Kernel-based methods for source identification using very small particles from carpet fibers

D. E. Armstrong, MSc.[a], C. Neumann, Ph.D.[a], C. P. Saunders, Ph.D.[a], D. Gantz, Ph. D.[b], J. Miller, Ph. D.[c], D. A. Stoney, Ph.D.[d]

[a] *Department of Mathematics and Statistics, South Dakota State University, Brookings, SD*
[b] *Department of Information Sciences and Technology, George Mason University, Fairfax, VA*
[c] *Department of Statistics, George Mason University, Fairfax, VA*
[d] *Stoney Forensic, Inc., 14101 Willard Road, Suite G, Chantilly, VA 20151*

Abstract

The objective comparison of complex signals in chemistry, and more particularly in forensic chemistry, with the view of inferring the source of a particular 'trace' object is an ongoing issue. In this paper, we propose a method that enables to assign a probability distribution to control material from any given source based on its chemical signal and to subsequently infer the source of a trace object using a simple Bayes classifier. Our method takes advantage of the dimension reduction and discriminative powers of kernels, and only requires the estimation of three parameters (once a kernel is chosen). We illustrate the application of our method to the inference of the source of trace objects based on very small particles (VSP) that can be found on their surfaces. VSPs are picked up in the environment(s) where the trace object has been in. These VSPs can (1) offer information about the geographic origins of the objects; (2) help discriminate between multiple mass-manufactured objects that would be otherwise identical. In this project, we use VSPs recovered from carpet fibers throughout the United States and apply our method to (1) reduce the complexity of compositional data obtained by SEM/EDS; (2) infer the source of the trace material. This method can be extended to VSPs found on other types of recovered forensic materials such as weapons, drugs, or IEDs (improvised explosive device), and to other types of chemical signals.

Keywords: Very small particles, Kernel based-methods, High-dimensional classification, Closed-set identification

Email address: Cedric.Neumann@me.com (C. Neumann, Ph.D.)

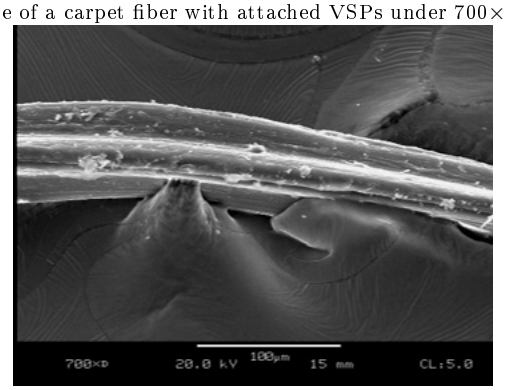
1. Introduction

Forensic chemists, and chemists in general, are often asked to compare the complex signals resulting from the chemical analyses of two objects in order to determine whether they have the same composition and to make inference on the commonality of their source. Objects of forensic interest are often small and degraded, resulting in noisy signals. Furthermore, many 'trace' objects may be recovered in connection with a crime, not to mention the multiple objects that are sampled from each suspected source to characterize its variability. Thus, the inference process involves the comparison of the trace objects to each other, and to the multiple objects sampled from each reference source. In this paper, we propose a method developed to compare complex chemical signals from multiple objects to each other in order to infer whether they all come from a unique source. Our method can be used for almost any type of signal with a minimum number of constraints. We illustrate our method by applying it to a particularly challenging forensic problem: the inference of the source of sets of very small particles.

Very small particles (VSP) (Figure 1) are airborne or contact particles ranging from about $0.1\mu m$-$100\mu m$ in size. VSP 'piggy-back ride' on any piece of evidence, such as drug packaging, weapons, contraband and clothing. They occur in complex mixtures, where any mixture may include a tremendous variety of particle types. The particles are acquired when manufactured materials are exposed to alternative environments. Particle mixtures reflect environmental conditions which affected the object of forensic interest [13]. The presence, identity and relative quantities of different particle types provide an untapped source of variability between mass-manufactured objects that can be used to augment the weight of associations or discriminations between multiple pieces of evidence, such as fibers, glass fragments, paint chips, or plastic. VSP has many advantages over traditional trace[1] examination:

1. VSPs can be recovered from virtually any material and analyzed directly by computer-assisted Scanning Electron Microscope with Energy Dispersive X-Ray Spectrometer (SEM/EDS), without involving different sequences of presumptive and confirmatory techniques;
2. VSPs can be used to determine the (lack of) commonality of the place of origin of objects of different types (e.g., the first piece of evidence is fiber and the second piece of evidence is plastic), and therefore extend the versatility of recovered material;
3. Under suitable conditions, VSPs can be used to constrain or infer the geographical location of origin of a particular object (e.g. drug shipment), even though the raw material (e.g. plastic material) has been mass-manufactured elsewhere.

Figure 1: Image of a carpet fiber with attached VSPs under 700× magnification.



Efforts to demonstrate in principle the usefulness of VSPs for identifying the source[2] of forensic evidence are limited [3, 13]. The surface of any object may be coated with thousands of VSPs; furthermore, the analysis of any one VSP by SEM/EDS results in a high-dimensional vector of compositional data representing the

[1] A trace is defined as material originating from an unknown source and recovered in connection with a crime

[2] In this case the source is defined as a place of origin.

relative proportions of targeted chemical elements. Each recovered object is then represented by an array of raw compositional data.

Inferring the source of a set of 'trace' VSPs involves characterizing the probability density functions of its many possible sources. We want to point out that the chemical composition of any given VSP is unlikely to be characteristic of a source. In other words, VSPs with similar chemical compositions can be categorized by type as determined by their dominant compounds and each source may display any mixture of VSP types. The discrimination power of VSPs arises from the relative proportions of the different particle types between different sources.

Previous efforts have attempted to reduce the complexity of the parameter space by defining the types of the VSPs [13]. Each source was then characterized by a vector of relative proportions of these target-particle types (TPTs) which could be used as parameter estimates in a multinomial model. The definition of TPTs involved the use of unsupervised clustering techniques with their inherent drawbacks (e.g. arbitrary choice of a limited number of TPTs to keep the dimension of the parameter space reasonable). Many of these drawbacks are due to the inherent difficulty of dealing with the noise resulting from the analytical process (e.g., VSPs are contaminated themselves by other VSPs and may show the presence of foreign elements) or by particles that do not correspond to any of the defined TPTs.

In this paper, we propose to circumvent several of these issues by applying a kernel directly to the raw data in order to propose a model with a 3-dimensional parameter space. This process enables us to draw the basis of an inferential system to associate or discriminate the sources of multiple sets of VSPs using a limited number of assumptions while retaining most of the discrimination associated with the variability of VSPs. We show that our method enables us to directly define a likelihood function for any given source and use it to estimate the 3 parameters for that source [3]. Contrary to other kernel-based methods proposed in forensic science and biometric literature [4, 1], our method captures the dependency structure between multiple objects originating from a given source and is therefore particularly suitable to our problem, where multiple sets of VSPs may be recovered in order to characterize the within-variability of a source. The dependency structure is important to consider as it contains an appreciable amount of information that is disregarded in the popular univariate similarity-based techniques.

The purpose of this paper is to describe a method that can be used to compare the complex chemical signals of multiple objects by using kernels and model the probability density function of the resulting similarity measures. Providing that a suitable kernel is used to compare pairs of chemical signals, our method can be generalized to any type of signal, as illustrate with our VSPs example. The purpose of this paper is not to optimize the choice or performance of kernels for our example. The paper is organized as follows: Section 2 will introduce and discuss the data being considered; Section 3 introduces the development of the parametric kernel model to be studied and the algorithm used to implement it for VSPs; Section 4 discusses the results of the application of the parametric model to the VSP data; Section 5 will end with conclusions.

2. Data

The data considered in this paper was collected by Stoney et. al [12] and consists of carpet fibers collected from 90 sources throughout the United States. In each source, anywhere between 3-12 fibers were sampled by plucking them from the carpet and are used as reference material to characterize the source, while 1-10 fibers were sampled by brushing and are used as trace material. VSPs were extracted from the fibers via ethanol extractions and elemental composition was analyzed via SEM-EDS (see [12, Section II] for full recovery details). The resulting compositional data of relative percentages for 18 elements was recorded. The 18 elements are given in table 1.

Table 1: 18 elements targeted in this study

Sodium (Na)	Silicon (Si)	Chlorine (Cl)
Titanium (Ti)	Manganese (Mn)	Nickel (Ni)
Magnesium (Mg)	Phosphorus (P)	Potassium (K)
Vanadium (V)	Iron (Fe)	Copper (Cu)
Aluminum (Al)	Sulfur (S)	Calcium (Ca)
Chromium (Cr)	Cobalt (Co)	Zinc (Zn)

The number of VSPs per fiber is random and varies widely, resulting in large differences between the numbers of particles recovered at each location. For instance, source 7 shows only 31 VSPs, while source 72 has 18,170 VSPs. Overall, 43 sources have less than 100 VSPs. In this paper, we selected and used the 20 sources with the most VSPs. A heat map of these 20 sources is given in figure 2 where lighter colors represent greater proportions of the corresponding element.

Figure 2: Heat map for top 20 sources, selected by most numerous samples of VSPs. Vertical black lines separate VSPs by source (x-axis). Lighter colors represent greater proportions of the corresponding elements (y-axis) for each VSP.

This heat map allows us to observe signals for each source and gives an idea of the inherent capability and difficulty of using VSPs for separating sources. It also shows the inherent hierarchical nature of the sample: for example, all sources contain particles with a calcium-dominant component; all sources also contain particles with a silicon-dominant component; however, we observe that some sources contain a higher relative proportion of the particles with a calcium-dominant component (sources 12, 23, 31, 47, 59, 72, 77, 81, 89) vs. a silicon-dominant component (sources 44, 45, 48, 49, 52) than other sources which enables to discriminate between these sources.

Additionally, figure 2 shows the unbalanced number of VSPs per source, represented by the space between the vertical black lines, resulting from the collection of the fibers by crime scene personnel under realistic crime scene conditions and not under a fully controlled lab setting. The counts for the number of VSPs for source and trace samples is given in table 2.

