Deconvolution of dust mixtures by latent Dirichlet allocation in forensic science

M. A. Ausdemore and C. Neumann

Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA

E-mail: Madeline.Ausdemore@gmail.com

Summary. Dust particles recovered from the soles of shoes may be indicative of the sites recently visited by an individual, and, in particular, of the presence of an individual at a particular site of interest, e.g., the scene of a crime. By describing the dust profile of a given site by a multinomial distribution over a fixed number of dust particle types, we can define the probability distribution of the mixture of dust recovered from the sole of a shoe via Latent Dirichlet Allocation. We use Variational Bayesian Inference to study the parameters of the model, and use their resulting posterior distributions to make inference on (a) the contributions of sites of interest to a dust mixture, and (b) the particle profiles associated with these sites.

Keywords: Latent Dirichlet Allocation; Posterior distribution; Topic models; Variational Bayesian Inference
1. Introduction

Dust particles recovered from beneath the soles of an individual’s shoes consist of a mixture of dust particles collected from different sources and may be indicative of the locations recently visited by that individual. In particular, this dust may reveal his presence at a location of interest, e.g., the scene of a crime. The contributions of these locations to the mixture may vary as a function of the amount of dust present at each location, the time spent by the individual at each location, the activity of the individual at each location, or how recently the individual visited each location. The profile of a given source of dust can be described by a multinomial distribution over a fixed number of dust particle types, which enables us to describe the mixture of dust recovered from the sole of a shoe by latent Dirichlet allocation.

In this paper, we describe an algorithm that resolves mixtures of dust to address two different questions of forensic interest. Given a set of samples recovered from one or more objects of forensic interest that consists in mixtures of dust from $M = Q + K$ sources, and samples of dust from $K$ known sources, we are interested in:

(a) inferring the dust profiles of the $Q$ unknown sources;
(b) inferring the proportions in which dust from each of the $Q + K$ sources are present in the samples.

An example of the first inference question may arise when, given an individual suspected of kidnapping with known home and workplace, we are interested in providing information on the dust composition of the unknown location where the victim is being held. Provided that the necessary data and a suitable inference framework exist, the second inference question may be useful to discuss issues such as (a) how long a person stayed at each location, (b) how recently the person visited each location, or (c) what type of activity the person had at each location.

We use latent Dirichlet allocation (LDA) (Blei et al. (2003)) to define the generative process that produces mixtures of dust samples. We use variational Bayesian inference (VBI) (Hoffman et al. 2013; Blei et al. 2016) to study the parameters of our model and to address these two inference questions.

Currently, the use of dust evidence is anecdotal and is limited to cases where rare and characteristic particles are observed (e.g., pollen, seeds, alloy traces). This model enables the widespread use of dust evidence and may trigger a paradigm change in the way forensic science contributes to criminal investigations. Indeed, most evidence types currently considered by forensic scientists result from the activity of criminals at crime scenes or their interactions with victims, while dust evidence arises from the mere presence of individuals at locations of interest.

2. Applying LDA to mixtures of dust

Topic models aim at discovering hidden semantic structures in a body of documents by grouping together words that are likely to have originated from the same themes or
Deconvolution of dust mixtures by latent Dirichlet allocation

authors. By generalising this concept, LDA can be extended to arbitrary mixtures of objects represented by categorical random vectors, such as particle types.

Dust sources generate particles that can be observed at different locations, such as a crime scene or the house of a suspect, or one or more objects of forensic interest, such as shoe soles. The model presented in this paper uses dust samples collected from relevant geographical locations and from the surfaces of objects of forensic interest to make inferences on the dust profiles of the dust sources, and on their respective contributions to the dust mixture observed on the objects.

In order to address the inference questions presented above, our model makes a distinction between the location where dust can be sampled, and the source from which the dust originates.† Our terminology also differentiates a geographical location, which corresponds to any place that an individual might have visited and where dust might be sampled from, from a trace location, which corresponds to a location or object where evidentiary dust samples are collected. Finally, our model considers that all the dust at a given geographical location originates exclusively from a single source, while the dust observed on the surfaces of trace objects originates potentially from more than one source (see assumptions (d) and (e) below). This constraint allows for learning the dust profiles of the different geographical locations that might have been visited by an individual and provides a basis to determine if dust from any of these locations is present in the mixture observed on the trace object.

Our problem is different than the one addressed by supervised topic models (Blei and McAuliffe, 2007; Lacoste-Julien et al., 2008; Wang et al., 2009; Zhu and Ahmed, 2012). We are not interested in associating a new set of dust samples with a single specific source, but are instead interested in determining the relative contributions of the $Q+K$ sources to the mixtures of dust observed in a sample.

Our problem also differs from that addressed by author-topic models (Rosen-Zvi et al., 2004; Steyvers et al., 2004; Rosen-Zvi et al., 2010). Although we can draw a parallel between authors and sampling locations, our inference questions diverge. The author-topic model proposed by Rosen-Zvi et al. (2004) aims at discussing which topics are preferred by each author in a set of authors by assuming a uniform contribution of each author to a single document; however, by construction, this model does not enable the study of the respective contribution of each author to a set of documents, which would be a comparable inference question to the one addressed in this paper.

Overall, our problem is similar to the one originally addressed by Blei et al. (2003). Nevertheless, while Blei et al. (2003) considered that a corpus consists in multiple documents composed of the same topics in the same proportions, we consider that a corpus consists in multiple documents composed of the same topics, but in varying proportions. Thus, we consider that the parameters that control the contributions of the different dust sources to the dust samples and the parameters that control the particle profiles of these

†Locations and sources correspond respectively to authors and topics in author-topic models, since authors generate documents, and documents are made out of topics. To help the reader with our development, we draw a parallel between the vocabulary used in this paper and the more familiar vocabulary used in topic modelling. This glossary of terms can be found in Appendix A.
sources are distributed according to asymmetric Dirichlet distributions. In our implementation, these hyper-parameters are represented by matrices rather than vectors. To help with identifiability, we use “pure” dust samples from known dust source to support the modelling of their dust profiles.

To develop our model, we make the following assumptions:

(a) Observations made on particles within a dust sample are exchangeable (Robert (2007), page 159).

(b) Observations made on dust samples collected at a given location are exchangeable.

(c) Sources yield dust with a fixed and constant profile. Dust sources do not cross-contaminate each other.

(d) The composition of dust samples recovered at a geographical location, such as a crime scene, workplace, or home, is considered to originate from a single source (i.e., the geographical location itself).

(e) The composition of dust samples recovered on trace objects may be influenced by more than one source.

