Parametric modelling of interpoint distance distributions, with an application to a mixture model for biosurveillance data

Marco Bonetti[1], Karen L. Olson[2], Kenneth D. Mandl[2], Marcello Pagano[3]

[1] Department of Decision Sciences, Bocconi University, Milan, Italy
[2] Children's Hospital Informatics Program, Children's Hospital Boston, Boston, MA, USA
[3] Department of Biostatistics, Harvard School of Public Health, MA, USA

Corresponding Author:
Marco Bonetti, Dept. of Decision Sciences, Bocconi University
Via Roentgen 1, 20136 Milano, Italy
E-mail: marco.bonetti@unibocconi.it

Summary

Objectives. The interpoint distance distribution can be used to analyze data consisting of inter-observation distances, i.e. all the pairwise distances arising from a random sample of *n* multivariate observations. Methods for the study of such distributions exist in the literature with applications to genetics, disease clustering, and biosurveillance problems. So far, techniques have been limited to nonparametric analyses. Here we illustrate how one can expand this set of tools to the use of parametric models.

Methods and Results. We assume a parametric model $f_D(d;\theta)$ for the random variable $D = d(X_1,X_2)$ where d(·) is a dissimilarity measure and X_1, X_2 two i.i.d. observations from a multivariate distribution. We describe the properties of a proposed estimator for θ ($\in\Re^k$), noting in particular its asymptotic normality. We compare the proposed estimator with two alternative estimators, both in general and within an analytically tractable case. We discuss the implementation of the methods to the construction of a parametric mixture model, and illustrate the use of that model for a preliminary analysis of data arising from a biosurveillance system.

Conclusions. Parametric models for interpoint distance distributions can be a valuable tool for the analysis of multivariate data ranging from geographic coordinates to highly dimensional vectors.

KEY WORDS: *syndromic surveillance, U-statistic, estimating equation, interpoint distance distribution*.

Introduction

Consider the positive random variable $D = d(X_1, X_2)$ obtained as the (Euclidean, say) distance between two independent and identically distributed vectors X_1 and X_2 arising from the distribution $F_X(x)$, $x\in\Re^p$. In what follows we use the term distance for *D*, but the discussion applies to any symmetric (non-negative) function of the two arguments X_1 and X_2, and in particular to any dissimilarity measure that may be relevant for the particular problem at hand. The distribution of the random variable *D* has been described analytically for a few simple cases in (1) and (2), and discussed further in (3) and (4). More recently there has been some renewed interest in the use of the distribution of *D* to describe multivariate i.i.d. samples $X_1,\ldots,X_n$ from F_X. In particular, the cumulative distribution function $F_D(d) = E1(d(X_1, X_2) \le d)$ of *D* can be estimated consistently from the set of dependent distances $\{d(X_i, X_j), i, j = 1,\ldots,n\}$ by $F_n(d) = (n(n-1)/2)^{-1}\sum_{i<j} 1(d(X_i, X_j) \le d)$. If one considers a grid of points $\{d_1,\ldots,d_K\}$ along the distance axis, then the vector $\sqrt{n}\{F_n(d_1) - F_D(d_1),\ \ldots, F_n(d_K) - F_D(d_K)\}$ converges in distribution as n tends to infinity to a zero-mean multivariate normal random variable, and a non parametric chi-square-like statistic based on $F_n(d)$ can be used to test for differences between the collection of all the interpoint distances observed in a sample and a null distribution, or between groups of observations. A general result de-

scribing the convergence of the estimator $F_n(d)$ to a Gaussian process is also available, see (2) and (5). This approach has been studied in the disease clustering setting in (6), (7), and in (8); in the genetics setting in (9); and in the biosurveillance setting in (10) and in (11). We refer to these papers for details on that nonparametric approach.

We now consider the estimation of parametric models for F_d. We assume that $D \sim f(d;\theta)$, where θ belongs to some parameter set $\Theta \subset \Re^p$, for some p. Letting $l(\theta;d)$ be the log-likelihood of D based on one observed value d of the interpoint distance, we focus on the estimation of the parameter θ from what one can call the U-score vector

$$\sum_{i<j} S_\theta(D_{(i,j)}) = 0 \,, \qquad [1]$$

or the unbiased estimating equation constructed from the marginal log-likelihood contributions, with $S_\theta(d)$ the score vector from the assumed distribution of D, and $D_{(i,j)} = d(X_i, X_j)$. The unbiasedness of [1] is immediate since the dependence among the distances has no effect on the expected values.