Unfortunately, the different reference sets of VSPs recovered at a given location were pooled together during the chemical analysis. This prevented us from characterizing the within-location variability using observed data and we had to resort to non-parametric bootstrap sampling (see section 4.2 for details). Furthermore, due to the difference in the sampling techniques for the source and trace sets of VSPs, we decided to create 3 datasets with the VSPs: (1) training dataset with 2/3 of the reference VSPs; (2) testing dataset with the remaining 1/3 of the reference VSPs; (3) validation dataset with the trace VSPs.

Table 2: Counts of VSPs per location for source and trace samples

Location	Source set size	Trace set size
10	1289	1216
12	1089	1209
19	538	3363
23	549	1735
24	567	924
31	716	553
32	674	1785
44	781	793
45	3782	1820
47	935	2443
48	515	2120
49	2153	2751
52	2273	3535
53	1774	2083
58	3867	898
59	9739	7102
72	18170	3235
77	4474	1937
81	5256	582
89	1136	997

3. Methodology

The core of our method relies on the use of kernels to measure the level of similarity between multiple pairs of objects from a given source. In this paper, we will define a kernel in an analogous manner to a U-statistic [7, 14], as a function κ such that for all $\mathbf{x}_i,\ \mathbf{x}_j \in X,\ i, j \in \mathbb{N}^+,\ \kappa : X \times X \mapsto \mathbb{R}$ such that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa(\mathbf{x}_j, \mathbf{x}_i)$. A more restrictive definition, commonly used in computer science [5, 8, 11], is to require that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ where ϕ is a mapping from the input space X to a new feature space F.

There exists different classes of kernels, which can be more or less flexible[3] or robust. Different classes will be suitable for different types of data [11, 8]. When the kernel is based on the inner product, closure properties of kernels allow for their combinations to create new kernels. Kernels may be summed, scaled, multiplied, or used within other kernels while still satisfying the definition above. These properties and rich classes of kernels are used in machine-learning to deal with high-dimension and complex data forms, such as the ones commonly encountered in forensic science. Often times, calculating the full inner product$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$is computationally intensive as each mapping $\phi(\mathbf{x}_i),\ \phi(\mathbf{x}_j)$ must be calculated. Instead, $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ can be replaced by a 'score', $s_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, and is referred to as the 'kernel trick' [5]. The kernel trick is more efficient because the explicit mapping, ϕ, does not need to be computed.

When multiple objects are compared pairwise, the resulting kernel scores are organized into a kernel matrix $\mathbf{K}_{n\times n}$:

$$\mathbf{K}_{n\times n} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \kappa(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

Using n objects from a single source of VSPs will result in $N = \binom{n}{2}$ similarity scores (the upper or lower triangle of $\mathbf{K}_{n\times n}$). Let $s_{ij} = \kappa(x_i, x_j)$ be a measure of similarity between two objects, i and j. While

[3] Flexibility is defined with respect to tuneable parameters, see [11, 8] for more information.

multiple VSPs may be assumed to be conditionally independent from each other given their source, the resulting scores are correlated. For example the scores $s_{12},\ s_{13}$, representing the comparison between 3 different objects (i.e., 1, 2 and 3), have one object in common (i.e., 1); thus, s_{12} and s_{13} are dependent. There are 3 possible values for the correlation:

1. $Cor(s_{ij},\ s_{kl}) = 1$ if $i = k$ and $j = l$ (same pair of objects compared in both scores)
2. $Cor(s_{ij},\ s_{kl}) = r$ if $i = k$ and $j \neq l$ (one object in common in the scores)
3. $Cor(s_{ij},\ s_{kl}) = 0$ if $i \neq k$ and $j \neq l$ (no object in common in the scores).

By modeling the joint distribution of pairwise scores between VSPs recovered from a source, we can define a likelihood function for that source. This likelihood function is related to the scores and not directly to the particles. We hope that by accounting for the dependencies between these scores, we retain most of the information that is lost when using the kernel function.

3.1. Parametric model

We choose to represent the score s_{ij} between two identically independently distributed (iid) objects from a given source as a linear random effects model. The model for the score, s_{ij}, between pairs of iid objects, i and j, from a common source is given in equation 1:

$$s_{ij} = \theta + a_i + a_j + e_{ij} \tag{1}$$

where θ is the grand mean of the scores for that source, $a_i,\ a_j$ are i.i.d. random variables associated with their respective objects and assumed to be distributed $N\left(0,\ \sigma_a^2\right)$, and ε_{ij} is assumed distributed $N\left(0,\ \sigma_e^2\right)$. Note that while we assume that the scores are normally distributed, we make no assumptions on the distribution of the raw VSP data. The expected value and variances for the score in the model are as follows:

$$\begin{aligned}
E\left(s_{ij}\right) &= \theta \\
Var\left(s_{ij}\right) &= Var\left(\theta + a_i + a_j + e_{ij}\right) \\
&= 2\sigma_a^2 + \sigma_e^2 \\
Cov\left(s_{ij},\ s_{kj}\right) &= Cov\left(s_{ij},\ s_{il}\right) \\
&= E\left(\left(\theta + a_i + a_j + e_{ij} - \theta\right)\left(\theta + a_k + a_j + e_{kj} - \theta\right)\right) \\
&= E\left(a_j^2\right) \\
&= \sigma_a^2 \\
Cov\left(s_{ij},\ s_{kl}\right) &= E\left(\left(\theta + a_i + a_j + e_{ij} - \theta\right)\left(\theta + a_k + a_l + e_{kl} - \theta\right)\right) \\
&= 0
\end{aligned}$$

We are interested in the joint distribution of $\mathbf{s}_n = \left(s_{1,2},\ s_{1,3},\ ...,\ s_{(n-1),n}\right)$, a vector of $N = \binom{n}{2}$ scores resulting from the comparison of n objects from a given source. For a set of n objects, the linear model in equation 1 can be rewritten as:

$$\mathbf{s}_n = \theta\mathbf{1}_N + \mathbf{P}\mathbf{a} + \mathbf{e}. \tag{2}$$

where $\mathbf{a}$ is the vector of the a_i, $\mathbf{1}_N$ is a vector of 1s of length N, $\mathbf{e}$ the vector of e_{ij} and $\mathbf{P}$ is the following design matrix:

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_{12}^t \\ \mathbf{p}_{13}^t \\ \vdots \\ \mathbf{p}_{(n-1)n}^t \end{pmatrix}$$

Each $\mathbf{p}_{ij}$ is a vector of 0's and 1's, with 1's in the i^{th} and j^{th} locations, corresponding to the comparison represented by s_{ij}. For example, if $n = 4$ the resulting $\mathbf{P}$ is

$$\mathbf{P}_{N\times n} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

The expected value and covariance of $\mathbf{s}_n$ is

$$\begin{aligned} E(\mathbf{s}_n) &= E(\theta\mathbf{1}_N + \mathbf{Pa} + \mathbf{e}) \\ &= \theta\mathbf{1}_N \\ Cov(\mathbf{s}_n) &= Cov(\theta\mathbf{1}_N + \mathbf{Pa} + \mathbf{e}) \\ &= 0 + Cov(\mathbf{Pa}) + Cov(\mathbf{e}) \\ &= \mathbf{P}Cov(\mathbf{a})\mathbf{P}^t + \sigma_e^2\mathbf{I}_N \\ &= \mathbf{P}\mathbf{P}^t\sigma_a^2 + \sigma_e^2\mathbf{I}_N \\ &= \mathbf{\Sigma} \end{aligned} \tag{3}$$

where $\mathbf{I}_N$ is a $N \times N$ identity matrix and $\mathbf{PP}^t$ is a $N \times N$ matrix with the form (for $n = 4$):

$$\mathbf{P}\mathbf{P}^t = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

Under the assumption of normality the scores, $\mathbf{s}_n$, will be distributed as $MVN(\theta\mathbf{1}_N,\ \mathbf{\Sigma})$.

3.2. Estimation of the parameters for a given source

To estimate the 3 parameters, θ, σ_a^2, σ_e^2, of our model, we choose to use restricted maximum likelihood estimates based off of the log-likelihood function of $\mathbf{s}_n$:

$$-2\,\ell\left(\theta,\, \sigma_a^2,\, \sigma_e^2|\mathbf{s}_n\right) = ln\left(|\mathbf{\Sigma}|\right) + \left(\mathbf{s}_n - \theta\mathbf{1}_N\right)^t \mathbf{\Sigma}^{-1}\left(\mathbf{s}_n - \theta\mathbf{1}_N\right) + Nln\left(2\pi\right) \tag{4}$$

3.2.1. $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$

Assuming that all eigenvalue λ_k are non-zero, and their corresponding eigenvectors $\mathbf{v}_k$,

$$|\mathbf{\Sigma}| = \prod \lambda_k \tag{5}$$

$$\mathbf{\Sigma}^{-1} = \sum \lambda_k^{-1}\mathbf{v}_k\mathbf{v}_k^t \tag{6}$$

3.2.2. Eigenstructure equivalence of $\mathbf{PP}^t$ and $\mathbf{\Sigma}$

If two matrices $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvector, $\mathbf{v}$, and respective eigenvalues λ_A, λ_B then,

$$(\mathbf{A} + \mathbf{B})\,\mathbf{v} = \mathbf{A}\mathbf{v} + \mathbf{B}\mathbf{v} = \lambda_A\mathbf{v} + \lambda_B\mathbf{v} = (\lambda_A + \lambda_B)\,\mathbf{v} \tag{7}$$

Furthermore, any eigenvector, $\mathbf{v}_k$ of $\mathbf{PP}^t$ and its corresponding non-zero eigenvalue λ_k, $\mathbf{v}_k$ is also an eigenvector of $\mathbf{I}_N$. Thus, we have:

$$
\begin{aligned}
\boldsymbol{\Sigma}\mathbf{v}_k &= \left(\sigma_a^2\mathbf{P}\mathbf{P}^t + \sigma_e^2\mathbf{I}_N\right)\mathbf{v}_k \\
&= \sigma_a^2\mathbf{P}\mathbf{P}^t\mathbf{v}_k + \sigma_e^2\mathbf{I}_N\mathbf{v}_k \\
&= \sigma_a^2\lambda_k\mathbf{v}_k + \sigma_e^2\mathbf{v}_k \\
&= \left(\sigma_a^2\lambda_k + \sigma_e^2\right)\mathbf{v}_k
\end{aligned}
$$

