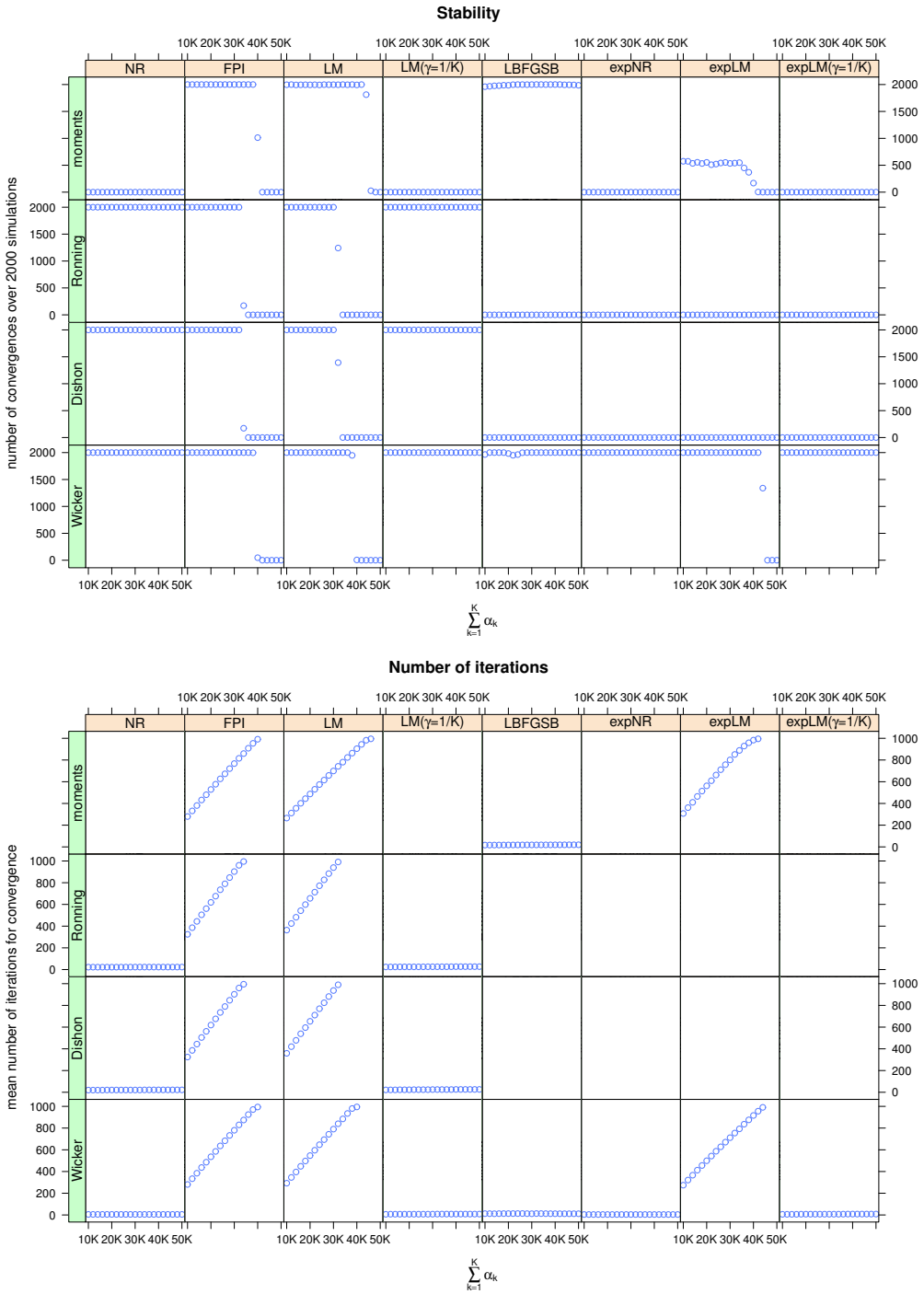

Figure 1: *Results from the simulation study. In the upper panel we show how many times we reached convergence for each combination of starting value and algorithm. Below we show instead the number of iterations used to reach convergence in the upper panel.* $\sum_{k=1}^{K} \alpha_k$ *indicates approximatively the sum of the parameters to be estimated.*