Thus the resulting estimator of θ is formally identical to the maximum likelihood estimator that one would compute if the distances were independent. Using results from the theory of U-statistics, in the next section we discuss the asymptotic properties of this estimator of θ, and we compare it to two alternative estimators obtained from i.i.d. reductions of the problem. We then discuss the analytically tractable case of the distance between independent and identically distributed observations arising from the bivariate normal distribution $N_2(0,\sigma^2 I_2)$. In the following section we then apply the methods to the construction of a mixture model within the context of biosurveillance data. We close with some discussion in the last section.

Methods

Estimation

The use of parametric models for the study of interpoint distance distributions requires that one be able to estimate the parameter θ of $f_D(d;\theta)$. Let us consider three possible estimators.

The first possibility is the estimator $\hat{\theta}_1 = \theta(\hat{\alpha})$, where $\hat{\alpha}$ is the maximum likelihood estimator of the vector parameter α of the underlying distribution $F_X(x;\alpha)$ of the original coordinates X. Very importantly, this requires the specification of the multivariate distribution of F_X, which might not be desirable (or feasible) if the dimensionality of X is indeed large. If, however, we allow for a moment for this possibility, then by also assuming that X is absolutely continuous with respect to Lebesgue measure we have an associated density function $f_X(x;\alpha)$ on $\Re^p$. As long as the function $\theta(\cdot)$ relating α to θ is known (and smooth enough), traditional likelihood theory applies since the individual observations $X_1,...,X_n$ are independent and identically distributed.

The second estimator is the maximum likelihood estimator $\hat{\theta}_2$ obtained by maximizing the likelihood function $L(\theta;D_1,...,D_{n/2})$ of D computed from $n/2$ independent distances obtained by pairing the observations, for example as in $(1,2),(3,4),\ ...,\ (n-1,n)$. This is equivalent to extracting from the $n(n-1)/2$ dependent distances $\{d(X_i,X_j), i,j = 1,...,n\}$ one particular set of $n/2$ independent distances (for simplicity we can assume n to be even). By independent distances here we mean that each element X_i, $i = 1,...,n$ can only

appear in exactly one of the distances, so that if one considers the distance matrix M_D of the distances $d(X_i, X_j)$, then the selection is equivalent to choosing an element (i, j) from M_D and then excluding the *ith* and *jth* rows and columns from subsequent consideration. Such a set of distances can then be written as $\{d(X_{\rho_1}, X_{\rho_2}), d(X_{\rho_3}, X_{\rho_4}), ..., d(X_{\rho_{n-1}}, X_{\rho_n})\}$ for a particular permutation $\rho = \{\rho_1, \rho_2, ..., \rho_n\}$ of the integers $\{1, 2, ..., n\}$, so that these distances can be indexed as $d_w, w = 1, ..., n/2$. The estimator $\hat{\theta}_2$ is thus defined as the solution of $\sum_{j=1}^{n/2} S_\theta\{d(X_{\rho_{(2j-1)}}, X_{\rho_{(2j)}})\} = 0$, and it can be identified as $\hat{\theta}_2 = \hat{\theta}_2(\rho)$. A combination of the possible estimators $\hat{\theta}_2(\rho)$ over the $n!$ permutations ρ or over the $n!/\{(n/2)!\, 2^{n/2}\}$ possible ways of constructing pairings could also be entertained.

The third estimator, $\hat{\theta}_3$, is the one that is obtained by solving the U-score estimating equation [1] above.

We now briefly review the well-known asymptotic results that apply to $\hat{\theta}_1$ and $\hat{\theta}_2$, and discuss asymptotic results that apply to $\hat{\theta}_3$. In what follows, the distributions that expectations refer to will be clear from the context, and thus they will be suppressed from the notation.

The asymptotic distributions of $n^{1/2}(\hat{\theta}_1 - \theta)$ and $n^{1/2}(\hat{\theta}_2 - \theta)$ (conditionally on the choice of the permutation ρ) are straightforward from likelihood theory. For $\hat{\theta}_1$ we first recall that for the maximum likelihood estimator $\hat{\alpha}$ of the parameter α of $F_X(x;\alpha)$, $n^{1/2}(\hat{\alpha} - \alpha)$ converges in distribution under regularity conditions to a normal random variable with mean zero and variance-covariance matrix

$$\left\{EH_\alpha(X)\right\}^{-1}\left[E\left\{S_\alpha(X)S_\alpha^T(X)\right\}\right]\left\{EH_\alpha^T(X)\right\}^{-1} = I_\alpha^{-1} \qquad [2]$$

where $S_\alpha(X) = \partial \log f_X(x;\alpha)/\partial\alpha^T$, $H_\alpha(X) = \partial S_\alpha(X)/\partial\alpha$, and with the expression simplifying to the inverse information $I_\alpha^{-1} = \left[E\left\{S_\alpha(X)S_\alpha^T(X)\right\}\right]^{-1}$ (see for example (12)).