Since $\left(\sigma_a^2\lambda_k + \sigma_e^2\right)$ is a constant, $\mathbf{v}_k$ is also an eigenvector of Σ. It is easier to study the eigenstructure of $\mathbf{P}\mathbf{P}^t$ using $\mathbf{P}^t\mathbf{P}$ to reduce the dimensionality of the problem;

$$
\begin{aligned}
\mathbf{P}\mathbf{P}^t\mathbf{v}_k &= \lambda_k\mathbf{v}_k \quad \Longleftrightarrow \\
\mathbf{P}^t\mathbf{P}\mathbf{P}^t\mathbf{v}_k &= \mathbf{P}^t\lambda_k\mathbf{v}_k \Longleftrightarrow \\
(\mathbf{P}^t\mathbf{P})\mathbf{P}^t\mathbf{v}_k &= \lambda_i\mathbf{P}^t\mathbf{v}_k \Longleftrightarrow \\
(\mathbf{P}^t\mathbf{P})\mathbf{v}_k' &= \lambda_k\mathbf{v}_k'
\end{aligned}
$$

We see that $\mathbf{P}\mathbf{P}^t$ has the same eigenvalues as $\mathbf{P}^t\mathbf{P}$, which implies that $\mathbf{P}\mathbf{P}^t$ has n non-zero eigenvalues and $N-n$ zero eigenvalues. The form of $\mathbf{P}^t\mathbf{P}$ is:

$$
\mathbf{P}^t\mathbf{P} = \begin{bmatrix} n-1 & 1 & \cdots & 1 \\ 1 & n-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & n-1 \end{bmatrix} = (n-2)\,\mathbf{I}_n + \mathbf{1}_n\mathbf{1}_n^t
$$

The matrix $\mathbf{1}_n\mathbf{1}_n^t$ has a single non-zero eigenvalue (or root) of value n with eigenvector $\mathbf{v}_1' = \mathbf{1}_n/\sqrt{n}$ and $n-1$ eigenvalues with value 0 and eigenvectors orthogonal to $\mathbf{1}_n$. Using the argument in equation 7, we see that $\mathbf{P}^t\mathbf{P}$ has one root equal to $2\,(n-1)$and $n-1$ roots equal to $(n-2)$.

This set of eigenvalues are the same as the first set of n eigenvalues for $\mathbf{P}\mathbf{P}^t$. Table 3 shows the build up of the eigenstructure for Σ starting with $\mathbf{P}^t\mathbf{P}$ on the right and working to the left.

Table 3: Eigenstructure of $\mathbf{\Sigma}$

$\mathbf{\Sigma} = \sigma_a^2 \mathbf{P}\mathbf{P}^t + \sigma_\varepsilon^2 \mathbf{I}_N$			$\mathbf{P}\mathbf{P}^t$			$\mathbf{P}^t\mathbf{P} = (n-2)\,\mathbf{I}_n + \mathbf{1}_n\mathbf{1}_n^t$		
Eigenvalue	# roots	Eigenvector	Eigenvalue	# roots	Eigenvector	Eigenvalue	# roots	Eigenvector
$2\,(n-1)\,\sigma_a^2 + \sigma_e^2$	1	$\mathbf{v}_1 = \frac{\mathbf{1}_N}{\sqrt{N}}$	$2\,(n-1)$	1	$\mathbf{v}_1 = \frac{\mathbf{1}_N}{\sqrt{N}}$	$2\,(n-1)$	1	$\mathbf{v}'_1 = \frac{\mathbf{1}_n}{\sqrt{n}}$
$(n-2)\,\sigma_a^2 + \sigma_e^2$	$n-1$	$\mathbf{v}_k$ s.t. $\mathbf{P}\mathbf{P}^t\mathbf{v}_k = (n-2)\,\mathbf{v}_k$	$(n-2)$	$n-1$	$\mathbf{v}_k$ s.t. $\mathbf{v}_k^t\mathbf{v}_{k'} = 0\ \forall\, k \neq k'$	$(n-2)$	$n-1$	$\mathbf{v}'_k$ s.t. $\mathbf{v}'^t_k\mathbf{v}'_{k'} = 0\ \forall\, k \neq k'$
σ_e^2	$N-n$	$\mathbf{v}_k$ is in the null space of $\mathbf{P}\mathbf{P}^t$	0	$N-n$	$\mathbf{v}_k$ s.t. $\mathbf{v}_k^t\mathbf{v}_{k'} = 0\ \forall\, k \neq k'$			

3.2.3. ANOVA table

Using table 3, we can rewrite equations 5 and 6 as:

$$\begin{aligned}
|\mathbf{\Sigma}| = \prod \lambda_k &= \left(2\left(n-1\right)\sigma_a^2+\sigma_e^2\right)\left(\left(n-2\right)\sigma_a^2+\sigma_e^2\right)^{(n-1)}\left(\sigma_e^2\right)^{N-n} \\
\mathbf{\Sigma}^{-1} = \sum \lambda_k^{-1}\mathbf{v}_k\mathbf{v}_k^t &= \frac{\mathbf{v}_1\mathbf{v}_1^t}{2\left(n-1\right)\sigma_a^2+\sigma_e^2}+\sum_{k=2}^{n}\frac{\mathbf{v}_k\mathbf{v}_k^t}{\left(n-2\right)\sigma_a^2+\sigma_e^2}+\sum_{k=n+1}^{N}\frac{\mathbf{v}_k\mathbf{v}_k^t}{\sigma_e^2} \\
&= \frac{\mathbf{v}_1\mathbf{v}_1^t}{\lambda_1}+\sum_{k=2}^{n}\frac{\mathbf{v}_k\mathbf{v}_k^t}{\lambda_2}+\sum_{k=n+1}^{N}\frac{\mathbf{v}_k\mathbf{v}_k^t}{\lambda_3},
\end{aligned}$$

where

$$\begin{aligned}
\lambda_1 &= 2\left(n-1\right)\sigma_a^2+\sigma_e^2 \\
\lambda_2 &= \left(n-2\right)\sigma_a^2+\sigma_e^2 \\
\lambda_3 &= \sigma_e^2,
\end{aligned}$$

and equation 4 as:

$$\begin{aligned}
-2\,\ell\left(\theta,\,\sigma_a^2,\,\sigma_e^2|\mathbf{s}_n\right) &= ln\left(|\mathbf{\Sigma}|\right)+\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)^t\mathbf{\Sigma}^{-1}\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)+Nln\left(2\pi\right) \\
&= ln\left(\lambda_1\right)+\left(n-1\right)ln\left(\lambda_2\right)+\left(N-n\right)ln\left(\lambda_3\right)+Nln\left(2\pi\right) \\
&+ \frac{\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)^t\mathbf{v}_1\mathbf{v}_1^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)}{\lambda_1} \\
&+ \sum_{k=2}^{n}\frac{\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)^t\mathbf{v}_k\mathbf{v}_k^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)}{\lambda_2} \\
&+ \sum_{k=n+1}^{N}\frac{\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)^t\mathbf{v}_k\mathbf{v}_k^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)}{\lambda_3}. \qquad (8)
\end{aligned}$$

We have some knowledge of the eigenvectors $\mathbf{v}_k$ that we can use to further simplify the last 3 terms of equation 8. We begin with the numerator of the first of these 3 terms, where $\mathbf{v}_1 = \mathbf{1}_N/\sqrt{N}$:

$$\begin{aligned}
\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)^t\mathbf{v}_1\mathbf{v}_1^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right) &= \left(\mathbf{v}_1^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)\right)^t\left(\mathbf{v}_1^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)\right) \\
&= \left(\mathbf{v}_1^t\left(\mathbf{s}_n-\theta\mathbf{1}_N\right)\right)^2 \\
&= \left(\mathbf{s}_n^t\mathbf{v}_1-\theta\mathbf{1}_N^t\mathbf{v}_1\right)^2 \\
&= \left(\frac{\mathbf{s}_n^t\mathbf{1}_N}{\sqrt{N}}-\frac{\theta\mathbf{1}_N^t\mathbf{1}_N}{\sqrt{N}}\right)^2 \\
&= \left(\frac{\mathbf{s}_n^t\mathbf{1}_N}{\sqrt{N}}\right)^2-\left(\frac{\mathbf{s}_n^t\mathbf{1}_N}{\sqrt{N}}\right)\left(\frac{\theta\mathbf{1}_N^t\mathbf{1}_N}{\sqrt{N}}\right)-\left(\frac{\theta\mathbf{1}_N^t\mathbf{1}_N}{\sqrt{N}}\right)\left(\frac{\mathbf{s}_n^t\mathbf{1}_N}{\sqrt{N}}\right)+\left(\frac{\theta\mathbf{1}_N^t\mathbf{1}_N}{\sqrt{N}}\right)^2 \\
&= \left(\frac{\sum_{i=1}^N s_i}{\sqrt{N}}\right)^2-\left(\frac{\sum_{i=1}^N s_i}{\sqrt{N}}\right)\left(\frac{N\theta}{\sqrt{N}}\right)-\left(\frac{N\theta}{\sqrt{N}}\right)\left(\frac{\sum_{i=1}^N s_i}{\sqrt{N}}\right)+\left(\frac{N\theta}{\sqrt{N}}\right)^2 \\
&= \frac{\left(\sum_{i=1}^N s_i\right)^2}{N}-2\theta\sum_{i=1}^{N}s_i+N\theta^2 \\
&= N\left(\frac{\left(\sum_{i=1}^N s_i\right)^2}{N^2}-\frac{2\theta\sum_{i=1}^N s_i}{N}+\theta^2\right) \\
&= N\left(\bar{s}^2-2\theta\bar{s}+\theta^2\right) \\
&= N\left(\bar{s}-\theta\right)^2 \qquad (9)
\end{aligned}$$