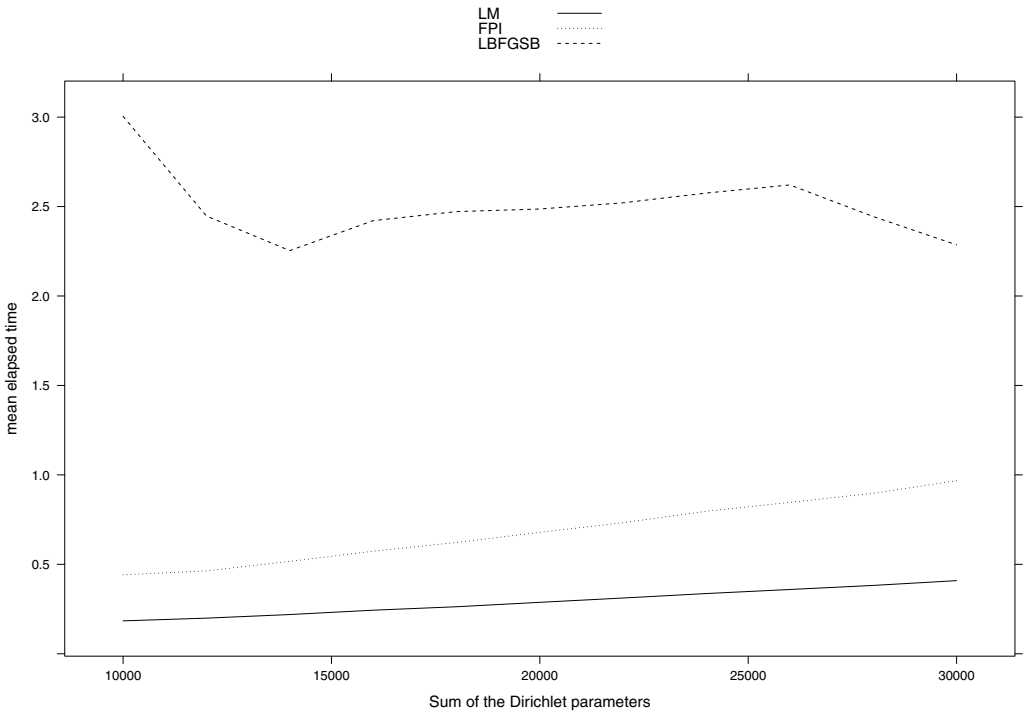

Figure 2: *Time comparison of the algorithms LM, FPI and L-BFGS-B over 100 simulations. Time is expressed in seconds.*

Wicker et al. (2008) are the only ones able to guarantee convergence for all the methods. The algorithms using the re-parametrization show poor performance, probably due to the possible lack of concavity and the fact that only the starting values of Wicker et al. (2008) seem to be often in a neighbourhood of the maximum. However, the price to pay for the highest stability is the high number of iterations required to reach convergence for both FPI and LM with $\gamma = 1$. NR was instead the fastest method. As expected, see Section 4.1, LM with $\gamma = 1/K$ and expLM with with $\gamma = 1/K$ have a performance close to those of NR and expNR, respectively.

The efficiency of the algorithms LM, FPI and L-BFGS-B cannot be compared only looking for differences in the number of iterations because their corresponding iteration times can be totally different. For these algorithms we therefore implemented also a comparison on the total time required to reach convergence. The settings for this simulation were similar to the ones used above with a range for the sum of the parameters going from 10000 to 30000 with step size of 2000 and a number of simulations equal to 2000. The results are reported in Figure 2. On average LM requires only half of the time employed by FPI. L-BFGS-B is clearly much slower than the other two competitors.

6. Apple data set

We have seen in the introduction that the Dirichlet distribution has strong independence properties that are often unrealistic for real data. However, for high-dimensional data once we focus on a single variable we expect that this is correlated only with a limited number of the other variables and uncorrelated with the rest. In this case it can be of interest to see the fit of a simple model as the Dirichlet distribution, that can be thought of as a raw approximation of the reality. We are going to consider a spike-in experiment that has been already investigated in the literature with the appropriate statistical tools. If we make the corresponding data compositional and we investigate them through the Dirichlet distribution we can judge if there is a discrepancy or an agreement between what is already known and what we can learn with the Dirichlet distribution. Since a spike-in experiment is a controlled experiment where we know the thruth, this comparison can indirectly tell us if the Dirichlet distribution can be used for high-dimensional data. This is not intended to be a proof to say that the Dirichlet distribution can safely be used for high-dimensional compositional data. For a stronger result, comparisons must be done with other distributions. Here we focus on the computational performance of different algorithms to get the maximum likelihood estimates and we offer a brief look at the possibilities of using the Dirichlet distribution to analyse high-dimensional data.