Therefore, the asymptotic distribution of $n^{1/2}\{\theta(\hat{\alpha}) - \theta(\alpha)\}$ can be obtained immediately via the delta method as being $N\left\{0, \frac{\partial}{\partial\alpha}\theta(\alpha) I_\alpha \frac{\partial}{\partial\alpha}\theta(\alpha)^T\right\}$.

The asymptotic distribution of the centred and scaled estimator $\hat{\theta}_2$ follows the same lines, if one replaces the score $S_\alpha(x)$ for α by the score $S_\theta(d)$ for θ. Also, *n* here needs to be replaced by $n/2$, because that is the number of independent pairs obtained from *n* individuals. In other words, $(n/2)^{1/2}(\hat{\theta}_2 - \theta)$ also converges in distribution to a zero-mean normal random variable, with variance-covariance matrix

$$\left\{EH_\theta(D)\right\}^{-1}\left[E\left\{S_\theta(D)S_\theta^T(D)\right\}\right]\left\{EH_\theta^T(D)\right\}^{-1} = \left[E\left\{S(\theta;D)S^T(\theta;D)\right\}\right]^{-1} = I_\theta^{-1}. \qquad [3]$$

We now turn to our proposed estimator $\hat{\theta}_3$. Solving [1] above is equivalent to solving $[n(n-1)/2]^{-1}\sum_{i<j} S_\theta\{D_{(i,j)}\} = 0$. (Note how for $n = 2$ the estimators $\hat{\theta}_2$ and $\hat{\theta}_3$ are algebraically identical.) We use the notation Pg and $P_n g$ to indicate the expectation of the function $g(Y)$ with respect to the distribution $P = P_Y$ of the random variable Y and with respect to the empirical distribution $P_n = (1/n)\sum_{i=1}^n \delta_{y_i}$ respectively, with δ_{y_i} the Dirac delta function at $Y = y_i$. Under regularity conditions, the estimator $\hat{\theta}_3$ maximizes the quantity $U_n(\theta) = P_n^2 l_\theta(d(X_1, X_2)) = \{n(n-1)/2\}^{-1}\sum_{i<j} l_\theta\{d(X_i, X_j)\}$, where $l_\theta\{d(X_1, X_2)\} = \log f\{d(x_1, x_2);\theta\}$ is the log-likelihood for the parameter θ from the distribution of the random variable *D*. Define the functional $U(\theta) = P^2 l_\theta\{d(X_1, X_2)\} = E l_\theta\{d(X_1, X_2)\}$. Consistency of $\hat{\theta}_3$ then follows if one assumes that: (i) $U(\theta)$ is uniquely maximized at the true parameter value $\theta = \theta_0$; (ii) $U(\theta)$ is continuous, and (iii) $U_n(\theta)$ converges uniformly in probability to $U_0(\theta_0)$ (see for example (13)). Further, $n^{1/2}(\hat{\theta}_3 - \theta_0)$ converges in distribution to a normal random variable with mean zero and variance

$$4\left\{EH_{\theta_0}(D)\right\}^{-1}\left[E\{S_{\theta_0}(D_{(1,2)})S_{\theta_0}^T(D_{(1,3)})\}\right]\left\{EH_{\theta_0}^T(D)\right\}^{-1}. \qquad [4]$$

When comparing [4] and [3] it should be noted that the convergence in [3] is based on the scaling constant $(n/2)^{1/2}$, while the convergence in [4] uses $n^{1/2}$. Also, the central terms in the two variance-covariance matrices are different.

This asymptotic result can be shown to hold in great generality by making use of some results from the theory of U-statistics and U-processes. In particular, the proof of the asymptotic normality is similar to that for the empirical semplicial median (see(14)). A general result about the asymptotic distribution of M-estimators based on a criterion function of several variables (as we have here) is stated as Theorem 5.5.7 in the same reference, where detailed regularity conditions are also described. In the appendix we detail the specialization of that result to the problem being considered here.

We can easily construct a consistent estimator for the variance-covariance matrix given in [4] above. From U-statistics theory, if $h(X_1, X_2)$ is a symmetric kernel such that $Eh^2(X_1, X_j) < \infty$, then

$$\sqrt{n}\left\{\binom{n}{2}^{-1}\sum_{i<j} h(X_1, X_2) - Eh(X_1, X_2)\right\} \xrightarrow{\text{d}} N\left(0, 4E\left\{h(X_1, X_2)h(X_1, X_3)^T\right\}\right)$$

as *n* tends to infinity (see for example (15) or (12)).