We now study the second term where $\mathbf{1}_N \mathbf{v}_k = 0$ for $k \geq 2$ (see table 3),

$$\sum_{k=2}^{n} (\mathbf{s}_n - \theta \mathbf{1}_N)^t \mathbf{v}_k \mathbf{v}_k^t (\mathbf{s}_n - \theta \mathbf{1}_N) = \mathbf{s}_n^t \left(\sum_{k=2}^{n} \mathbf{v}_k \mathbf{v}_k^t \right) \mathbf{s}_n = \sum_{k=2}^{n} \left(\mathbf{v}_k^t \mathbf{s}_n \right)^2 . \tag{10}$$

We are interested in the structure of $\sum_{k=2}^{n} \mathbf{v}_k \mathbf{v}_k^t$.

$$\begin{aligned}
\sum_{k=2}^{n} \lambda_k \mathbf{v}_k \mathbf{v}_k^t &= \mathbf{P}\mathbf{P}^t - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^t \\
\sum_{k=2}^{n} (n-2) \mathbf{v}_k \mathbf{v}_k^t &= \mathbf{P}\mathbf{P}^t - 2(n-1) \mathbf{v}_1 \mathbf{v}_1^t \\
\sum_{k=2}^{n} \mathbf{v}_k \mathbf{v}_k^t &= \frac{\mathbf{P}\mathbf{P}^t - 2(n-1) \mathbf{v}_1 \mathbf{v}_1^t}{n-2} \\
&= \frac{1}{n-2} \left(\mathbf{P}\mathbf{P}^t - \frac{2(n-1) \mathbf{1}_N \mathbf{1}_N^t}{N} \right) \\
&= \frac{1}{n-2} \left(\mathbf{P}\mathbf{P}^t - \frac{2(n-1) \mathbf{1}_N \mathbf{1}_N^t}{\frac{n(n-1)}{2}} \right) \\
&= \frac{(n-1)^2}{(n-1)^2} \frac{1}{n-2} \left(\mathbf{P}\mathbf{P}^t - \frac{4}{n} \mathbf{1}_N \mathbf{1}_N^t \right) \\
&= \frac{(n-1)^2}{n-2} \frac{1}{(n-1)^2} \left(\mathbf{P}\mathbf{P}^t - \frac{4}{n} \mathbf{1}_N \mathbf{1}_N^t \right) \\
&= \frac{(n-1)^2}{n-2} \left(\frac{\mathbf{P}\mathbf{P}^t}{(n-1)^2} - \frac{4}{n(n-1)^2} \mathbf{1}_N \mathbf{1}_N^t \right) \\
&= \frac{(n-1)^2}{n-2} \left(\frac{\mathbf{P}\mathbf{P}^t}{(n-1)^2} + \left(\frac{4n}{n^2(n-1)^2} - \frac{8}{n(n-1)^2} \right) \mathbf{1}_N \mathbf{1}_N^t \right) \\
&= \frac{(n-1)^2}{n-2} \left(\frac{1}{(n-1)^2} \mathbf{P}\mathbf{P}^t - \frac{4}{N(n-1)} \mathbf{1}_N \mathbf{1}_N^t + \frac{n}{N^2} \mathbf{1}_N \mathbf{1}_N^t \right) \\
&= \frac{(n-1)^2}{n-2} \left(\frac{1}{n-1} \mathbf{P} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_n^t \right) \left(\frac{1}{n-1} \mathbf{P}^t - \frac{1}{N} \mathbf{1}_n \mathbf{1}_N^t \right) .
\end{aligned} \tag{11}$$

Replacing equation 11 back into 10, we get

$$\mathbf{s}_n^t \left(\sum_{k=2}^{n} \mathbf{v}_k \mathbf{v}_k^t \right) \mathbf{s}_n = \frac{(n-1)^2}{n-2} \mathbf{s}_n^t \left(\frac{1}{n-1} \mathbf{P} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_n^t \right) \left(\frac{1}{n-1} \mathbf{P}^t - \frac{1}{N} \mathbf{1}_n \mathbf{1}_N^t \right) \mathbf{s}_n . \tag{12}$$

We note that

$$\frac{1}{n-1} \mathbf{P}^t \mathbf{s}_n = \begin{bmatrix} \bar{s}^{(1)} \\ \bar{s}^{(2)} \\ \vdots \\ \bar{s}^{(n)} \end{bmatrix}$$

is the vector of column averages where $\bar{s}^{(k)} = \frac{1}{n-1} \sum_{i \, or \, j = k} s_{ij}$, $k = 1, 2, ..., n$ and that

$$\frac{1}{N} \mathbf{1}_N^t \mathbf{s}_n = \bar{s}$$

This results in

$$\frac{1}{n-1}\mathbf{P}^t\mathbf{s}_n - \frac{1}{N}\mathbf{1}_n\mathbf{1}_N^t\mathbf{s}_n = \begin{bmatrix} \bar{s}^{(1)} - \bar{s} \\ \bar{s}^{(2)} - \bar{s} \\ \vdots \\ \bar{s}^{(n)} - \bar{s} \end{bmatrix}.$$

Substituting this result into equation 12, we obtain

$$\mathbf{s}_n^t\left(\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k^t\right)\mathbf{s}_n = \sum_{k=2}^{n}\left(\mathbf{v}_k^t\mathbf{s}_n\right)^2 = \frac{(n-1)^2}{n-2}\sum_{k=1}^{n}\left(\bar{s}^{(k)} - \bar{s}\right)^2 = SS_a, \tag{13}$$

which is the sum of squares distances between the overall average score for the source and the average pairwise scores between each object, taken in turn and kept fix, and all the other ones. We denote this sum of squares, SS_a.

Considering that $SS_t = SS_a + SS_e$, we can use SS_t to find SS_e. The corrected SS_t can be calculated by $SS_t = \mathbf{s}_n^t\left(\mathbf{I}_N - \mathbf{v}_1\mathbf{v}_1^t\right)\mathbf{s}_n$. Using the spectral decomposition of $\mathbf{I}_N$, we have:

$$\mathbf{I}_N = \mathbf{v}_1\mathbf{v}_1^t + \sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k^t + \sum_{k=n+1}^{N}\mathbf{v}_k\mathbf{v}_k^t \Longleftrightarrow$$

$$\mathbf{I}_N - \mathbf{v}_1\mathbf{v}_1^t = \sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k^t + \sum_{k=n+1}^{N}\mathbf{v}_k\mathbf{v}_k^t$$

Therefore,

$$SS_t = \mathbf{s}_n^t\left(\mathbf{I}_N - \mathbf{v}_1\mathbf{v}_1^t\right)\mathbf{s}_n = \mathbf{s}_n^t\left(\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k^t\right)\mathbf{s}_n + \mathbf{s}_n^t\left(\sum_{k=n+1}^{N}\mathbf{v}_k\mathbf{v}_k^t\right)\mathbf{s}_n, \tag{14}$$

which gives us that $SS_e = \mathbf{s}_n^t\left(\sum_{k=n+1}^{N}\mathbf{v}_k\mathbf{v}_k^t\right)\mathbf{s}_n$.

Rewriting the log-likelihood in equation 8, we arrive at:

$$\begin{aligned} -2\,\ell\left(\theta,\,\sigma_a^2,\,\sigma_e^2|\mathbf{s}\right) &= ln\left(\lambda_1\right) + (n-1)\,ln\left(\lambda_2\right) + (N-n)\,ln\left(\lambda_3\right) + Nln\left(2\pi\right) \\ &+ \frac{N\left(\bar{s}-\theta\right)^2}{\lambda_1} + \frac{SS_a}{\lambda_2} + \frac{SS_e}{\lambda_3}. \end{aligned}$$

By Cochran's Theorem [9], SS_a and SS_e are independent with degrees of freedom $n-1$ and $N-n$, respectively. Each sums of squares is independent of the sample mean, giving unbiased estimators of σ_a^2 and σ_e^2. In order to estimate σ_a^2 and σ_e^2, we calculate the expected mean sum of squares for SS_a:

$$\begin{aligned} SS_a &= \sum_{k=2}^{n}\left(\mathbf{v}_k^t\mathbf{s}_n\right)^2 \\ E\left(SS_a\right) &= \sum_{k=2}^{n}E\left[\left(\mathbf{v}_k^t\mathbf{s}_n\right)^2\right] \\ &= \sum_{k=2}^{n}Var\left(\mathbf{v}_k^t\mathbf{s}_n\right) \\ &= (n-1)\left((n-2)\,\sigma_a^2 + \sigma_e^2\right) \\ E\left(MS_a\right) &= \frac{(n-1)\left((n-2)\,\sigma_a^2 + \sigma_e^2\right)}{(n-1)} \\ &= (n-2)\,\sigma_a^2 + \sigma_e^2. \end{aligned}$$

using that the distribution of $\mathbf{v}_k^t\mathbf{s}$, for $k = 2, ..., n$, is:

$$
\begin{aligned}
\mathbf{s}_n &\sim MVN\left(\theta \mathbf{1}_N,\ \boldsymbol{\Sigma}\right) && \Longrightarrow \\
\mathbf{v}_k^t \mathbf{s}_n &\sim N\left(\theta \mathbf{v}_k^t \mathbf{1}_N,\ \mathbf{v}_k^t \boldsymbol{\Sigma} \mathbf{v}_k\right) && \Longleftrightarrow \\
\mathbf{v}_k^t \mathbf{s}_n &\sim N\left(0,\ \mathbf{v}_k^t \left(\mathbf{P}\mathbf{P}^t \sigma_a^2\right) \mathbf{v}_k + \mathbf{v}_k^t \mathbf{v}_k \sigma_e^2\right) && \Longleftrightarrow \\
\mathbf{v}_k^t \mathbf{s}_n &\sim N\left(0,\ \sigma_a^2 \mathbf{v}_k^t \mathbf{P}\mathbf{P}^t \mathbf{v}_k + \sigma_e^2\right) && \Longleftrightarrow \\
\mathbf{v}_k^t \mathbf{s}_n &\sim N\left(0,\ \lambda_k \sigma_a^2 + \sigma_e^2\right)
\end{aligned}
$$

where λ_k is the k^{th} root and is equal to $n-2$ (table 3). $E\left(MS_e\right)$ is calculated in the same way with $\mathbf{v}_k^t \mathbf{s} \sim N\left(0,\ \sigma_e^2\right)$ since $\lambda_k = 0$ for $k = n+1, ...,\ N$ (table 3). Thus,

$$E\left(MS_e\right) = \sigma_e^2.$$

These results are organized in the ANOVA table 4

Table 4: ANOVA table for parametric model

Source	df	SS	MS	E(MS)
A	$n-1$	SS_a	$\frac{SS_a}{(n-1)}$	$(n-2)\,\sigma_a^2 + \sigma_e^2$
Error	$N-n$	SS_e	$\frac{SS_e}{(N-n)}$	σ_e^2
Total	$N-1$	SS_t	$\frac{SS_t}{(N-1)}$	

Using the hat to denote the estimates of the parameters $\theta,\ \sigma_a^2,\ \sigma_e^2$ we get

$$
\begin{aligned}
\hat{\theta} &= \bar{s} \\
\hat{\sigma}_a^2 &= \frac{MS_a - MS_e}{n-2} \\
\hat{\sigma}_e^2 &= MS_e
\end{aligned}
$$

which are closely related to REML (restricted maximum likelihood) estimates and can easily be estimated using the results in equations 13 and 14. If the above formulas yield a negative result for $\hat{\sigma}_a^2$, the new results are that $\hat{\sigma}_a^2 = 0$ and $\hat{\sigma}_e^2 = MS_t$.