Specifically we apply the previous algorithms to a data set from the field of metabolomics, available in the R package BioMark (Wehrens and Franceschi, 2012). The data set consists of mass spectrometric measurements on apples and is fully described in Franceschi et al. (2012). We consider the positive ionization data for the 10 control samples, and the first group of 10 spiked-in samples. There are 1632 variables in total. We delete the variables with missing values and normalize the data to give 1 as the sum of the elements of each unit. This corresponds to have a total intensity for each unit that is redistributed through the variables. The final data set has 1602 variables. The results are summarised in Table 1.

For these data, the initialisation of Wicker and co-workers is able to give convergence in the correct range of the parameters for five methods. In two cases (LM with $\gamma = 1/K$, NR) the result is outside the correct range. The other initialisations fail for expLM ($\gamma = 1$), expLM ($\gamma = 1/K$) and expNR. The initialisation based on the method of moments fails also for LM ($\gamma = 1$). L-BFGS-B is not able to reach convergence with any initialisation method. The only method that is always able to reach convergence in the correct range is FPI. LM ($\gamma = 1$) and LM ($\gamma = 1/K$) reach convergence in three out of four cases, while NR converges in two out of four cases. Using the exponential parametrization the other methods reach convergence only with the Wicker initialisation, but in these cases very few iterations are needed.

Table 1: *Number of iterations required for convergence in the Apple data set. We report with a bar the cases where convergence is not reached or the result is outside the correct range for the parameters.*

	NR	FPI	LM	LM($\gamma = \frac{1}{K}$)	L-BFGS-B	expNR	expLM	expLM($\gamma = \frac{1}{K}$)
moments	—	543	—	—	—	—	—	—
Ronning	30	495	543	32	—	—	—	—
Dishon	21	494	529	23	—	—	—	—
Wicker	—	434	447	—	—	7	411	8

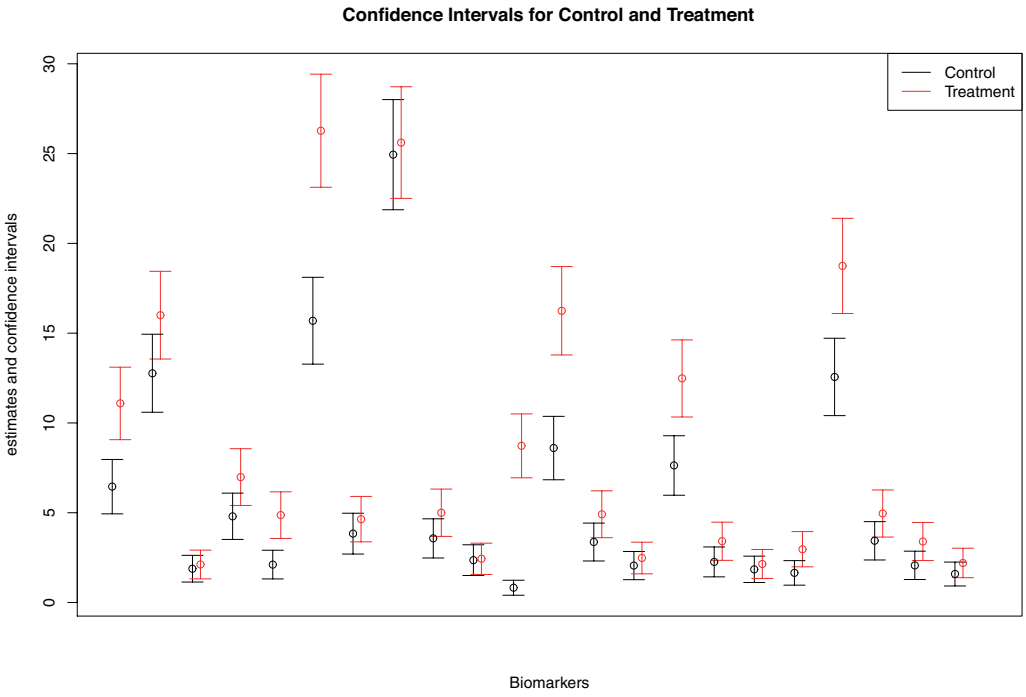

Figure 3: *Each pair in the figure represents the confidence intervals for control and treatment respectively, for 22 biomarkers.*