Assumptions (a) and (b) are identical to the assumptions made to develop the original model proposed by Blei et al. (2003). Assumption (a) is reasonable since particles are not organised in any particular order in a dust sample. In practice, an appropriate sampling procedure will ensure that the second assumption holds. Assumption (c) considers that the dust profile of any given dust source is characterised by the “dust output” of that source, and thus accounts for any prior cross-contamination between sites. This assumption may not be appropriate and may be investigated in future work. Finally, assumptions (d) and (e) are critical for the inferences we want to make with this model: they allow us to make inferences on the origin of the dust recovered from trace objects of forensic interest in terms of geographical sampling locations from mixtures of dust sources.

We describe the generative process of a dust sample as follows. Notation is summarised in tables 4 and 5 in Appendix B. The top part of figure 1 provides a graphical representation of this process.

(a) Choose an $M \times T$ matrix $H$ to represent the relative contributions of the different particle types to the dust profiles of each of the $M = Q + K$ sources that have the potential to contribute to the mixture. The $m$-th row of $H$ corresponds to the parameters of a $T$–Dirichlet distribution that drives the mixing proportions of the $T$ particle types that characterise source $m$.

(b) Choose an $L \times M$ matrix $A$ to represent the relative contributions of the $M$ different sources to each of the $L$ locations from which we obtain dust samples. The $l$-th row, $\alpha_l$, of $A$ corresponds to the parameters of an $M$–Dirichlet distribution that drives the mixing proportions of the $M$ sources in samples obtained from location $l$.

(c) For a set of dust samples obtained from known locations and on evidentiary objects, sample an $M \times T$ matrix $B$ from $H$ to obtain the mixing proportions of the types of
Fig. 1. Generative process for a sample of dust particles (top part) and update process for the Variational Bayesian Inference algorithm (bottom part).

dust particles for each source of dust \( m \in \{1, \ldots, M\} \).

(d) For a sample \( x_{ls} \) taken from location \( l \):

(i) Sample \( \theta_{ls} \sim \text{Dirichlet}(\alpha_l) \) to obtain a vector of mixing proportions of the dust sources for the samples obtained from location \( l \).

(ii) For each of the \( N_{ls} \) particles, \( x_{lsn} \), in sample \( x_{ls} \):

(a) Sample a source of dust \( z_{lsn} \sim \text{Multinomial}(1, \theta_{ls}) \).

(b) Sample a particle \( x_{lsn} \sim \text{Multinomial}(1, \beta_{z_{lsn}}) \), where \( \beta_{z_{lsn}} \) represents the row of matrix \( B \) for the source defined by \( z_{lsn} \).

We see that the model makes no assumptions pertaining to any sort of ordering or grouping of the particles or the locations in a dust sample. This is synonymous to the bag of words assumption (i.e., exchangeability [Robert (2007), page 159] that is commonly associated with topic modelling.

This process can be represented by means of the Directed Acyclic Graph (DAG) depicted in figure 1. By making use of the DAG and the generative process described above, the probability of a sample of dust particles is given by:

\[
p(x_{ls}|\alpha_l, H) = \int \int p(\theta_{ls}|\alpha_l) p(B|H) \prod_{n=1}^{N} \sum_{z_{lsn}} p(z_{lsn}|\theta_{ls}) p(x_{lsn}|z_{lsn}, B) \, d\theta_{ls} \, dB.
\]  

The joint probability of a set of samples collected across multiple locations is then given
by:
\[ p(X|A, H) = \prod_{l=1}^{L} \prod_{s=1}^{S_l} \left[ \int \int p(\theta_{ls}|\alpha_l) p(B|H) \prod_{n=1}^{N} \sum_{z_{lns}} p(z_{lns}|\theta_{ls}) p(x_{lns}|z_{lns}, B) d\theta_{ls}dB \right]. \] (2)

The distributions represented by each node in the top part of figure [1] are given by:

\[ p(\theta_{ls}|\alpha_l) = \frac{\Gamma \left( \sum_{j=1}^{M} \alpha_{lj} \right)}{\prod_{m=1}^{M} \Gamma (\alpha_{lm})} \prod_{m=1}^{M} \theta_{lm}^{\alpha_{lm} - 1} \] (3)

\[ p(B|H) = \prod_{m=1}^{M} \Gamma \left( \sum_{t=1}^{T} \eta_{mt} \right) \prod_{t=1}^{T} \eta_{mt}^{\gamma_{mt} - 1} \] (4)

\[ p(a_{sn}|\theta_{ls}) = \prod_{m=1}^{M} \theta_{lm}^{\alpha_{lnm} - 1} \] (5)

\[ p(x_{lns}|z_{lns}, B) = \prod_{m=1}^{M} \prod_{t=1}^{T} \beta_{lm}^{T_{lns} z_{lm} - 1} \] (6)

3. Assigning the model parameters

We use VBI to assign the approximate posterior distribution \( p(\Theta, B|X) \) of a set of exchangeable observations \( X \) obtained from \( L \) geographical locations and trace objects. This is achieved by maximising the lower bound function defined by the negative Kullback-Leibler divergence of the joint distribution \( p(X, \Theta, B, Z) \) and the variational distribution \( q(\Theta, B, Z) \) (Bishop (2006), Chapter 10). We introduce the variational parameters, \( \Gamma, \Lambda, \) and \( \Phi \) to break the dependencies that exist between \( \Theta, B, \) and \( Z \) (see figure [1]), and we define \( q(\Theta, B, Z) \) as:

\[ q(\Theta, B, Z|\Gamma, \Lambda, \Phi) = \prod_{l=1}^{L} \prod_{s=1}^{S_l} \left[ q_{ls}(\theta_{ls}|\gamma_{ls}) \prod_{n=1}^{N} q_{ls}(x_{lns}|\phi_{ls}) \prod_{m=1}^{M} q_{ms}(\beta_{lm}|\lambda_{lm}) \right]. \] (7)

The “E-Step” of our implementation of VBI maximises the lower bound function with respect to each of the variational parameters, \( \Gamma, \Lambda, \) and \( \Phi \), while maintaining fixed values of \( \Theta, B, \) and \( Z \); the “M-Step” maximises the lower bound function with respect to the global latent parameters \( H \) and \( A \), while keeping the variational parameters obtained in the E-step fixed. Each step is itself an iterative process that repeats until some convergence criteria is satisfied. From equation (7), we note that \( \gamma_{ls}, \lambda_{ls}, \) and \( \phi_{ls} \) can be updated independently to minimise the KL divergence between \( q(\Theta, B, Z|\Gamma, \Lambda, \Phi) \) and \( p(X, \Theta, B, Z) \) for each sample. For ease of notation, we now suppress the explicit conditioning on the variational parameters, and shorten \( q(\Theta, B, Z|\Gamma, \Lambda, \Phi) \) to \( q(\Theta, B, Z) \).