Letting $h(X_1, X_2) = H_\theta(D)$, it follows that as long as $E\{H_\theta(D)\}^2 < \infty$, as n tends to infinity one has that $\hat{E}H_{\theta_0}(D) = (n(n-1)/2)^{-1}\sum_{i<j} H_{\theta_0}(D_{(i,j)}) \to EH_{\theta_0}(D)$ in probability. Note that this result actually holds (as long as $E\,|\,H_\theta(D)\,| < \infty$) with the convergence being almost surely and in L^1 (see (14)). After symmetrisation of the kernel $S_{\theta_0}(D_{(i,j)})S_{\theta_0}^T(D_{(i,k)})$, from this result we obtain

$$\hat{E}S_{\theta_0}(D_{(1,2)})S_{\theta_0}^T(D_{(1,3)}) = \frac{1}{6\binom{n}{3}}\sum_{i<j<k} h(i,j,k) \xrightarrow{\text{p}} E\left[S_{\theta_0}(D_{(1,2)})S_{\theta_0}^T(D_{(1,3)})\right]$$

as *n* tends to infinity, where we have defined

$$h(i,j,k) = \sum_{(\rho_1,\rho_2,\rho_3)\in R(i,j,k)} S_{\theta_0}(D_{(\rho_1,\rho_2)})S_{\theta_0}^T(D_{(\rho_2,\rho_3)}),$$

with $R(i,j,k)$ being the set of the $3! = 6$ permutations of (i,j,k).

Both estimators are square matrices with the same dimension as θ_0, and they can be computed with $\hat{\theta}_3$ replacing θ_0 as long as $S_\theta(D)$ and $H_\theta(D)$ can be expanded around θ_0 with a first-order Taylor expansion. Thus an approximate large-sample variance estimator for $\hat{\theta}_3$ is

$$\frac{4}{n}\{\hat{E}H_{\hat{\theta}_3}(D)\}^{-1}[\hat{E}\{S_{\hat{\theta}_3}(D_{(1,2)})S_{\hat{\theta}_3}^T(D_{(1,3)})\}]\{\hat{E}H_{\hat{\theta}_3}(D)^T\}^{-1}. \qquad [5]$$

[Indeed, a similar estimator can be constructed for the asymptotic variance-covariance matrix of the non-parametric estimator of the interpoint distance distribution described in (5). Details of that construction are provided in the appendix].

Note that there is an interesting connection between the estimating equations for $\hat{\theta}_2$ and $\hat{\theta}_3$ that further motivates the construction of $\hat{\theta}_3$ (in addition to the unbiasedness of [1] above). We show in the appendix that

$$\sum_{\rho}\sum_{j=1}^{n/2} S_\theta\{d(X_{\rho_{(2j-1)}}, X_{\rho_{(2j)}})\} = n(n-2)!\sum_{i<j} S_\theta\{d(X_i, X_j)\}, \qquad [6]$$

so that solving [1] above is equivalent to solving $(n!)^{-1}\sum_{\rho}\sum_{j=1}^{n/2} S_\theta\{d(X_{\rho_{(2j-1)}}, X_{\rho_{(2j)}})\} = 0$, or the mean of the estimating equations that produce $\hat{\theta}_2$ over all possible permutations ρ of the integers $\{1,...,n\}$. This is similar to the relationship that expresses the sample variance from a univariate sample $Y_1,...,Y_n$ as the average of the estimators built on the distances $(Y_i - Y_j)^2$ from a specific permutation of $\{1,...,n\}$, with the average taken over all possible permutations.

In the next section, we discuss a specific situation in which we assume knowledge of the underlying distribution F_X of the coordinates from which Euclidean distances $D = d(X_1, X_2)$ are computed. The specification of F_X allows in this case for the analytic derivation of all quantities described above, and thus their analytical examination.