3.3. Assigning a probability density to a trace

Once we have estimated the parameter for the joint distribution of the pairwise scores between all objects from a given source,$\mathbf{s}_n$, we are interested in assigning the density of a new set of scores, $\mathbf{s}_m$, representing the pairwise comparisons between the trace objects, and the trace and source objects.

For example, if we have 2 objects from the trace, denoted as x_1 and x_2, and 4 objects from the reference source denoted as x_3 to x_6, $\mathbf{s}_m$ and $\mathbf{s}_n$ would take the following forms:

$$
\mathbf{s}_m = \begin{pmatrix} s_{12} \\ s_{13} \\ \vdots \\ s_{16} \\ s_{23} \\ \vdots \\ s_{26} \end{pmatrix},\ \mathbf{s}_n = \begin{pmatrix} s_{34} \\ s_{35} \\ s_{36} \\ s_{45} \\ s_{46} \\ s_{56} \end{pmatrix}.
$$

In effect, the probability density of $\mathbf{s}_m$ is the conditional probability density $f\left(\mathbf{s}_m | \mathbf{s}_n\right)$ which requires the joint probability density function $f\left(\mathbf{s}_m,\ \mathbf{s}_n\right)$. The joint probability of $\mathbf{s}_m$ and $\mathbf{s}_n$, when the trace and source objects are assumed to arise from the same source, simply requires the same design matrix $\mathbf{P}$ presented in

section 3.1 with size $\binom{m+n}{2} \times (m+n)$ where all the scores involving a trace object are stacked together. The covariance matrix of the stacked vector $\mathbf{s} = \begin{pmatrix} \mathbf{s}_m \\ \mathbf{s}_n \end{pmatrix}$ is a block matrix with the form:

$$\mathbf{\Sigma}_{\boldsymbol{s}} = \begin{bmatrix} \mathbf{\Sigma}_{s_m s_m} & \mathbf{\Sigma}_{s_m s_n} \\ \mathbf{\Sigma}_{s_n s_m} & \mathbf{\Sigma}_{s_n s_n} \end{bmatrix}$$

where each block corresponds to a section of the covariance matrix in equation 3 with the design matrix for the stacked vector $\mathbf{s}$. Continuing our example, $\mathbf{\Sigma}_{\boldsymbol{s}}$ would take the following form with $\sigma_c^2 = 2\sigma_a^2 + \sigma_e^2$:

$\mathbf{\Sigma}_{\boldsymbol{s}} =$

	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{23}	s_{24}	s_{25}	s_{26}	s_{34}	s_{35}	s_{36}	s_{45}	s_{46}	s_{56}
s_{12}	σ_c^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	0	0	0	0	0	0
s_{13}	σ_a^2	σ_c^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	0	0	0	σ_a^2	σ_a^2	σ_a^2	0	0	0
s_{14}	σ_a^2	σ_a^2	σ_c^2	σ_a^2	σ_a^2	0	σ_a^2	0	0	σ_a^2	0	0	σ_a^2	σ_a^2	0
s_{15}	σ_a^2	σ_a^2	σ_a^2	σ_c^2	σ_a^2	0	0	σ_a^2	0	0	σ_a^2	0	σ_a^2	0	σ_a^2
s_{16}	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_c^2	0	0	0	σ_a^2	0	0	σ_a^2	0	σ_a^2	σ_a^2
s_{23}	σ_a^2	σ_a^2	0	0	0	σ_c^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	0	0	0
s_{24}	σ_a^2	0	σ_a^2	0	0	σ_a^2	σ_c^2	σ_a^2	σ_a^2	σ_a^2	0	0	σ_a^2	σ_a^2	0
s_{25}	σ_a^2	0	0	σ_a^2	0	σ_a^2	σ_a^2	σ_c^2	σ_a^2	0	σ_a^2	0	σ_a^2	0	σ_a^2
s_{26}	σ_a^2	0	0	0	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_c^2	0	0	σ_a^2	0	σ_a^2	σ_a^2
s_{34}	0	σ_a^2	σ_a^2	0	0	σ_a^2	σ_a^2	0	0	σ_c^2	σ_a^2	σ_a^2	σ_a^2	σ_a^2	0
s_{35}	0	σ_a^2	0	σ_a^2	0	σ_a^2	0	σ_a^2	0	σ_a^2	σ_c^2	σ_a^2	σ_a^2	0	σ_a^2
s_{36}	0	σ_a^2	0	0	σ_a^2	σ_a^2	0	0	σ_a^2	σ_a^2	σ_a^2	σ_c^2	0	σ_a^2	σ_a^2
s_{45}	0	0	σ_a^2	σ_a^2	0	0	σ_a^2	σ_a^2	0	σ_a^2	σ_a^2	0	σ_c^2	σ_a^2	σ_a^2
s_{46}	0	0	σ_a^2	0	σ_a^2	0	σ_a^2	0	σ_a^2	σ_a^2	0	σ_a^2	σ_a^2	σ_c^2	σ_a^2
s_{56}	0	0	0	σ_a^2	σ_a^2	0	0	σ_a^2	σ_a^2	0	σ_a^2	σ_a^2	σ_a^2	σ_a^2	σ_c^2

where θ, σ_a^2, σ_e^2 are estimated based on the sole vector $\mathbf{s}_n$. Obtaining $f(\mathbf{s}_m|\mathbf{s}_n)$ using $\mathbf{\Sigma}_{\mathbf{s}}$is trivial [2, 6, 10].

3.4. Choice of the kernel for VSPs

Our problem is as follows: given a finite population of potential reference sources, each defined by multiple sets of VSPs, we want to build an inferential model that will provide information on the source of a 'trace' set of VSPs. In this context, we assume that, given a source, VSP sets are identically independently distributed. Our parametric model enables us to characterize the probability density function of the pairwise scores between multiple sets of VSPs representing a given source and to estimate its 3 parameters, providing that we can define a kernel that satisfies the normality assumption of the scores[4]. The kernel defined below was chosen after initial work showed that well known kernels [5] did not meet assumptions of the model and did not correctly classify more than 50% of the sources in the validation dataset.

The comparison of two VSP sets involves comparing sets that do not necessarily contain an equivalent number of particles. Therefore the first step in defining a kernel to measure their simlarity requires us to map the sets of VSPs into a common space.

Let $\mathbf{X}_i^l$ and $\mathbf{X}_j^l$ be two $p_i \times 18$ and $p_j \times 18$ matrices representing two sets of VSPs from reference source l in $1, ..., 20$, with p_i and p_j particles respectively, and 18 chemical elemental relative concentrations. Let $\mathbf{x}_{ir}^l$ be the set of measurements for the r^{th} $(r = 1, 2, ..., 18)$ chemical element across all particles on the i^{th} $(i = 1, ..., n_l)$ VSP set of source l. For a pair of VSP sets, our kernel compares the empirical cummulative distribution functions of the relative concentration of each element taken in turn and then aggregates the elementwise scores to obtain s_{ij} for that pair of VSP sets.

We define our kernel κ as the function $\kappa\left(\mathbf{F}_{\mathbf{Y}_i^l}, \mathbf{F}_{\mathbf{Y}_j^l}\right)$ where $\mathbf{F}_{\mathbf{Y}_i^l}$, $\mathbf{F}_{\mathbf{Y}_j^l}$ are the sets of empirical cummulative distribution functions (ECDF) based off of $\mathbf{Y}_i^l$ and $\mathbf{Y}_j^l$, where

[4]The model proposed in this paper is based on the pairwise comparisons between sets of VSPs. It could be tempting to build a model that directly uses pairwise comparisons of the particles themselves. However, we explained above that it is not so much the composition of the particles that contains discriminative information but is the relative proportions of the different particle types found in a given source. It is therefore desireable to represent each object from a given source by arrays of VSPs. This renders difficult the use of classic machine learning methods in this context.

$$\mathbf{Y}_i^l = \mathbf{A}\mathbf{X}_i^l.$$

$\mathbf{A}$ is a matrix of bases that can be suitably chosen to weight the original data or reduce the dimension of the number of measurements taken on each particle. The outside operator of $\kappa\left(\mathbf{F}_{\mathbf{Y}_i^l}, \mathbf{F}_{\mathbf{Y}_j^l}\right)$ is:

$$\kappa\left(\mathbf{F}_{\mathbf{Y}_i^l}, \mathbf{F}_{\mathbf{Y}_j^l}\right) = \sum_{r=1}^{18} log\left(\int \left(F_{\mathbf{y}_{ir}^l}(t) - F_{\mathbf{y}_{jr}^l}(t)\right)^2 dt\right)$$

where $\mathbf{y}_{ir}^l$ is the r^{th} $(r = 1, 2, ..., 18)$ column of $\mathbf{Y}_i^l$ corresponding to the transformed measurements of the r^{th} chemical element in matrix $\mathbf{X}_i^l$. In practice we do not have an analytical solution for the integral and we approximate it numerically.