3.1. The lower bound function for mixtures of dust particles

The lower bound function mentioned above is the sum of expectations of each of the latent parameters, taken with respect to the variational distribution, \( q(\Theta, B, Z) \), as shown in
The first four expectations are developed using equations (3) - (6) above:

\[ \mathcal{L}(\Theta, \Theta, B, Z) = \int q(\Theta, B, Z) \log p(X, \Theta, B, Z) d\Omega - \int q(\Theta, B, Z) \log q(\Theta, B, Z) d\Omega \]

\[ = E_q[\log p(X | Z, B)] + E_q[\log p(\Theta | A)] + E_q[\log p(Z | \Theta)] + E_q[\log p(B | H)] - E_q[\log q(\Theta)] - E_q[\log q(Z)] - E_q[\log q(B)] \]  \tag{8}

The first four expectations are developed using equations (3) - (6) above:

\[ E_q[\log p(X | Z, B)] = \sum_{l=1}^{L} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{m=1}^{M} x_{lsm} \phi_{lsm} \left( \Psi(\lambda_{lsm}) - \Psi\left( \sum_{j=1}^{M} \lambda_{lsmj} \right) \right) \]

\[ E_q[\log p(\Theta | A)] = \sum_{l=1}^{L} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{m=1}^{M} \phi_{lsm} \left( \Psi(\gamma_{lsm}) - \sum_{j=1}^{M} \gamma_{lsmj} \right) \]

\[ E_q[\log p(Z | \Theta)] = \sum_{l=1}^{L} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[ \log \Gamma \left( \sum_{j=1}^{M} \alpha_{lj} \right) - \sum_{m=1}^{M} \log \Gamma (\alpha_{lm}) + \sum_{m=1}^{M} (\alpha_{lm} - 1) \left( \Psi(\gamma_{lsm}) - \Psi\left( \sum_{j=1}^{M} \gamma_{lsmj} \right) \right) \right] \]

\[ E_q[\log p(B | H)] = \sum_{l=1}^{L} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[ \log \Gamma \left( \sum_{j=1}^{M} \beta_{lj} \right) - \sum_{m=1}^{M} \log \Gamma (\beta_{lm}) + \sum_{m=1}^{M} (\beta_{lm} - 1) \left( \Psi(\lambda_{lsm}) - \Psi\left( \sum_{j=1}^{M} \lambda_{lsmj} \right) \right) \right] \]

The last three expectations are developed using the entropies of the distributions corresponding to each of the latent parameters, given by equations (3) - (5):

\[ -E_q[\log q(\Theta)] = \sum_{l=1}^{L} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{m=1}^{M} \phi_{lsm} \log \phi_{lsm} \]

By maximising equation (8) with respect to each of the variational parameters, we obtain update equations for \( \phi_{lsm}, \gamma_{lsm}, \) and \( \lambda_{lsm} \):

\[ \phi_{lsm} \propto \exp \left\{ \sum_{t=1}^{T} x_{lsm} \left( \Psi(\lambda_{lsm}) - \Psi\left( \sum_{j=1}^{M} \lambda_{lsmj} \right) \right) + \Psi(\gamma_{lsm}) \right\} \]  \tag{9}

\[ \gamma_{lsm} = \alpha_{lm} + \sum_{n=1}^{N} \phi_{lsmn} \]  \tag{10}

\[ \lambda_{lsm} = \eta_{lm} + \sum_{n=1}^{N} x_{ln}\phi_{lsmn} \]  \tag{11}

These updates for the variational parameters are used in the E-Step of the VBI algorithm described in algorithm [1]. Assigning values to the global parameters \( H \) and \( A \) in the M-Step requires using optimisation techniques, since tractable maximum likelihood solutions do not exist. We use the L-BFGS-B method [Byrd et al., 1994; Zhu et al., 1994] to obtain the matrices \( H \) and \( A \). For more information pertaining to the L-BFGS-B method and the M-Step, see Appendix C.
Algorithm 1: E-Step - Updating the variational parameters

for each sample do
    Initialise $\Gamma := A + \left( \sum_{n=1}^{N} x_{lsmn} \right) / \left( M \star \sum_{l=1}^{L} S_{l} \right)$;
    Initialise $\Phi_{ls} := 1 / M$;
    Initialise $\Lambda_{ls} := H$;
    while $\mathcal{L}(q(\Theta, B, Z))$ has not converged do
        $\phi_{lsmn} \propto \exp \left\{ \sum_{t=1}^{T} x_{lsmn} \left( \Psi (\lambda_{lstm}) - \Psi \left( \sum_{j=1}^{T} \lambda_{lsmj} \right) \right) + \Psi (\gamma_{lsm}) \right\}$;
        $\gamma_{lsm} = N^{-1} \left( \alpha_{lm} + \sum_{n=1}^{N} \phi_{lsmn} \right)$;
        $\lambda_{lstm} = N^{-1} \left( \eta_{ml} + \sum_{n=1}^{N} x_{lst} \phi_{lsmn} \right)$;
    end
end

3.2. Initialisation of the model

By assumption (d), our model considers that the dust observed at any given geographical location originates from a single source. Hence, the strategy behind the proposed model is to learn the dust profile of the $K$ known sources by obtaining “pure” samples from the corresponding locations, and to infer the profile of the remaining $Q$ unknown sources (corresponding to $Q$ locations that cannot be studied) through the deconvolution of the samples recovered on the surface of the trace objects. This process also provides information on the respective contributions of the $M = Q + K$ sources to the mixtures observed on the trace objects.

To ensure that each geographical location (known and unknown) is uniquely associated with a single source, we constrain, $A$, the $L \times M$ matrix of Dirichlet parameters controlling the mixing proportions of dust sources at each location. Each column of $A$ corresponds to one of the $M$ sources that may have contributed dust to the different locations, and $K$ of the $L$ rows of $A$ correspond to known locations where “pure” samples were collected. Constraining $A$ requires to heavily weigh the $m$-th element of the $l$-th row of $A$, when $m = l$ and $m, l \leq K$. For rows of $A$ corresponding to the evidentiary samples, we use a flat $M$–Dirichlet distribution to reflect that, before running the algorithm, we consider that all dust sources are equally likely to have contributed to these samples. The values of the rows associated with the $K$ known locations are not updated by the algorithm, while the rows associated with the $Q$ unknown locations and the evidentiary samples are updated during the “M-step” of the algorithm.

Matrix $H$ is initialised with flat Dirichlet distributions for all rows. All rows are updated during the M-step of the algorithm to learn the dust profiles of the different sources potentially contributing to the evidentiary samples.
4. Inferences on sources’ dust profiles and mixing proportions

Following convergence of the algorithm, we obtain updated Dirichlet distribution parameter matrices $A$ and $H$. The marginal distribution of each of the Dirichlet distributions in the rows of $A$ and $H$ gives the posterior distributions for the multinomial parameters $\Theta_{ls}$ and $B$, respectively. Hence,

(a) the contribution of the $m$-th source to the $s$-th sample from location $l$ is
\[ \theta_{ism} \sim \text{Beta}(\alpha_{lm}, \sum_{i \neq m} \alpha_{li}). \]
Note that, by construction, the expectation of $\theta_{ism}$ will be very close to 1 if location $l$ is one of the $K$ known locations and $l = m$.