While FPI is very stable, it also requires a large number of iterations to reach convergence with each initialisation (mean for the four initialisations $= 491.5$). The same is true for LM ($\gamma = 1$) where the mean for three initialisations that reach convergence is 506.3 and for the initialisation with the method of moments we do not have convergence because we reach the maximum number of iteration allowed (1000 in our setting). However FPI can require a longer time than LM because FPI requires for each iteration a second inner iterative algorithm. For example, comparing FPI and LM ($\gamma = 1$) using the Wicker initialisation we have an elapsed time of 2.259 and 0.678 seconds respectively.

With Wicker initialisation expLM ($\gamma = 1$) requires 411 iterations to reach convergence while expLM ($\gamma = 1/K$) requires only 8 steps; similarly expNR requires only

7 steps. LM ($\gamma = 1/K$) and NR also required a low number of iterations when they reached convergence.

Fitting different Dirichlet distributions to control and treatment data it is possible to compare them. We report the confidence intervals for 22 biomarkers in Figure 3. These biomarkers are known to correspond to spike-in compounds (see Franceschi et al., 2012). Since these confidence intervals are not simultaneous we cannot use them for identifying 'directly' statistically significant biomarkers. We use them for ranking these biomarkers, visualizing the order of magnitude of the differences between control and treatment. There are several cases where the differences are clear. This is in agreement with the previous findings in Franceschi et al. (2012).

7. Conclusions

In this paper we have compared the computational performance of eight different algorithms and four different starting value strategies to estimate the Dirichlet distribution through maximum likelihood. Such a comparison provides indications about the methods to use in order to analyse high-dimensional compositional data with the Dirichlet distribution.

The Newton-Raphson algorithm is very fast, but can lead to estimated parameters outside the allowed region. On the other hand, the FPI algorithm has a slow convergence but is very stable. The other algorithms have a performance between these two extremes.

To have parameters always in the correct range we considered a re-parametrization and a box-constraints algorithm, L-BFGS-B. The re-parametrization allows us to have parameters always in the correct range but possibly loses the characteristic of being a concave function. This means that good convergence is assured only in a neighbourhood of the maximum and that convergence cannot be guaranteed. In practice from our study we can see that the re-parametrization is useful only if coupled with the initialisation method of Wicker et al. (2008). L-BFGS-B is more stable than the re-parametrization but less than FPI and moreover its iteration time is huge.

The proposed modifications to the Levenberg-Marquardt algorithm consider a prudent step compared to the Newton-Raphson algorithm and therefore can offer a good trade-off between speed and stability. Newton-Raphson and the proposed algorithms have local convergence characteristics and therefore the starting values are very important even if the function to be optimized is concave. These features are particularly relevant in a high-dimensional setting where the number of parameters largely exceed the number of units. From the simulations and the real study only the Wicker et al. (2008) approach seems able to provide convergence for high-dimensional data.

Considering both simulations and the real data example the combination of the Levenberg-Marquardt methods or fixed point iteration method with the starting values of Wicker appear to be the most promising. However, the Levenberg-Marquardt methods leave room for improvements. In this paper we have been able to prove convergence

properties for a fixed damping parameter γ. If this parameter can be made adaptive, as in the original Levenberg-Marquardt algorithm, it is not unreasonable to expect a higher stability and a lower mean number of iterations for convergence.

Appendix

In what follows the notation refers to the quantities previously introduced in the paper.

Lemma 1 *At the point of maximum, $\hat{\mathbf{x}}$, the differential of the iteration map (26) is given by*

$$d\mathbf{M}(\hat{\mathbf{x}}) = \mathbf{I} - \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \mathrm{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1}\mathbf{H}(\hat{\mathbf{x}}).$$

Proof. We treat γ as a positive constant, but the proof holds even if γ is a positive function that does not depend on $\mathbf{x}$.