In our example, we would obtain the following vector of scores related to the 4 sets of VSPs from the reference source l and the two sets of VSPs from the trace:

$$\mathbf{s}_n = \begin{pmatrix} \kappa\left(\mathbf{F}_{\mathbf{Y}_3^l}, \mathbf{F}_{\mathbf{Y}_4^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_3^l}, \mathbf{F}_{\mathbf{Y}_5^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_3^l}, \mathbf{F}_{\mathbf{Y}_6^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_4^l}, \mathbf{F}_{\mathbf{Y}_5^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_4^l}, \mathbf{F}_{\mathbf{Y}_6^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_5^l}, \mathbf{F}_{\mathbf{Y}_6^l}\right) \end{pmatrix}, \ \mathbf{s}_m = \begin{pmatrix} \kappa\left(\mathbf{F}_{\mathbf{Y}_1^l}, \mathbf{F}_{\mathbf{Y}_2^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_1^l}, \mathbf{F}_{\mathbf{Y}_3^l}\right) \\ \vdots \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_1^l}, \mathbf{F}_{\mathbf{Y}_6^l}\right) \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_2^l}, \mathbf{F}_{\mathbf{Y}_3^l}\right) \\ \vdots \\ \kappa\left(\mathbf{F}_{\mathbf{Y}_2^l}, \mathbf{F}_{\mathbf{Y}_6^l}\right) \end{pmatrix}.$$

4. Implementation of the model for VSPs

4.1. Basis selection

To obtain the following results, we use a basis matrix $\mathbf{A}$ that consisted in the 12 first eigenvectors from the spectral decomposition of the covariance of all 3 datasets of particles pooled together[5]. We show the results of the separation between the different sources when they are projected using the two first eigenvectors in figures 3 and 4. In both figures we highlighted a specific source (in blue) and its corresponding trace from the validation dataset (in red). Figure 3 illustrates a situation where the ECDFs of the trace and its corresponding source are very similar, while figure 4 illustrate a situation where they are not. Situations such as that will result in the correct source being harder to infer, if at all possible (see table 6 below).

We also show in figure 5 that some traces (here trace from source '58') can have very similar ECDFs to other sources (here source '47') than their true source.

[5]Note that the identity matrix can be used to work directly with the raw observations.

Figure 3: PCA Score ECDFs for all sources (first two components) with the ECDF for source '24' highlighted in blue and the trace from source '24' added in red. We can observe that the trace and source ECDFs are very similar to each other. This results in our system being able to correctly associate them (table 6)

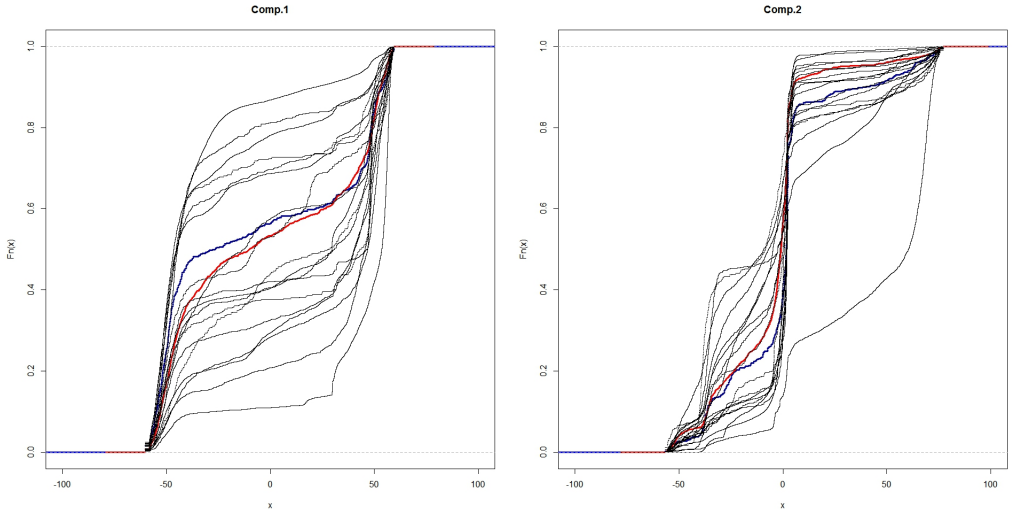


Figure 4: PCA Score ECDFs for all sources (first two components) with the ECDF for source '32' highlighted in blue and the trace from source '32' added in red. We can observe that the trace and source ECDFs are not as similar as one would expect given their common origin. Our system was not able to correctly associate them (table 6).

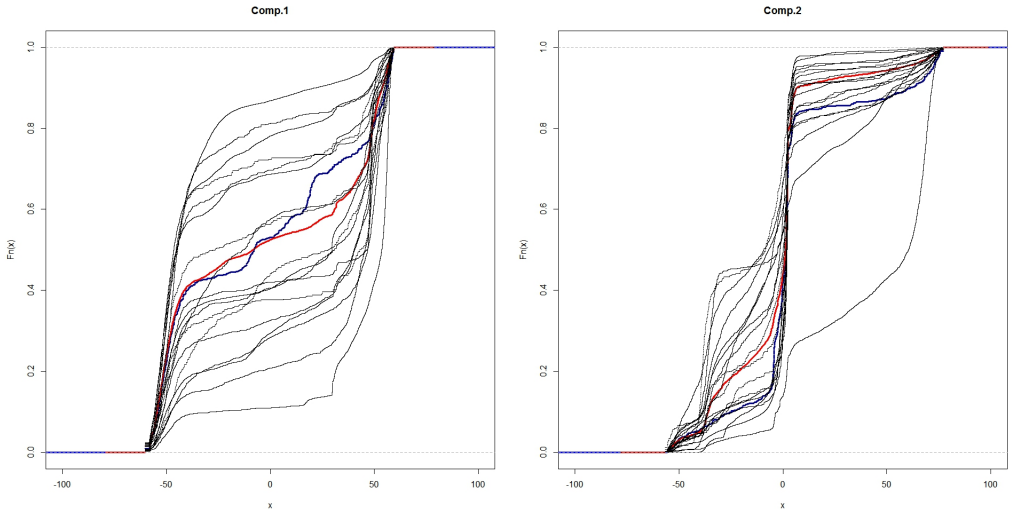
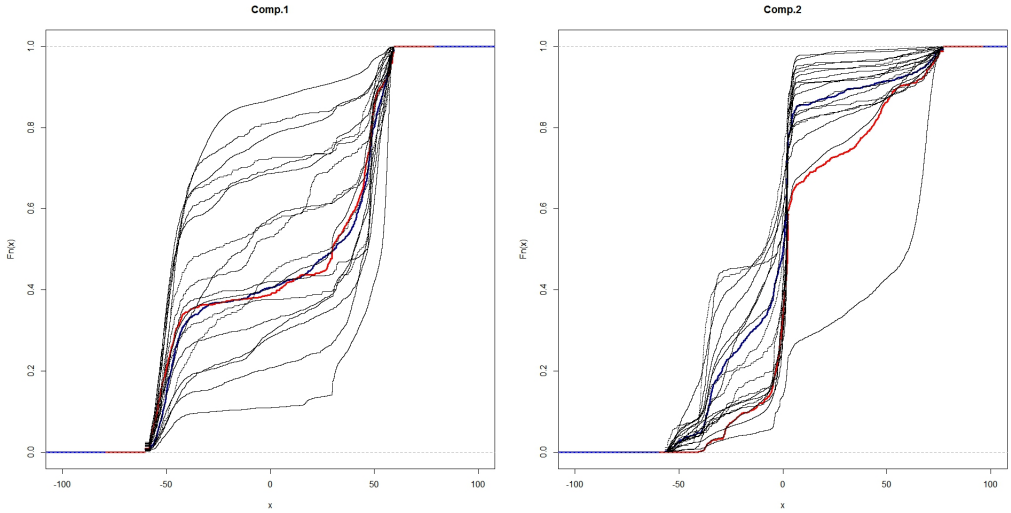


Figure 5: PCA Score ECDFs for all sources (first two components) with the ECDF for source '47' highlighted in blue and the trace from source '58' added in red. We can observe that the trace and source ECDFs are much more similar than one would expect given their different origins. Our system incorrectly associated them (table 6).



4.2. Generation of multiple pseudo-sets of VSPs

Because the particles, for each source, were pooled together during their analysis, we were unable to use multiple sets of VSPs per source to estimate the parameters of $f(\mathbf{s}_n)$ as explained above. To estimate the parameters for the distribution of a particular source, a non-parametric bootstrap was used to create pseudo-sets of VSPs. Using the bootstrap allows for obtaining multiple ECDFs for a given source and a given trace in order to create $\mathbf{s}_n$ and $\mathbf{s}_m$. In this project, we created 30 pseudo-sets of 300 VSPs for each source and 2 pseudo-sets of 500 VSPs for each trace.

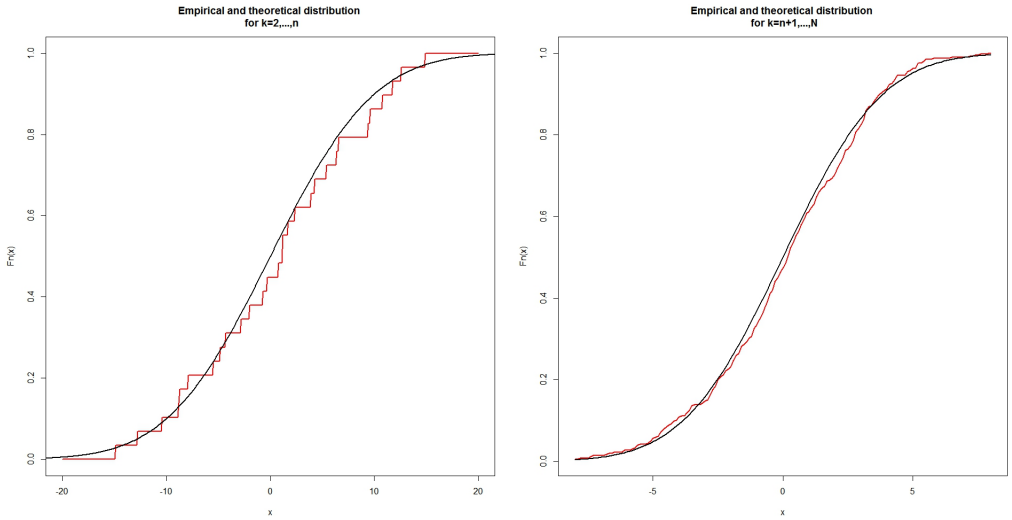
4.3. Estimation of parameters for source '24'

The calculations for source 24 are presented below. The calculations for all other sources are similar. The full set of results can be found in the next section.