(b) the proportion of the $t$-th particle type in the $m$-th source is $\beta_{mt} \sim \text{Beta}(\eta_{mt}, \sum_{i \neq t} \eta_{mi})$.

5. Worked example

As an example, the algorithm presented above is used to resolve two mixtures of dust particles provided by Stoney Forensic, Inc. (Chantilly, VA, USA). The data set is composed of “pure” and mixed samples of dust from two locations, labeled “AT” and “LQ”, consisting in (1) twelve samples of dust knowingly obtained from each of the two locations, and (2) two trace samples consisting in mixtures of dust obtained by mixing known proportions of dust from AT and LQ. A dust sample is characterised by a vector of counts for fourteen particle types. The data set is summarised in tables 1, 2 and 3. Our model is used to resolve the dust mixtures in the trace samples presented in Table 3 under three different scenarios:

(a) In the first scenario, both sources are considered known and can be sampled from $(K = 2, Q = 0)$;

(b) In the second scenario, location AT is known and can be sampled from, while LQ is unknown and cannot be studied $(K = 1, Q = 1)$;

(c) In the last scenario, location LQ is known and can be sampled from, while AT is unknown and cannot be studied $(K = 1, Q = 1)$.
Table 1: Twelve samples obtained from location “AT”.

|                         | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| Alkali Feldspar         | 189| 182| 200| 184| 254| 182| 181| 178| 139| 193| 229| 204|
| Alterite                | 21 | 20 | 9  | 20 | 15 | 11 | 21 | 19 | 12 | 28 | 32 | 20 |
| Biotite                 | 1  | 4  | 1  | 3  | 0  | 1  | 3  | 0  | 1  | 0  | 0  | 0  |
| Epidote                 | 3  | 7  | 6  | 11 | 12 | 3  | 7  | 12 | 5  | 4  | 7  |    |
| High Index              | 3  | 2  | 1  | 7  | 3  | 3  | 2  | 2  | 3  | 2  | 1  | 4  |
| Hornblende              | 0  | 2  | 2  | 4  | 5  | 4  | 2  | 3  | 2  | 1  | 0  | 2  |
| Iron Oxides             | 9  | 4  | 1  | 5  | 5  | 7  | 7  | 9  | 5  | 6  | 7  | 6  |
| Lithic Fragments        | 3  | 0  | 7  | 12 | 3  | 2  | 3  | 2  | 3  | 4  | 0  |    |
| Muscovite               | 0  | 0  | 1  | 1  | 0  | 3  | 0  | 0  | 3  | 0  | 0  | 0  |
| Opaques                 | 16 | 14 | 10 | 9  | 37 | 18 | 25 | 20 | 42 | 16 | 9  | 15 |
| Plagioclase             | 5  | 0  | 2  | 5  | 7  | 1  | 0  | 2  | 0  | 10 | 5  | 10 |
| Quartz                  | 74 | 74 | 75 | 63 | 112| 90 | 62 | 71 | 101| 56 | 54 | 94 |
| Titanite                | 0  | 0  | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  |
| Other                   | 11 | 11 | 5  | 12 | 17 | 23 | 3  | 6  | 8  | 9  | 9  | 8  |

Table 2: Twelve samples obtained from location “LQ”

|                         | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| Alkali Feldspar         | 18 | 26 | 29 | 20 | 31 | 33 | 28 | 30 | 39 | 22 | 30 | 22 |
| Alterite                | 4  | 4  | 4  | 6  | 7  | 7  | 10 | 4  | 5  | 12 | 9  | 10 |
| Biotite                 | 16 | 10 | 22 | 13 | 10 | 12 | 26 | 13 | 11 | 25 | 18 | 20 |
| Epidote                 | 10 | 5  | 13 | 11 | 7  | 7  | 8  | 19 | 6  | 11 | 13 | 5  |
| High Index              | 3  | 0  | 1  | 2  | 2  | 0  | 0  | 1  | 0  | 0  | 2  | 2  |
| Hornblende              | 73 | 55 | 64 | 68 | 61 | 91 | 93 | 68 | 51 | 73 | 82 | 75 |
| Iron Oxides             | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 2  | 0  | 0  |
| Lithic Fragments        | 5  | 7  | 6  | 6  | 2  | 4  | 0  | 6  | 3  | 6  | 7  | 11 |
| Muscovite               | 0  | 3  | 0  | 2  | 2  | 0  | 1  | 4  | 0  | 2  | 5  | 1  |
| Opaques                 | 5  | 0  | 2  | 4  | 8  | 10 | 8  | 5  | 3  | 7  | 4  | 3  |
| Plagioclase             | 46 | 37 | 47 | 45 | 52 | 39 | 14 | 16 | 13 | 33 | 35 | 27 |
| Quartz                  | 153| 159| 151| 145| 174| 150| 128| 161| 195| 134| 137| 156|
| Titanite                | 2  | 4  | 6  | 1  | 4  | 6  | 5  | 6  | 2  | 2  | 4  | 2  |
| Other                   | 3  | 5  | 1  | 4  | 9  | 6  | 8  | 2  | 3  | 5  | 2  | 5  |

Table 3: Two samples obtained by mixing dust from “AT” and “LQ” in known proportions. The proportions are indicated in the last column.

| Trace 1 | Trace 2 |
|---------|---------|
| 312     | 104     |
| 31      | 16      |
| 5       | 23      |
| 12      | 16      |
| 16      | 200     |
| 9       | 2       |
| 7       | 3       |
| 1       | 13      |
| 32      | 48      |
| 12      | 48      |
| 151     | 240     |
| 1       | 5       |
| 17      | 10      |
| 0.10/0.90 | 0.80/0.20 |

M. A. Ausdemore et al.
5.1. Resolving the mixtures in table when both locations are known

In this example, the algorithm is provided with all 26 samples described above: 12 “pure” samples from each location and 2 mixed samples. We initialise the algorithm by designing the matrices $H$ and $A$. To allow the model to freely determine the particle profiles of the two sources in the dust mixtures, the matrix $H$ is set to:

$$H_{\text{initial}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

To ensure that the algorithm correctly learns the dust profiles of the two known sources from the samples obtained from locations AT and LQ, the rows of matrix $A$ associated with these samples are heavily weighted in the dimension corresponding to sources AT and LQ, while the rows of matrix $A$ corresponding to the trace samples are set to 1:

$$A_{\text{initial}} = \begin{bmatrix} \alpha_{\text{AT},1} & \cdots & \alpha_{\text{AT},12} & \alpha_{\text{LQ},1} & \cdots & \alpha_{\text{LQ},12} & \alpha_{e_1} & \cdots & \alpha_{e_1} & \alpha_{e_2} & \cdots & \alpha_{e_2} \\ 150 & \cdots & 150 & 1 & \cdots & 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix}$$