Let us denote by $\mathbf{G}(\mathbf{x})$ the matrix $\{\mathbf{H}(\mathbf{x}) + \gamma \mathrm{diag}\mathbf{H}(\mathbf{x})\}^{-1}$. We know from Section 4 that $\mathbf{G}(\mathbf{x})$ is well defined for every $\mathbf{x}$ in the original parametrization and at least in a neighbourhood of the maximum for the re-parametrization. For such cases the elements of $\mathbf{G}(\mathbf{x})\nabla f(\mathbf{x})$ can be written as $\sum_j g_{ij}(\mathbf{x})\nabla f_j(\mathbf{x})$. To prove the lemma we have to show that the partial derivatives of this expression are well defined. Using the product rule it is patent that the difficult part is to prove that the partial derivatives of the elements $g_{ij}(\mathbf{x})$ are well defined. If we are able to prove that the l_{ij} and the elements of the diagonal matrix $\mathbf{D}$ are derivable it follows that also the g_{ij} are derivable and therefore the lemma easily follows. Using standard rules for derivation we see that these terms are derivable if the trigamma function is derivable. For positive reals the trigamma function can be expressed as a positive series dominated by $\sum n^{-2}$. Therefore by the Weierstrass M-test there is uniform convergence and the trigamma function is derivable. Moreover, the series form assures that the trigamma function is strictly decreasing for positive reals.

We can summarize the results in matrix form. At the point of maximum $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}$ and therefore we get $d\mathbf{M}(\hat{\mathbf{x}}) = \mathbf{I} - \mathbf{G}(\hat{\mathbf{x}})\mathbf{H}(\hat{\mathbf{x}}) = \mathbf{I} - \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \mathrm{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1}\mathbf{H}(\hat{\mathbf{x}})$. ■

Theorem 1 *The proposed Levenberg-Marquardt algorithms based upon equations (24) and (25) are locally attracted to the maximum $\hat{\boldsymbol{x}}$ at a linear rate equal to the spectral radius of*

$$\mathbf{I} - \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \mathrm{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1}\mathbf{H}(\hat{\mathbf{x}})$$

or at a better rate.

Proof. The point of maximum for $f(\mathbf{x})$ is a fixed point for $\mathbf{M}(\mathbf{x})$. According to Proposition 15.3.1 in Lange (2010), it suffices to show that all eigenvalues of the differential

$d\mathbf{M}(\hat{\mathbf{x}})$ lie on the open interval $(0,1)$. By Lemma 1 the following equalities hold:

$$\begin{aligned} d\mathbf{M}(\hat{\mathbf{x}}) &= \mathbf{I} - \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1} \mathbf{H}(\hat{\mathbf{x}}) \\ &= \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1} \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}}) - \mathbf{H}(\hat{\mathbf{x}})\} \\ &= \{\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})\}^{-1} \{\gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})\}. \end{aligned}$$

The maximum and minimum eigenvalues of $d\mathbf{M}(\hat{\mathbf{x}})$ are determined by the maximum and minimum values of the Rayleigh quotient ($\mathbf{v} \neq \mathbf{0}$):

$$\begin{aligned} R(\mathbf{v}) &= \frac{\mathbf{v}^\mathsf{T} [\gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})] \mathbf{v}}{\mathbf{v}^\mathsf{T} [\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})] \mathbf{v}} \\ &= 1 - \frac{\mathbf{v}^\mathsf{T} \mathbf{H}(\hat{\mathbf{x}}) \mathbf{v}}{\mathbf{v}^\mathsf{T} [\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})] \mathbf{v}}. \end{aligned}$$

If the quantities $\mathbf{H}(\hat{\mathbf{x}})$ and $\gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})$ are definite negative then also $\mathbf{H}(\hat{\mathbf{x}}) + \gamma \text{diag}\mathbf{H}(\hat{\mathbf{x}})$ is definite negative and it follows that $0 < R(\mathbf{v}) < 1$.