Since our model assumes a multivariate normal distribution for $\mathbf{s}_n$, we visually checked the marginal normality of $\mathbf{V}^t\mathbf{s}_n$ where $\mathbf{V}$ is a matrix of the eigenvectors $\mathbf{v}_k$ (figure 6). By construction, $\mathbf{V}^t_{2:n}\mathbf{S}_n \sim MVN\left(\mathbf{0},\, (n-2)\,\sigma_a^2\mathbf{I} + \sigma_e^2\mathbf{I}\right)$ and $\mathbf{V}^t_{(n+1):N}\mathbf{S}_n \sim MVN\left(\mathbf{0},\, \sigma_e^2\mathbf{I}\right)$. We also performed a Shapiro-Wilk normality test[6]. For source '24', the Shapiro-Wilk normality test resulted in a $p - value = 0.2959$ for $\mathbf{V}^t_{2:n}\mathbf{s}_n$ and $p - value = 0.1988$ for $\mathbf{V}^t_{(n+1):N}\mathbf{s}_n$.

[6]This test is not a consistent test for multivariate normality of the vector of scores. Nevertheless, in this context it suggests the assumption of multivariate normality is reasonable.

Figure 6: Empirical against theoretical cummulative distribution functions for marginal projections. The empirical distribution of $\mathbf{V}_{2:n}^t \mathbf{s}_n$ (left) and $\mathbf{V}_{(n+1):N}^t \mathbf{s}_n$ (right) is represented by the red line and their respective theoretical distribution as the black line.



The left plot in figure 6 shows the resulting approximate normal distribution of $k = 2, ..., n$ projections. The plot on the right shows the same but for the $k = n + 1, ..., N$ projections. The variance of the right distribution is approximately equal to σ_e^2 in the parametric model while the variance of the left distribution is approximately equal to $(n-2)\,\sigma_a^2 + \sigma_e^2$ 4. Both ECDFs are overlaid by the CDFs from the parametric model using the estimated parameters for source '24', which are:

$$\begin{aligned} \hat{\theta} &= -20.60895 \\ \hat{\sigma}_a^2 &= 3.850046 \\ \hat{\sigma}_e^2 &= 6.728073. \end{aligned}$$

These estimates are used in equations 2 and 3 in order to build the conditional distribution $f(\mathbf{s}_m|\mathbf{s}_n, \hat{\theta}\mathbf{1}_N, \mathbf{\Sigma})$ described in section 3.3. We evaluated the conditional distribution when the VSP sets from the trace and the reference source originate from the same location and when they were not (trace from location 24 and sources from the other locations).

The results for the comparison between traces from source '24' and sets of VSPs from all sources are given in tables 5 and 6), where the source with the highest probability density for each test set is bolded. The results show that the VSP set from location 24 in the testing dataset has been associated to its corresponding set in the training dataset (table 5 - row labelled T24). In addition, the 'trace' VSP set from location 24 in the validation dataset has also been correctly associated to its corresponding source (table 6 - row labelled T24).

Table 5: Log_{10} of the probability densities of pseudo-trace VSP sets (from the testing dataset) given the distribution of potential sources (from the training dataset). Highest probability densities in each row are bolded. 'True' sources are in the diagonal. The $-\infty$ terms are due to numerical zeroes arising when assigning probability densities of some traces in the distributions of some sources.

	R10	R12	R19	R23	R24	R31	R32	R44	R45	R47	R48	R49	R52	R53	R58	R59	R72	R77	R81	R89
T10	**-90.0**	-278.9	-218.7	-127.9	-170.7	-212.3	-228.0	-191.1	-254.0	-146.1	-181.3	-281.8	-267.2	-176.2	-274.2	-254.6	$-\infty$	-189.5	-197.4	-198.1
T12	-226.8	**-83.5**	-143.1	-256.7	-218.7	-145.2	-267.8	-255.4	$-\infty$	$-\infty$	-247.1	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-262.1	-222.7	-127.9	-298.4
T19	-185.8	-144.5	**-78.4**	-223.1	-179.3	-136.3	-231.4	-208.1	$-\infty$	-321.1	-200.0	$-\infty$	$-\infty$	$-\infty$	-318.6	$-\infty$	-269.2	-155.0	-143.0	-232.6
T23	-122.2	-281.9	-256.9	**-78.4**	-145.3	-210.3	-206.9	-235.7	-221.9	-137.5	-206.8	-251.6	-225.0	-154.1	-269.8	-224.5	$-\infty$	-241.6	-228.9	-189.5
T24	-143.6	-250.9	-202.4	-146.3	**-86.9**	-169.0	-216.8	-230.6	$-\infty$	-217.0	-201.6	$-\infty$	$-\infty$	-292.7	-299.6	-273.4	-303.1	-179.4	-190.2	-172.9
T31	-146.7	-156.7	-111.7	-179.9	-147.2	**-93.1**	-174.3	-231.1	$-\infty$	-226.3	-218.4	$-\infty$	$-\infty$	$-\infty$	-278.3	-263.8	-297.0	-120.6	-115.3	-204.1
T32	-252.8	$-\infty$	$-\infty$	-263.1	-309.8	$-\infty$	**-97.7**	$-\infty$	-271.1	-303.6	-301.8	$-\infty$	$-\infty$	$-\infty$	-267.5	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-311.0
T44	-173.8	$-\infty$	-273.8	-254.9	-251.6	-321.3	-307.4	**-82.4**	$-\infty$	$-\infty$	-107.5	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-295.4
T45	-140.0	$-\infty$	-314.6	-163.2	-262.8	$-\infty$	-155.6	-260.6	**-83.3**	-149.3	-223.8	-218.7	-216.6	-172.0	-240.8	$-\infty$	$-\infty$	-286.9	-322.3	-212.5
T47	-97.3	-272.8	-232.8	-106.6	-167.3	-208.0	-156.1	-231.1	-147.6	**-81.6**	-214.9	-262.0	-201.1	-131.5	-256.6	-235.3	$-\infty$	-193.1	-215.6	-143.9
T48	-210.6	$-\infty$	-322.2	-274.9	-281.2	$-\infty$	$-\infty$	-122.3	$-\infty$	$-\infty$	**-83.4**	-308.6	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
T49	-182.3	$-\infty$	-322.4	-170.0	-267.2	$-\infty$	-251.5	-235.4	-258.7	-287.0	-191.6	**-77.2**	-235.7	-284.2	-269.0	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-272.4
T52	-130.8	$-\infty$	-317.6	-157.0	-218.3	-304.9	-224.3	-248.3	-200.2	-162.0	-207.9	-182.6	**-78.8**	-119.5	-257.7	-320.9	$-\infty$	-280.2	$-\infty$	-250.9
T53	-98.3	-277.6	-247.6	-111.5	-172.5	-222.3	-199.9	-209.1	-185.7	-126.4	-180.0	-216.2	-123.3	**-73.5**	-263.5	-274.5	$-\infty$	-225.0	-258.4	-217.6
T58	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-300.5	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**-82.7**	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
T59	-202.5	$-\infty$	$-\infty$	-214.2	-269.4	-308.2	-248.3	-291.7	$-\infty$	-271.5	-257.1	$-\infty$	$-\infty$	$-\infty$	-246.9	**-75.7**	$-\infty$	-305.0	$-\infty$	-308.9
T72	-243.9	-203.0	-221.3	-311.9	-221.8	-220.8	-310.5	-274.9	$-\infty$	$-\infty$	-254.4	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**-90.9**	-211.3	-175.2	-302.7
T77	-139.2	-179.7	-124.4	-206.7	-143.2	-97.8	-196.9	-254.1	$-\infty$	-215.2	-239.4	$-\infty$	$-\infty$	-316.1	-298.1	-274.2	-224.8	**-97.2**	-103.0	-209.5
T81	-167.3	-117.1	-128.6	-200.8	-175.7	-101.7	-209.7	-254.9	$-\infty$	-264.5	-237.1	$-\infty$	$-\infty$	$-\infty$	-307.4	-292.8	-206.6	-118.2	**-80.7**	-222.8
T89	-187.3	$-\infty$	-286.2	-213.1	-226.2	-306.4	-260.2	-276.3	-318.2	-242.4	-258.9	$-\infty$	$-\infty$	$-\infty$	-295.9	$-\infty$	$-\infty$	-303.3	-290.8	**-88.7**

The strength of the association can easily be seen in figure 3 by the closeness of the red and blue lines.

4.4. Performance of the model as a classifier

Table 5 shows the full set of results obtained by comparing each pseudo-trace from the testing dataset to all sources in the training dataset. Our method has perfect accuracy. This is expected as the sets of VSPs in the testing dataset are very similar to the ones in the training dataset. Using the validation dataset, which is more representative of a real-world scenario, table 6 shows that the accuracy drops to 55% correctly classified traces. This is better than chance alone (5%) and is a slight improvement over Stoney et. al [13], who achieved a 50% correct classification rate with this dataset.

Table 6: Log_{10} of the probability densities of trace VSP sets (from the validation dataset) given the distribution of potential sources (from the training dataset). Highest probability densities in each row are bolded. 'True' sources are in the diagonal. The $-\infty$ terms are due to numerical zeroes arising when assigning probability densities of some traces in the distributions of some sources.