Upon introducing the samples representing the known sources and the two trace samples into the model and observing convergence, we obtain (note that, by design of the algorithm, the rows corresponding to $\alpha_{\text{AT},s}$ and $\alpha_{\text{LQ},s}$ are not updated):

$$H_{\text{converged}} = \begin{bmatrix} 150.00 & 14.94 & 1.23 & 5.36 & 2.49 & 1.67 & 4.90 & 2.77 & 0.58 & 14.70 & 2.72 & 59.54 & 0.25 & 8.00 \\ 26.12 & 6.90 & 16.18 & 9.61 & 1.22 & 70.18 & 0.34 & 5.35 & 1.78 & 4.86 & 32.49 & 150.00 & 3.91 & 4.54 \end{bmatrix}$$

$$A_{\text{converged}} = \begin{bmatrix} \alpha_{e_1} & \alpha_{e_2} \\ 178.46 & 21.54 \\ 38.85 & 161.15 \end{bmatrix}.$$

The rows of these matrices are the parameters of the posterior marginal distributions described in section and enable to study the distributions of $\beta_{\text{AT}}$ and $\beta_{\text{LQ}}$, which represent the particle profiles of the sources present in the dust samples, and of $\theta_{e_1}$ and $\theta_{e_2}$, which represent the mixing proportions of the two sources in the evidentiary samples. The resulting marginal posterior distributions of $\beta_{\text{AT}}$ and $\beta_{\text{LQ}}$ are displayed in figure. The resulting marginal posterior distributions of $\theta_{e_1}$ and $\theta_{e_2}$ are displayed in figure.
Fig. 2. Plots of posterior distributions corresponding to the elements of $B$ for location AT (top) and location LQ (bottom) when both locations are known. Each plot is associated with one of the fourteen particle types. The vertical blue line corresponds to the point estimates of the mixing proportions, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.

Fig. 3. Plots of posterior distributions corresponding to the elements of $A$ when both AT and LQ are known and "pure" samples can be obtained from both. The first pair of plots corresponds to a trace mixture where 90% of the particles originate from location AT and 10% of the particles originate from location LQ. The second pair of plots corresponds to a trace mixture where 20% of the particles originate from location AT, and 80% originate from location LQ. The vertical blue line corresponds to the known mixing proportion, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.
Figure 2 shows that the model is able to effectively extract the dust profiles of the sources. All posterior distributions are sharply centred on their mean and mode.

Figure 3 shows that the model is also able to extract the mixing proportions of the locations within the dust mixtures. The known mixing proportions are within the 95% Highest Posterior Density Intervals (HPDI’s), and the posterior distributions show little variance.

5.2. Resolving the mixtures in table 3 when location AT is known

In this example, the algorithm is provided with 14 samples described above: 12 “pure” samples from location AT and 2 mixed samples. Matrix $H$ is initialised as in the previous example, since we are still interested in learning the dust profiles of both sources.

However, matrix $A$ has a different number of rows to reflect that no sample representing source LQ has been observed. Hence, only the rows of matrix $A$ associated with the samples obtained from location AT are heavily weighted in the dimension corresponding to source AT. As before, the rows of $A$ corresponding to the trace samples are set to 1:

$$A_{\text{initial}} = \begin{bmatrix} \alpha_{\text{AT},1} \\ \vdots \\ \alpha_{\text{AT},12} \\ \alpha_{e1,1} \\ \alpha_{e2,1} \end{bmatrix} = \begin{bmatrix} 150 \\ \vdots \\ 150 \\ 1 \\ 1 \end{bmatrix}.$$

Upon introducing the twelve samples from location AT and the two trace samples into the model and observing convergence, we obtain (note that, by design of the algorithm, the rows corresponding to $\alpha_{\text{AT},s}$ are not updated):

$$H_{\text{converged}} = \begin{bmatrix} 150.00 & 14.86 & 1.19 & 5.31 & 2.49 & 1.64 & 4.93 & 2.72 & 0.51 & 14.60 & 2.66 & 59.49 & 0.24 & 7.92 \\ 36.96 & 9.03 & 17.03 & 11.41 & 1.55 & 71.45 & 0.76 & 6.83 & 2.69 & 6.86 & 34.55 & 150.00 & 4.21 & 6.34 \end{bmatrix}.$$

$$A_{\text{converged}} = \begin{bmatrix} \alpha_{e1,1} \\ \alpha_{e2,1} \end{bmatrix} = \begin{bmatrix} 177.13 \\ 29.65 \end{bmatrix}.$$

The resulting marginal posterior distributions of $\theta_{e1,1}$ and $\theta_{e2,1}$ are displayed in figure 5. Even though, the algorithm is only provided with “pure” samples from one single source, figure 4 shows that the model remains capable of effectively extracting the profiles of the sources. That said, by comparing figures 2 and 4, we note that the modes/means of the posterior distributions for the profile of source LQ are not as well aligned with the proportion estimates of the particle types when no sample from source LQ is observed.

Figure 5 shows that the model is able to extract accurately the mixing proportions of the locations within the dust mixture dominated by location AT, and less accurately the mixing proportions in the dust mixtures dominated by location LQ. These results seem to indicate that the accuracy of the predicted particle profile for the unobserved source impacts the detection of the mixing proportions when the unobserved source is present in large quantity in the mixtures.
Fig. 4. Plots of posterior distributions corresponding to the elements of $B$ for location AT (top) and location LQ (bottom) when location AT is known and location LQ is unknown. Each plot is associated with one of the fourteen particle types. The vertical blue line corresponds to the point estimates of the mixing proportions, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.

Fig. 5. Plots of posterior distributions corresponding to the elements of $A$ when AT is known and LQ is unknown. The first pair of plots corresponds to a trace mixture where 90% of the particles originate from (known) location AT, and 10% originate from location LQ. The second pair of plots corresponds to a trace mixture where 20% of the particles originate from (known) location AT, and 80% originate from location LQ. The vertical blue line corresponds to the true mixing proportion, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.
5.3. Deconvolving the mixtures in Table 3 when source LQ is known

In the final example, we assess the model’s ability to deconvolve the trace mixtures in table 3 when location LQ is known, and location AT is unknown.