For the original parametrization we can show that $\gamma \text{diag}\mathbf{H}(\mathbf{x})$ is always definite negative. We use the parametric form of $\mathbf{H}$ for the Dirichlet distribution. We have $\mathbf{v}^\mathsf{T} \gamma \text{diag}\mathbf{H}(\mathbf{x}) \mathbf{v} = \gamma \sum v_i^2 h_{ii}$ where $h_{ii} = N(\Psi'(\sum \alpha_j) - \Psi'(\alpha_i))$.We have seen that for positive reals the trigamma function is strictly decreasing and therefore $h_{ii} < 0$. Therefore $\mathbf{v}^\mathsf{T} \gamma \text{diag}\mathbf{H}(\mathbf{x}) \mathbf{v} < 0$. Being $f(\cdot)$ concave $\mathbf{H}$ is semi-definite negative, but $\mathbf{H}$ is also invertible and therefore is definite negative.

For the re-parametrization we cannot assure that the Hessian matrix is negative definite for every $\mathbf{x}$. However, to apply Proposition 15.3.1 in Lange (2010) we need only to prove that this is true at $\hat{\mathbf{x}}$. For the diagonal matrix we observe that:

$$[\lambda \text{diag}\mathbf{H}(\hat{\mathbf{x}})]_{kk} = \lambda q_{kk} + \lambda \exp(2\beta_k) N \Psi' \left(\sum_k \exp(\beta_k) \right) \tag{37}$$

$$= \lambda N [\nabla f(\boldsymbol{\beta})]_k \tag{38}$$

$$+ \lambda \exp(2\beta_k) N \left[\Psi'(\sum_k \exp(\beta_k)) - \Psi'(\exp(\beta_k)) \right]. \tag{39}$$

By the properties of the trigamma function $\Psi'(\sum_k \exp(\beta_k)) - \Psi'(\exp(\beta_k)) < 0$ and therefore at $\hat{\mathbf{x}}$ the matrix $\lambda \text{diag}\mathbf{H}$ is negative definite. Moreover at $\hat{\mathbf{x}}$ the Hessian of the re-parametrization is semi-definite negative and invertible and therefore is definite negative. ■

Acknowledgement

The authors thank Federico Vaggi for useful comments on an earlier version of the manuscript.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on statistics and applied probability. Chapman and Hall.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16, 1190–1208.

Connor, R. J. and Mosiman, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194–206.

Dishon, M. and Weiss, G. (1980). Small sample comparison of estimation methods for the beta distribution. *Journal of Statistics Computation and Simulation*, 11, 1–11.

Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F. and Wehrens, R. (2012). A benchmark spike-in data set for biomarker selection in metabolomics. *Journal of Chemometrics*, 26, 16–24.

Huang, J. (2005). Maximum likelihood estimation of Dirichlet distribution parameters. http://www.stanford.edu/ jhuang11/research/dirichlet/, Robotics Institute, Carnegie Mellon University.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer, second edition.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2, 164–168.

Madsen, K., Nielsen, H. B. and Tingleff, O. (2004). *Methods for Non-Linear Least Squares Problems* (2nd ed.). Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11, 431–441.

Minka, T. P. (2000). Estimating a Dirichlet distribution. http://research.microsoft.com/ en-us/um/people/minka/papers/, M.I.T.

Narayanan, A. (1991a). Algorithm AS266: maximum likelihood estimation of parameters of the Dirichlet distribution. *Applied Statistics*, 40, 365–374.

Narayanan, A. (1991b). Small sample properties of parameter estimation in Dirichlet distribution. *Communications in Statistics Simulations and Computation*, 20, 647–666.

Nielsen, H. B. (1998). *Damping Parameter in Marquardt's Method*. Technical Report IMM-REP-1999-05. Department of Mathematical Modelling, DTU.

Rayens, W. S. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the Simplex. *Journal of the American Statistical Association*, 89, 1465–1470.

Ronning, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistics Computation and Simulation*, 32, 215–221.

Wehrens, R. and Franceschi, P. (2012). Meta-statistics for variable selection: the R package BioMark. *Journal of Statistical Software*, 51, 1–18.

Wicker, N., Muller, J., Kalathur, R. K. R. and Poch, O. (2008). A maximum likelihood approximation method for Dirichlet's parameter estimation. *Computational Statistics and Data Analysis*, 52, 1315–1322.