	R10	R12	R19	R23	R24	R31	R32	R44	R45	R47	R48	R49	R52	R53	R58	R59	R72	R77	R81	R89
T10	**-81.4**	-239.0	-184.9	-122.2	-165.6	-165.6	-162.2	-160.0	-144.8	-108.4	-198.2	-125.5	-107.6	-86.2	-182.9	-198.1	$-\infty$	-174.4	-238.4	-166.5
T12	-128.0	-124.6	-107.2	-112.2	-92.7	**-92.3**	-129.3	-166.8	-195.0	-124.0	-233.0	-167.8	-214.7	-137.1	-184.0	-175.3	-277.9	-100.0	-112.0	-133.0
T19	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**-270.5**	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
T23	-105.5	-200.8	-149.9	-90.8	-106.8	-124.8	-125.6	-176.5	-149.2	**-81.0**	-252.8	-163.9	-164.1	-102.7	-171.0	-153.6	$-\infty$	-122.0	-167.3	-131.5
T24	-119.8	-171.6	-140.2	-93.5	**-92.6**	-112.3	-129.4	-170.9	-167.4	-93.9	-242.9	-157.8	-163.3	-105.6	-178.5	-166.8	$-\infty$	-133.2	-170.3	-117.3
T31	-197.8	-94.7	-103.2	-156.8	-122.1	**-94.2**	-161.7	-168.5	-299.6	-239.1	-236.1	-221.8	-284.2	-208.5	-204.7	-197.7	-226.4	-125.7	-122.1	-189.7
T32	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**-289.5**	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
T44	-143.5	-153.4	-110.1	-152.4	**-109.9**	-118.3	-167.0	-127.5	-268.0	-204.3	-169.2	-211.2	-259.8	-181.7	-196.1	-188.1	-265.9	-122.7	-168.3	-153.5
T45	-284.3		-313.3	-239.5	-312.7	-302.7	-236.8	-267.7	-314.6	$-\infty$	$-\infty$	-209.7	-234.9	-239.3	**-179.1**	-279.9	$-\infty$	$-\infty$	$-\infty$	-310.3
T47	-116.3	-249.1	-208.2	-127.7	-181.2	-183.7	-156.4	-195.3	-116.8	-97.0	-243.9	-131.5	-92.9	**-81.3**	-174.8	-190.9	$-\infty$	-184.2	-261.7	-181.3
T48	-138.0	-237.0	-200.2	-132.9	-169.6	-175.4	-167.6	-193.6	-134.2	-128.9	-237.3	-149.0	-117.1	**-99.5**	-178.6	-203.6	$-\infty$	-191.9	-249.1	-179.6
T49	-137.5	-296.1	-205.7	-160.3	-204.8	-252.3	-215.8	-128.6	-240.3	-300.9	-145.7	**-113.0**	-228.5	-196.3	-204.5	-265.0	$-\infty$	-249.0	-310.0	-245.0
T52	-144.1	-302.3	-242.8	-166.2	-250.5	-243.4	-180.9	-211.6	-151.4	-152.2	-276.4	-136.1	**-80.8**	-105.6	-172.9	-222.1	$-\infty$	-225.5	$-\infty$	-221.2
T53	-111.6	-243.1	-207.0	-136.9	-175.1	-194.2	-174.2	-179.4	-141.0	-125.2	-217.2	-124.8	-94.0	**-91.5**	-183.3	-206.1	$-\infty$	-199.2	-272.7	-191.6
T58	-105.3	-192.9	-164.5	-112.4	-140.1	-135.9	-135.1	-173.6	-140.1	**-77.6**	-246.8	-171.5	-129.8	-91.7	-180.8	-160.7	$-\infty$	-132.6	-192.7	-137.3
T59	-222.3	-253.4	-215.2	-167.9	-202.6	-196.0	-190.4	-220.5	-303.6	-248.3	-289.2	-242.6	-265.0	-201.4	-183.7	**-99.0**	$-\infty$	-210.0	-273.2	-239.1
T72	-264.3	-170.5	-159.7	-260.3	-212.0	-168.1	-230.3	-220.8	$-\infty$	$-\infty$	$-\infty$	-322.5	$-\infty$	-318.1	-243.9	-244.2	**-92.5**	-159.0	-159.8	-235.2
T77	-156.9	-147.5	-111.9	-160.9	-121.9	-89.9	-146.0	-187.2	-267.2	-191.1	-265.3	-238.8	-261.4	-193.1	-195.7	-171.8	-200.4	**-80.0**	-121.5	-172.1
T81	-186.7	-109.8	-112.1	-160.3	-131.0	-93.2	-164.0	-188.6	-293.4	-233.0	-256.2	-225.8	-302.6	-219.0	-201.3	-196.3	-179.7	-105.3	**-89.3**	-175.4
T89	-208.8	-260.9	-206.9	-156.3	-214.2	-204.4	-180.8	-227.2	-224.8	-199.6	$-\infty$	-196.1	-256.3	-215.8	-192.5	-205.8	$-\infty$	-225.5	-275.4	**-91.6**

5. Discussion and Conclusions

The method proposed in this paper has many advantages over the methods currently used to compare complex chemical signals and infer the source of forensic traces: (1) it only requires to estimate 3 univariate parameters for each source (for a given choice of kernel); (2) it can be applied to a wide range of types of chemical signals (but also images) provided that a suitable kernel is used; and (3) it is designed to account for the dependencies between multiple samples from a same source, which is a situation that commonly arises in forensic science (e.g., glass fragments, plastic bags, illegal drugs, tapes or wires used to restrain victims, bullets and cartridge casings).

Our model relies on a very limited number of assumptions, which all directly arise from the design of the model in equation (1). More specifically, our model does not make any assumption on the type or distribution of the data in the raw chemical signals; however, it requires that the choice of the kernel is such that the resulting scores for each source are approximately normally distributed (although in practice, the model appears to be robust to the violation of this assumption). Kernel selection is not only important to meet the normality assumption discussed above, but also to ensure that variability between signals from objects from a common source is minimized, while variability between signals from objects from different sources is maximized. In our VSP example, the kernel measured distances between ECDFs for each of the considered chemical elements from pairs of VSP sets. However, there exist many different forms of kernels that can be considered [5, 8, 11]. Some possible improvements to our application of the model to VSPs may stem from the kernel or kernels used. For example, a correlation-based kernel could be used to measure the similarity of ECDF shapes and then leverage the closure properties of kernels to combine the multiple scores into a single similarity measure. In addition, kernels may have parameters themselves, which can be used to rescale or normalize the resulting scores. However, since our objective for this research paper was the development and implementation of the parametric score model and to illustrate its implementation, we did not optimize the choice, construction, or parameters of our kernel.

The implementation of our model as a classifier for the sets of VSPs was impeded by three main elements: (1) the sets of VSPs considered as 'trace' and 'reference' were not sampled in the same manner: the 'trace' VSP sets were obtained by brushing a carpet area, while the 'source' VSP sets were obtained by plucking fibers. It may be that the resulting differences in the particle profiles of these two datasets originate from a difference between VSPs found at the surface of carpets and VSPs found deeper into the carpets. Furthermore, the raw fibers were originally collected by crime scene technicians in [13]. Despite their instructions, they may not have collected them in the same area of the carpets. (2) Our method requires multiple samples to characterize the within-variability of each source, however, all fibers collected in [13] were pooled together (for each location) before being washed to extract the VSPs. We had to resort to non-parametric bootstrap to create multiple pseudo-sets of VSPs for each source. When calculating our results, it appeared that our method was sensitive to the amount of VSPs used to create the pseudo-sets. (3) Some locations were characterized by many thousands of particles, while some were only characterized by a few hundreds. Previous research has shown that the inference of the source of an object based on VSPs is very sensitive to the number of particles collected from the suspected sources [13]. All of these elements can explain the lack of performance of our method when used to infer the source of the 'trace' VSP sets. Further research is needed to investigate the multiple source of variability affecting the presence of VSPs on carpets (e.g., within a single location on a carpet, between multiple locations on a carpet, between multiple carpets in a house), or on objects in general. Similarly, further research is needed to investigate the effect of the number and size of VSP sets on the performance of our method.

Currently, our model can only provide the likelihood of observing one or more trace objects given a specific source. This allows us to use it as a classifier in a closed-set context or to test hypotheses such as 'same source' vs. 'different sources' in a frequentist context. We plan to expand our model in order to be able to assign the likelihood of observing one or more trace objects by chance in a population of potential sources, with the ultimate goal of being able to quantify the weight of forensic evidence using a formal Bayesian paradigm. In any case, we believe that our model allows for leveraging the versatility, and dimension reduction and discriminative powers of kernels, while it retains the dependency structure between scores calculated between multiple objects originating from a single source.

6. Acknowledgement

This research presented was supported by the National Institute of Justice, Office of Justice Programs, US Department of Justice under Awards No. 2015-R2-CX-0028, 2014-IJ-CX-K088, 2012-DN-BX-K041, 2009-DN-BX-K234.

[1] I. Alberink, A. Jongh, and C. Rodriguez. Fingermark evidence evaluation based on automated fingerprint identification system matching scores: The effect of different types of conditioning on likelihood ratios. *Journal of Forensic Sciences*, 59(1):70–81, 2014.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3 edition, 2003.

[3] D. Gantz and C. P. Saunders. Quantifying the effects of database size and sample quality on measures of individualization validity and accuracy in forensics. Final Grant Report 248670, National Institute of Justice, March 2015.

[4] A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia. Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219(1):129–140, 2012.

[5] Thomas Hofmann, Bernhard Scholkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[6] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, 2008.

[7] A. J. Lee. *U-Statistics Theorey and Practice*. Marcel Dekker Inc., 1990.

[8] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[9] Scheffé. *The Analysis of Variance*. John Wiley and Sons, 1959.

[10] George A. F. Seber. *Multivariate Observations*. Wiley, 2004.

[11] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 6 edition, 2012.

[12] David A. Stoney, Paul L Stoney, and Cedric Neumann. Exploitation of very small particles to enhance the probative value of carpet fibers. Technical Report 248904, U.S. Department of Justice, 2012.

[13] David A. Stoney, Cedric Neumann, Kim E. Mooney, J. Matney Wyatt, and Paul L Stoney. Exploitation of very small particles to enhance the probative value of carpet fibers. *Forensic Science International*, 252:52–68, 2015.

[14] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2007.