Illustration: the bivariate normal distribution with Euclidean distance

Consider $X_1,...,X_n$, $X_i = \{X_{(i,1)}, X_{(i,2)}\}^T \overset{i.i.d.}{\sim} N(0,\sigma^2 I_2)$. The optimal estimation of σ^2 requires maximization of the log-likelihood $\sum_{i=1}^{n}\sum_{j=1}^{2}\left[\log(2\pi)^{-1/2}\} - (2\sigma^2)^{-1}X_{(i,j)}^2 - (1/2)\log(\sigma^2)\right]$. Since the two coordinates are independent and have the same distribution, we can simplify the notation and re-define the coordinates to be called $X_1,...,X_{2n}$. Differentiation of the log-likelihood produces the score $S_{\sigma^2}(X_1,...,X_{2n}) = -n/\sigma^2 + \sum_{i=1}^{2n} X_i^2/(2\sigma^4)$. This derivative can be set equal to zero to yield $\hat{\sigma}_1^2 = (2n)^{-1}\sum_{i=1}^{2n} X_i^2$. Note how this estimator requires the use of the original coordinates X_i, and the explicit knowledge of their probabilistic model F_X. The (exact) variance of this estimator is obtained immediately as $\mathrm{var}(\hat{\sigma}_1^2) = (2n)^{-2}\sum_{i=1}^{2n}\mathrm{var}(X_i^2) = (2n)^{-1}\mathrm{var}(X_1^2) = \sigma^4/n$ since $X_1^2 \sim \sigma^2\chi_1^2$, and therefore $\mathrm{var}(X_1^2) = 2\sigma^4$.

Note that this same result can be obtained by computing the asymptotic variance of the estimator from the model's information as in [2] above. First, note that $EX_1^2X_2^2 = (EX_1^1)^2 = \sigma^4$. Also, integration by parts yields $EX_1^4 = 3\sigma^4$. Easy algebra then shows that the expected value of the squared score (the information) $I_{\sigma^2} = ES_{\sigma_2}^2$ is equal to n/σ^4. The inverse of I_{σ^2} is therefore also equal to σ^4/n, as one should expect, and the variance of $\hat{\sigma}_1^2$ can be approximated by σ^4/n.

Now, consider $\hat{\theta}_2 = \hat{\sigma}_2^2$ and $\hat{\theta}_3 = \hat{\sigma}_3^2$. Given the definition of Euclidean distance, under the assumed model for the coordinates X, the distribution of the squared interpoint distance between two randomly selected points X_1 and X_2 follows the exponential distribution with parameter $\lambda = (4\sigma^2)^{-1}$, as $D^2 = \{X_{(1,1)} - X_{(2,1)}\}^2 + \{X_{(1,2)} - X_{(2,2)}\}^2$ is the sum of two independent random variables each with distribution $2\sigma^2\chi_1^2$ and is therefore distributed as a $2\sigma^2\chi_2^2$ random variable. We work directly on estimating the parameter $\theta = \sigma^2$, as the function $\theta(\alpha)$ here is the identity function. Since $D^2 \sim \exp\{1/(4\sigma^2)\}$, the density function of *D* is equal to $f_D(d;\sigma^2) = (2\sigma^2)^{-1} d\exp\left(d^2/(4\sigma^2)\right)$.

For the distance *D* the derivative of the log-likelihood is equal to $S_{\sigma^2}(D) = -(\sigma^2)^{-1} + D^2/(4\sigma^4)$, which after summing over the $n/2$ distances $d_1,...,d_{n/2}$ and setting equal to zero, yields the estimator $\hat{\sigma}_2^2 = (2n)^{-1}\sum_{w=1}^{n/2} d_w^2$. The (exact) variance of this estimator is immediately found to be equal to $\mathrm{var}(\hat{\sigma}_2^2) = (4n^2)^{-1}(n/2)\mathrm{var}(D_1^2) = (8n)^{-1}\left(4\sigma^2\right)^2 = 2\sigma^4/n.$

The direct verification of the expression of the asymptotic variance for $\hat{\sigma}_2^2$ in [3] requires the term $EH_{\sigma^2}(D) = E\left((\sigma^4)^{-1} - D^2/(2\sigma^6)\right) = (\sigma^4)^{-1} - 4\sigma^2/(2\sigma^6) = -(\sigma^4)^{-1}$. It is easy to verify that this is also equal to $ES_{\sigma^2}^2$, so that the approximate variance of $\hat{\sigma}_2^2$ is indeed equal to $(2/n)I_{\sigma^2}(D)^{-1} = (2/n)\sigma^4$.