Matrix $\mathbf{H}$ remains the same as in the two previous examples. The known samples that are introduced into the model are now from location LQ. We account for this difference in information by weighting the elements of matrix $\mathbf{A}$ corresponding to location LQ, rather than to location AT:

$$\mathbf{A}_{\text{initial}} = \begin{bmatrix} \alpha_{LQ,1} \\ \vdots \\ \alpha_{LQ,12} \\ \alpha_{e_1,1} \\ \alpha_{e_2,1} \end{bmatrix} = \begin{bmatrix} 1 & 150 \\ \vdots & \vdots \\ 1 & 150 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Upon observing convergence, we obtain the updated matrices $\mathbf{H}_{\text{converged}}$ and $\mathbf{A}_{\text{converged}}$:

$$\mathbf{H}_{\text{converged}} = \begin{bmatrix} 150.00 & 16.16 & 2.45 & 6.40 & 2.93 & 5.83 & 4.82 & 4.01 & 0.91 & 16.58 & 5.41 & 68.63 & 0.734 & 9.23 \\ 26.07 & 6.85 & 16.13 & 9.55 & 1.18 & 70.03 & 0.25 & 5.29 & 1.75 & 4.79 & 32.34 & 150.00 & 3.88 & 4.49 \end{bmatrix}$$

$$\mathbf{A}_{\text{converged}} = \begin{bmatrix} \alpha_{e_1,1} \\ \alpha_{e_2,1} \end{bmatrix} = \begin{bmatrix} 192.82 & 7.18 \\ 42.54 & 157.46 \end{bmatrix}.$$

The resulting marginal posterior distributions of $\mathbf{\beta}_{AT}$ and $\mathbf{\beta}_{LQ}$ are displayed in figure 6. The resulting marginal posterior distributions of $\mathbf{\theta}_{e_1,1}$ and $\mathbf{\theta}_{e_2,1}$ are displayed in figure 7.

As in the previous example, figure 6 shows that the model is able to extract the two location profiles present in the evidentiary samples: the mean and mode of the distributions prove to be reasonably similar to the proportion estimates of particle types. As before, the model is able to deconvolve the second trace mixture where the trace mixtures is mostly composed of dust originating from the observed location (figure 7(right)), but struggles to appropriately deconvolve the mixture in the sample mostly composed of the unobserved location AT (figure 7(left)).
Fig. 6. Plots of posterior distributions corresponding to the elements of $B$ for location AT (top) and location LQ (bottom) when location AT is unknown and location LQ is known. Each plot is associated with one of the fourteen particle types. The vertical blue line corresponds to the proportion estimates of the particle types, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.

Fig. 7. Plots of the posterior distributions corresponding to the elements of $A$ when location LQ is known and location AT is unknown. The first pair of plots corresponds to a mixture where 90% of the particles originate from location AT, and 10% originate from (known) location LQ. The second pair of plots corresponds to a mixture where 20% of the particles originate from known location AT, and 80% originate from (known) location LQ. The vertical blue line corresponds to the known mixing proportion, the vertical green line corresponds to the mean of the resulting posterior distribution, and the vertical red line corresponds to the mode of the resulting posterior distribution. The grey shaded region corresponds to the 95% HPDI.
6. Model performance

The results presented in sections 5.2 and 5.3 are obtained by constraining the range of values that can be considered by the L-BFGS-B optimisation method for $H$ and $A$ in the M-Step of our algorithm. The following constraints are used: $\alpha_{lm} \in [1, 200]$, $\sum_{m=1}^{M} \alpha_{lm} = 200$ and $\eta_{mt} \in [0.1, 150]$. Depending on the constraints, we noted that our algorithm does not necessarily approximate the correct solution and may suffer from a lack of identifiability. This is particularly flagrant when “pure” samples from only one of the two sources are used, or when a low number of “pure” samples per source are used. A search of the literature for discussions on the identifiability of LDA models indicated that this issue has only been considered by very few authors (Rabani et al., 2013; Vandermeulen and Scott, 2015) without providing satisfactory solutions. Furthermore, LDA-based models typically proposed in the literature do not compare their results to ground truth, which prevents assessing their accuracy and identifiability.

To study the behaviour of our algorithm, we simulate dust mixtures in which the particle counts and mixing proportions vary. We simulate two cases:

(a) both sources are known and can be sampled from;

(b) source AT is known and can be sampled from. Source LQ is not known and its profile is left to be learned by the model from the mixtures.

We used the point estimates for the proportions of the particle types for sources AT and LQ obtained from tables 1 and 2 to obtain dust samples from these sources before mixing them. In each simulation, we consider two evidentiary sets of five samples each, composed of mixtures of dust from the sources AT and LQ. The number of samples was selected to correspond to a realistic forensic scenario where both trace and control locations would be sampled several times to study their respective variability, but where the number of samples would not be so high that their examination would be too time consuming. The mixing proportions associated with the second set of five samples remain fixed such that $\theta_{e_2,s} = (0.51, 0.49)$, for $s \in \{1, 2, 3, 4, 5\}$. For each of the nine mixing proportions considered for $\theta_{e_1,s}$, the particle count in each sample of dust mixture varies so that $N_s \in \{100, 300, 500, \ldots, 1500, 1700, 1900\}$, $s \in \{1, 2, 3, 4, 5\}$. This results in a total of 90 different simulated scenarios, which are executed ten times each. The average model performance is evaluated.

6.1. Two known sources

The simulations performed under scenario (a) show that the model is able to deconvolve the dust mixtures with minimal error with respect to the known values used to create the mixtures. Figure 3 shows that the accuracy of the prediction for the proportions of
the different sources in the mixtures is not a function of the number of particles, and that overall good precision can be achieved when both sources are known. The accuracy of the prediction is better for balanced mixtures than for extreme mixtures. The mean appears to be a better predictor than the mode; however, the difference between the mean/mode predicted values and the known values is never greater than 5%. The same observations can be made regarding the accuracy of the prediction for the proportion of particle types in the dust profiles for both locations, when samples from both locations are observed. The results are comparable to those displayed in figure 10 and are not presented here.

Fig. 8. Average predicted values (over 10 simulations, with two known locations) for $\theta_{e1,s,AT}$ and $\theta_{e2,s,AT}$ using the means (left) and the modes (right) of the distributions of the parameters. The dashed black lines correspond to the mixing proportions that are used to generate the mixtures.

6.2. One known source

The results obtained by our simulations in scenario (b) are presented in figures 9, 10, and 11. Figure 9 shows that the accuracy of the predicted mixing proportions when samples from location LQ are not observed is lower than when both locations are observed. The lower accuracy is observed in all mixtures, including balanced ones, and in the control mixture in Trace 2. While running the simulations, we noted that our model has a strong tendency to over estimate the proportion of the major contributing source, irrespectively of whether it is the observed source, and even to drive the proportion of the minor contributor to zero. As previously, the accuracy of the predictions does not appear to be a function of the number of particles. Figure 10 shows the accuracy of the prediction
Deconvolution of dust mixtures by latent Dirichlet allocation

Fig. 9. Average predicted values (over 10 simulations, with one known location) for $\theta_{e_1,s,AT}$ and $\theta_{e_2,s,AT}$ using the means (left) and the modes (right) of the distributions of the parameters. The dashed black lines correspond to the mixing proportions that are used to generate the mixtures.

of the dust profile for known location AT. Since samples from location AT are observed, the accuracy is unsurprisingly good. On the contrary, the accuracy of the prediction for the dust profile of unknown location LQ presented in figure [11] is not as impressive. We believe that the lack of ability of our model to accurately infer the dust profile of unobserved sources of dust is one of the key reasons behind the lack of accuracy in the prediction of the mixing proportions observed in figure [9].

Other series of simulations, with increasing numbers of trace and control samples did not permit to observe a pattern of increasing accuracy, even when the number of control samples is much larger (ratio of 10:1) than the number of trace samples. Finally, series of simulations that considered both sources unknown were attempted but showed very low accuracy. This was surprising given that LDA was designed with the idea of detecting topics in corpus of documents without requiring to “learn” each topic individually.

The main difference between our model and that originally proposed by [Blei et al. 2003] is that we use asymmetric Dirichlet distributions instead of symmetric ones ([Blei et al. 2003], p. 1006, footnote 2). Overall, we are not sure whether the lack of accuracy of our model originates from a general lack of identifiability of LDA models; from the large number of parameters to be assigned in $H$ and $A$, which may require a much larger number of samples than the number considered by our simulations; or from some instability of the numerical optimisation methods used in the M-step of our algorithm. Future developments of our method may involve comparing it to a Gibbs sampler implementation of the model in order to determine the role of the optimisation process in the lack of ac-
accuracy of the algorithm. Additional validation of the model is also needed to determine its ability to resolve mixtures of more than two sources, as well as dust sources with more or less similar profiles.

Fig. 10. Average predicted values of the means of the distributions of $\beta_{AT,t}$ when “pure” samples from location AT are used to discover the dust profile of source AT.
Dust particles recovered from beneath the soles of an individual’s shoes consist of a mixture of dust particles collected from different sources and may be indicative of the locations recently visited by that individual. In particular, this dust may reveal his presence at a location of interest, e.g., the scene of a crime. In this paper, we propose a model for the deconvolution of mixtures of dust originating from $M$ sources. Our goal is to infer the particle profiles of the $M$ sources, as well as their respective contribution to the mixture. Our overarching purpose is to enable the use and interpretation of dust evidence in order to determine, for example, if the dust recovered from under a pair of shoes contains particles originating from a given crime scene.

### 7. Conclusion
We describe the profiles of each of the $M$ dust sources using a multinomial distribution over a fixed number of particle types. We use latent Dirichlet allocation (LDA) to define the probability distribution of the dust mixture. We use variational Bayesian inference (VBI) to study the source mixing proportions and particle profiles of each of the $M$ sources present in the dust mixture. Finally, we propose a method to constrain our model to learn the dust profiles of known sources using control samples collected at locations of interest (such as crime scenes, houses of suspects, etc.), while retaining the model’s ability to learn the dust profiles of sources that are present in the mixtures but cannot be directly observed (such as the unknown location where a body is buried).

We test the performance of the model using real and simulated data. We find that our model is able to effectively extract the particle profiles of the sources in the mixtures present in the real data set when “pure” samples from all sources present in the mixtures are used to resolve them. The accuracy of our model decreases with the number of contributing sources that can be studied. Our simulations indicate that the accuracy of the model is not a function of the number of particles in the dust mixtures, and that no clear pattern emerges in terms of its accuracy as a function of the number of control and trace samples.

We observe that our model behaves very differently depending on the constraints used for the numerical optimisation of its Dirichlet parameters. The lack of consistency of our model may be rooted in a lack of identifiability of LDA models in general. Very little has been published on the subject of identifiability of LDA models. Furthermore, most models proposed in the literature are not tested using datasets with known parameters and, therefore, their accuracy cannot be assessed. This is clearly an open field for future research.

The performance of our model in various situations needs to be extensively tested before it can be used in forensic practice. That said, it is capable of resolving mixtures of dust sources, and thus, of enabling forensic examiners to quantitatively support their inference of the presence of a suspect/object at a location of interest by examining dust evidence. While the transfer and examination of dust evidence was only considered as a theoretical concept by the founding fathers of forensic science, our model shows that dust particles have a great potential as a forensic tool in the near future.

Acknowledgements

This project was supported in part by Award No. 2016-DN-BX-0146 awarded by the National Institute of Justice (Office of Justice Programs, U.S. Department of Justice) to Stoney Forensic, Inc. (Chantilly, VA, USA), and Award No. 2014-IJ-CX-K088 awarded by the National Institute of Justice to South Dakota State University (Brookings, SD, USA). The opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the Department of Justice or the National Science Foundation. We would like to thank Dr. David Stoney for bringing up this interesting problem and providing the data.
References

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer Texts in Statistics.

Blei, D., A. Kucukelbir, and J. McAuliffe (2016). Variational inference: A review for statisticians. [arXiv:1601.00670](http://arxiv.org/abs/1601.00670).

Blei, D. and J. McAuliffe (2007, March). Supervised topic models. *Proceedings of NIPS*, 121–128.

Blei, D., A. Ng, and M. Jordan (2003, January). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Byrd, R., P. Lu, J. Nocedal, and C. Zhu (1994). A limited memory algorithm for bound constrained optimisation. Technical Report NAM-08, Northwestern University.

Hoffman, M., D. Blei, C. Wang, and J. Paisley (2013, May). Stochastic variational inference. *Journal of Machine Learning Research 14*, 1303–1347.

Lacoste-Julien, S., F. Sha, and M. Jordan (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, Volume 22.

Rabani, Y., L. Schulman, and C. Swamy (2013). Learning mixtures of arbitrary distributions over large discrete domains. [arXiv:1212.1527v3](http://arxiv.org/abs/1212.1527v3).

Robert, C. (2007). *The Bayesian Choice*. Springer Texts in Statistics.

Rosen-Zvi, M., C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers (2010, January). Learning author-topic models from text corpora. *ACM Transaction on Information Systems 28*(1).

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 487–494. AUAI Press.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths (2004). Probabilistic author-topic models for information discovery. *10th ACM SigKDD conference knowledge discovery and data mining*.

Vandermeulen, R. and C. Scott (2015). On the identifiability of mixture models from grouped samples. [arXiv:1502.06644](http://arxiv.org/abs/1502.06644).

Wang, C., D. Blei, and L. Fei-Fei (2009). Simultaneous image classification and annotation. *IEEE Conference on Computer Vision and Pattern Recognition*, 1903–1910.

Zhu, C., R. Byrd, P. Lu, and J. Nocedal (1994). L-BFGS-B - fortran subroutines for large-scale bound constrained optimisation. Technical Report NAM-11, Northwestern University.

Zhu, J. and A. Ahmed (2012, August). MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research 13*, 2237–2278.
Appendices

A. Topic modelling and dust modelling: parallel terms

(a) *Dust particle*: a dust particle corresponds to a word in a topic model.

(b) *Sample*: a collection of dust particles. A sample of dust corresponds to a document in a topic model.