The estimator $\hat{\theta}_3 = \hat{\sigma}_3^2$ treats all $n(n-1)/2$ dependent distances as if they were independent. It follows that $\hat{\sigma}_3^2 = \{2n(n-1)\}^{-1} \sum_{i<j} d(X_i, X_j)^2$. Note that by the argument in the previous paragraph, if one assumed that the $n(n-1)/2$ distances were really independent, then one would (mistakenly) compute the variance of $\hat{\sigma}_3^2$ as being equal to $2\sigma^4 / \{n(n-1)\}$. This would grossly underestimate the true variance, which we now show to be equal to σ^4 / n. The asymptotic variance formula in [4] requires the same term $EH_{\sigma^2}(D)$ above, as well as the additional term $ES_{\sigma^2}(D_{(1,2)})S_{\sigma^2}(D_{(1,3)})$. Since in this example the underlying distribution F_X is known, we can also compute the true theoretical value for this latter quantity: in particular, one can check that $E\left\{S_{\sigma^2}(D_{(1,2)})S_{\sigma^2}(D_{(1,3)})\right\} = (4\sigma^4)^{-1}$ (please see the appendix), so that from [5], the approximate large sample variance of $\hat{\sigma}_3^2$ is equal to $(4/n)(-\sigma^4)(4\sigma^4)^{-1} - \sigma^4 = 4\sigma^4/(4n) = \sigma^4/n$.

Lastly, note that for $n = 2$ the two estimators $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$ are algebraically identical (and they must therefore have the same variance) but that as $n \to \infty$, $\mathrm{var}(\hat{\sigma}_1^2) \sim \mathrm{var}(\hat{\sigma}_3^2) \sim (0.5)\,\mathrm{var}(\hat{\sigma}_2^2)$.

A mixture model for biosurveillance

One of the aims of biosurveillance is the identification of geographic patterns that represent aberrations from an assumed null behaviour of recorded health events (16). Information recorded at most health care encounters includes patient home address, a demographic variable that, when geocoded, may have value for surveillance. Outbreak detection using geographic coordinates requires that a baseline distribution of patients be established and that population density be accounted for. It has previously been observed that the distribution of the pairwise interpoint distance among patients tracked in the surveillance system being studied shows remarkable stability over time (see (17), (18)). The approach that we have discussed above can be used to entertain parametric models to describe these interpoint distance distributions. The data that we consider consists of all visits for patients with respiratory illness presenting to the emergency department of an urban academic tertiary care paediatric hospital from December 21st, 1998 through January 12th, 2002. Patients living more than 50 miles from the hospital were excluded. The chief complaint and International Classification of Diseases (ICD) codes of eligible patients were used to select those with respiratory illness (see (19)). Patients' home addresses were translated to geographic coordinates using geocoding software (ArcGIS 8.2, Environmental Systems Research Institute, Redlands, CA). Addresses were cleaned prior to geocoding using software (ZP4, Semaphore Corp, Aptos, CA) that matched addresses to the August 2002 US Postal Service ZIP+4 database and made corrections.

We now show how the techniques described above specialize to a parametric model suggested by this data (for a complete discussion of the biosurveillance data system we refer to (16) and (19)). Observation of the histogram of the interpoint distances computed among the patients' addresses suggested that a reasonable model for the interpoint distance could be a mixture of a lognormal and a normal distribution. Clearly, the support of such a distribution also includes (impossible) negative distance values, but the location of the normal component of the mixture appears to be such that this model can be expected to provide a good fit. Equivalently, we assume that D has the density function $f_D(d;\theta) = \alpha f_1(d;\mu_1,\sigma_1^2) + (1-\alpha) f_2(d;\mu_2,\sigma_2^2)$, with $f_1(d)$ the lognormal density with parameter (μ_1,σ_1^2) and $f_2(d)$ the normal density with parameter (μ_2,σ_2^2), with the mixing parameter $\alpha \in (0,1)$.

For a general two-component mixture $f_D\{d;\theta = (\theta_1,\theta_2,\alpha)^T\} = \alpha f_1(d;\theta_1) + (1-\alpha) f_2(d;\theta_2)$ of two densities $f_1(d;\theta_1)$ and $f_2(d;\theta_2)$ with separate parameters θ_1 and θ_2, we let $S_1 = S_1(\theta_1) = \partial \log f_1(d;\theta_1)/\partial\theta_1^T$

and $S_2 = S_2(\theta_2) = \partial \log f_2(d;\theta_2)/\partial\theta_2^T$ be the two score vectors. Differentiation of the log-likelihood $\log f_D(d;\theta)$ with respect to θ^T yields the score vector $S_D(d) = [S_D^T(\theta_1), S_D^T(\theta_2), S_D(\alpha)]^T$, where for ease of notation we drop the parameters and the argument *d*:

$$S_D(\theta_1) = \frac{\partial}{\partial\theta_1^T}\log f_D(d;\theta) = \frac{\alpha f_1}{f_D}S_1$$
$$S_D(\theta_2) = \frac{\partial}{\partial\theta_2^T}\log f_D(d;\theta) = \frac{(1-\alpha) f_2}{f_D}S_2$$
$$S_D(\alpha) = \frac{\partial}{\partial\alpha}\log f_D(d;\theta) = \frac{f_1 - f_2}{f_D}.$$

Let $H_1 = H_1(\theta_1,\theta_1) = \partial^2 \log f_1(d)/(\partial\theta_1\partial\theta_1^T)$ and $H_2 = H_2(\theta_2,\theta_2) = \partial^2 \log f_1(d)/(\partial\theta_2\partial\theta_2^T)$ be the two Hessian matrices for $f_1(d)$ and $f_2(d)$. After taking partial mixed derivatives of the log-likelihood, the Hessian matrix $H_D(\theta,\theta^T) = \partial^2 \log f_D(d)/(\partial\theta\partial\theta^T)$ for $f_D(d)$ is equal to

$$H_D(\theta,\theta^T) = \begin{bmatrix} \alpha\left\{\frac{(1-\alpha)f_1 f_2}{f_D^2}S_1S_1^T + \frac{f_1}{f_D}H_1\right\} & -\frac{\alpha(1-\alpha)f_1 f_2}{f_D}S_1S_2^T & \frac{f_1 f_2}{f_D^2}S_1 \\ -\frac{\alpha(1-\alpha)f_1 f_2}{f_D}S_2S_1^T & (1-\alpha)\left\{\frac{\alpha f_1 f_2}{f_D^2}S_2S_2^T + \frac{f_2}{f_D}H_2\right\} & -\frac{f_1 f_2}{f_D^2}S_2 \\ \frac{f_1 f_2}{f_D^2}S_1^T & -\frac{f_1 f_2}{f_D^2}S_2^T & -\frac{(f_1 - f_2)^2}{f_D^2} \end{bmatrix}.$$

In the specific mixture model suggested by our data we have $\theta_1 = (\mu_1,\sigma_1^2)$ and $\theta_2 = (\mu_2,\sigma_2^2)$ as the parameters in the two densities $f_1(d;\theta_1) = (d\,\sigma_1)^{-1}(2\pi)^{-1/2}\exp\left\{-(\log d - \mu_1)^2/(2\sigma_1^2)\right\}$ and $f_2(d;\theta_2) = (\sigma_2^2 2\pi)^{-1/2}\ \exp\left\{-(d-\mu_2)^2/(2\sigma_2^2)\right\}$.
The score vector and Hessian matrix for f_1 are

$$S_1 = \begin{bmatrix} \frac{1}{\sigma_1^2}(\log d - \mu_1) \\ -\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}(\log d - \mu_1)^2 \end{bmatrix}; H_1 = \begin{bmatrix} -\frac{1}{\sigma_1^2} & -\frac{1}{\sigma_1^4}(\log d - \mu_1) \\ -\frac{1}{\sigma_1^4}(\log d - \mu_1) & \frac{1}{2\sigma_1^4} - \frac{1}{\sigma_1^6}(\log d - \mu_1)^2 \end{bmatrix}.$$

For f_2 one has

$$S_2 = \begin{bmatrix} \frac{d-\mu_2}{\sigma_2^2} \\ -\frac{1}{2\sigma_2^2} + \frac{(d-\mu_2)^2}{2\sigma_2^4} \end{bmatrix}; H_2 = \begin{bmatrix} -\frac{1}{\sigma_2^2} & -\frac{d-\mu_2}{\sigma_2^4} \\ -\frac{d-\mu_2}{\sigma_2^4} & \frac{1}{2\sigma_2^4} - \frac{(d-\mu_2)^2}{\sigma_2^6} \end{bmatrix}.$$

The U-score estimator $\hat{\theta}_3$ is obtained by solving [1], or equivalently by maximizing the likelihood, which can be done using standard software. In particular, we used the SAS/OR procedure NLP, after reparametrization of the mixing parameter α through the logistic function.

We fit the lognormal-normal mixture model to the biosurveillance data consisting of all interpoint distances among the home addresses of 708 patients with respiratory illness who visited the paediatric emergency room

during the last two consecutive weeks in the period December 30th, 2001 trough January 12th, 2002. We obtained the estimate $[\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\alpha}] = [1.721, 0.878, 19.563, 78.363, 0.837]$, with associated standard errors [0.062,0.055,2.261,9.625,0.042], obtained with the estimator described in [5] above. Note that any default standard errors that are produced by the maximization software should be discarded, as they are based on the wrong assumption that the distances constitute a random sample of independent and identically distributed distances that follow the assumed model.