(c) *Source*: a process or geographical area that yields dust. A source corresponds to a topic in a topic model.

(d) *Sampling Location*: a geographical area or an object where a set of samples of dust is obtained. A location may resemble an author in a topic model in the sense that both may generate samples.

(e) *Particle types*: a pre-defined set of categories that are used to classify the dust particles. A list of particle types corresponds to the vocabulary or dictionary of words in a topic model.

B. Tables of notation, variables and parameters used in development

| Description | Variable Type |
|-------------|---------------|
| $l$ | Indicates which location a sample corresponds to, where $l \in \{1, \ldots, L\}$ | Observed variable |
| $s$ | Indicates which sample is being considered, where $s \in \{1, \ldots, S_l\}$ | Observed variable |
| $n$ | Indicates which particle is being considered, where $n \in \{1, \ldots, N_s\}$ | Observed variable |
| $m$ | Indicates which source produced a particle, where $m \in \{1, \ldots, M\}$, where $M = Q + K$ | Latent variable |
| $t$ | Indicates which particle type is being considered, where $t \in \{1, \ldots, T\}$ | Observed variable |

Table 4: Table of subscripts and superscripts used to describe dust particles and dust samples
Table 5: Table of variables and parameters used in development

| Description | Variable Type |
|-------------|---------------|
| X           | Observed variable |
| $x_{ls}$    | Observed variable |
| $x_{lsnt}$  | Observed variable |
| Z           | Latent variable |
| $z_{lsnm}$  | Latent variable |
| $\Theta$    | Latent variable |
| $\theta_{ls}$ | Latent parameter |
| $\theta_{lsm}$ | Latent parameter |
| $B$         | Latent parameter |
| $\beta_{m}$ | Latent parameter |
| $\beta_{mt}$ | Latent parameter |
| $A$         | Latent parameter |
| $\alpha_{l}$ | Latent parameter |
| $\alpha_{lm}$ | Latent parameter |
| $H$         | Latent parameter |
| $\eta_{m}$  | Latent parameter |
| $\eta_{mt}$ | Latent parameter |
L-BFGS-B is a limited-memory quasi-Newton method that incorporates box constraints into the optimisation process \cite{Byrd1994, Zhu1994}. The method serves to minimise (or, likewise, maximise) a function, \( f(x_k) \), subject to the condition that \( l \leq x_k \leq u \), where \( l \) and \( u \) represent the lower and upper bounds specified for the current iterate, \( x_k \). L-BFGS-B avoids the computational cost of explicitly computing a Hessian matrix, \( \nabla^2 f(x_k) \), and instead, approximates \( \tilde{\nabla}^2 f(x_k) \) using the gradient of the function to be optimised, \( \nabla f(x_k) \), and is thus efficient for large-scale problems. The algorithm proceeds by defining a quadratic function in terms of the original function, \( f(x_k) \), the gradient, \( \nabla f(x_k) \), and a positive definite limited-memory Hessian approximation, \( \tilde{\nabla}^2 f(x_k) \). Minimising this quadratic function provides an approximate solution for the next iterate, \( \tilde{x}_{k+1} \), from which we can obtain a search direction. This search direction allows us to find the next iterate \( x_{k+1} \). Given \( x_{k+1} \), a new gradient \( \nabla f(x_k) \) and limited-memory Hessian \( \tilde{\nabla}^2 f(x_k) \) are computed, and, pending satisfaction of some convergence criterion, the next iteration begins.

In deconvolving mixtures of dust particles, L-BFGS-B can be used to obtain updates for the matrices \( H \) and \( A \) by maximising the lower bound function \( \mathcal{L}(q(\Theta, B, Z)) \) with respect to each \( \eta_{mt} \) and \( \alpha_{lm} \). Clearly, the gradient plays a central role in this method, and so we use this appendix to define the gradients used in the L-BFGS-B algorithm to obtain the updates for \( H \) and \( A \). For further discussion on the L-BFGS-B method, see \cite{Byrd1994} and \cite{Zhu1994}.

It is convenient to begin by defining the lower bound in terms of the considered parameters. We note that the lower bound functions \( \mathcal{L}_{\eta_{mt}} \) and \( \mathcal{L}_{\alpha_{lm}} \) can be optimised independently, since neither function depends on the parameters of the other:

\[
\mathcal{L}_{\eta_{mt}} = \log \Gamma \left( \sum_{j=1}^{T} \eta_{jt} \right) - \sum_{t=1}^{T} \log \Gamma (\eta_{mt}) + \sum_{t=1}^{T} (\eta_{mt}) \left( \Psi (\lambda_{smt}) - \Psi \left( \sum_{j=1}^{T} \lambda_{smj} \right) \right),
\]

\[
\mathcal{L}_{\alpha_{lm}} = \log \Gamma \left( \sum_{j=1}^{M} \alpha_{lj} \right) - \sum_{m=1}^{M} \log \Gamma (\alpha_{lm}) + \sum_{m=1}^{M} (\alpha_{lm}) \left( \Psi (\gamma_{lsm}) - \Psi \left( \sum_{j=1}^{M} \gamma_{lsj} \right) \right).
\]

Indeed, specifying \( \mathcal{L}_{\alpha_{lm}} \) and \( \mathcal{L}_{\eta_{mt}} \) makes it straightforward to define the gradients:

\[
\nabla \mathcal{L}_{\eta_{mt}} = \left( \frac{\partial \mathcal{L}_{\eta_{mt}}}{\partial \eta_{m1}}, \ldots, \frac{\partial \mathcal{L}_{\eta_{mt}}}{\partial \eta_{mT}} \right),
\]

\[
\nabla \mathcal{L}_{\alpha_{lm}} = \left( \frac{\partial \mathcal{L}_{\alpha_{lm}}}{\partial \alpha_{l1}}, \ldots, \frac{\partial \mathcal{L}_{\alpha_{lm}}}{\partial \alpha_{lM}} \right).
\]

(12)
where each $\frac{\partial \mathcal{L}}{\partial \alpha_i}$ and $\frac{\partial \mathcal{L}}{\partial \eta_m}$ is given by:

$$\frac{\partial \mathcal{L}_{\eta_{mt}}}{\partial \eta_{mt}} = \Psi \left( \sum_{j=1}^{T} \eta_{mj} \right) - \Psi(\alpha_{mt}) + \Psi(\gamma_{smt}) - \Psi \left( \sum_{j=1}^{T} \gamma_{smj} \right)$$

$$\frac{\partial \mathcal{L}_{\alpha_{lm}}}{\partial \alpha_{lm}} = \Psi \left( \sum_{j=1}^{M} \alpha_{lj} \right) - \Psi(\alpha_{lm}) + \Psi(\gamma_{slm}) - \Psi \left( \sum_{j=1}^{M} \gamma_{slj} \right).$$

(14)