As an exploratory tool, we have compared the average of the variance estimates [5] computed on 20 independent samples obtained from the three-year data with the empirical variance of the U-score estimates computed on 1000 independent samples of 700 locations, sampled from the same three years of respiratory data. The ratio of the standard errors of the estimators to the empirical standard errors ranged between 0.92 and 1.33, thus showing reasonable agreement. Since the theoretical variance estimator is based on the assumed mixture parametric model while the empirical standard errors are based on resampling from the actual distribution of addresses, this comparison also indirectly increases the confidence in the parametric model that we have assumed.

Further analyses will study the relative merits of this new parametric approach when compared to the non-parametric one, but one would expect a gain in efficiency from using the former if the model describes the data reasonably accurately.

Conclusions

We have shown in detail how parametric models can be constructed for the analysis of distance data. We have motivated the study of the estimating equation estimator with the study of the bivariate normal case, and have provided both general formulas for inference and specialized formulas for mixture distributions.

Consideration of the interpoint distance distribution for the analysis of multivariate data may prompt the question of identifiability of the underlying distribution of the coordinates from the interpoint distance distribution, especially for lower-dimensional problems. The question of the description of the class of distributions $F_X(x)$, $x \in \Re^2$ that produce a particular distribution F_D of the interpoint distance D is a difficult one as one cannot exclude the possibility that several distributions F_X may produce the same distribution for D. For example, when working with Euclidean distances, any translation or rotation of the axes produces the same $F_D(d)$, so that one only needs consider equivalence classes up to these transformations. This (and other) invariance properties, however, do not prevent one from modelling the interpoint distance distribution directly without having to worry about the interpretation of features of that distribution in terms of the underlying distribution F_X. For example, the mixture structure that we have illustrated in the previous section is useful to describe the observed interpoint distance distribution, regardless of the fact that it would not be easy (if at all possible) to produce an inverse mapping to the original coordinates.

Note that while we have not explicitly elaborated on this, it should be clear that performing hypothesis testing and constructing confidence intervals are both straightforward from knowledge of the asymptotic distribution of the estimators of the model parameters. In applications, one may for example be interested in testing the hypothesis of equality of the interpoint distance distributions between two groups of observations and chi-square-distributed quadratic forms can easily be constructed for that purpose from the estimated pa-

rameter estimates and their estimated variance-covariance matrices. For example, in the bivariate data that we have discussed earlier, this would allow the testing of the hypothesis of no deviation of a recently observed sample from a known historical norm against the general alternative hypothesis of a distributional change in the distribution of the cases. (Note that this is not limited to alternative hypotheses of clustering).

The methods that we have discussed can be used in a variety of settings, ranging from low-dimensional data (e.g. geographic coordinates, as in the illustration discussed above) to very highly dimensional data. As a matter of fact, it should be stressed again that while here we have focused on the Euclidean distance on the plane, the same discussion also applies to any symmetric dissimilarity measure between observations. Consideration of these distances instead of the individual coordinates allows one to entertain the analysis of data in large and complicated spaces without the need for the specification of the multivariate distribution of the coordinates, and with a great reduction in the dimensionality of the problem.

There are many settings where dissimilarities between high dimensional quantities are of interest (e.g. clustering, functional data analysis, nonparametric tests, etc.), but we mention two in particular. We cited above the use of interpoint distance distributions when distances are genetic distances (see (9)). Another area of potential application is the analysis of life courses in demography and of sequences in general (see (20) and (21)). Distances between sequences can be constructed, for example, from optimal matching techniques as in (22).

Appendix

On the asymptotic normality of $\hat{\theta}_3$

To apply Theorem 5.5.7 in (14) we need the following development for $U(\theta)$ near θ_0 (taken without loss of generality to be zero) to hold: $U(\theta) = U(0) - (1/2)\theta A\theta^T + o(|\theta|^2)$. If we assume that the log-likelihood can be expanded around $\theta_0 = 0$ via Taylor expansion then the requirement is satisfied immediately, as $l_\theta(d) = l_0(d) + l'_0(d)\theta^T + (1/2)\theta\{l''_0(d)\}\theta^T + o(|\theta|^2)$ implies

$$U(\theta) = U(0) + E\{l'_0(D)\}\theta^T + \frac{1}{2}\theta E\{l''_0(D)\}\theta^T + o(|\theta|^2)$$

$$= E\{l_0(D)\} - \frac{1}{2}\theta[-E\{l''_0(D)\}]\theta^T + o(|\theta|^2)$$

by the properties of score functions, since $l'_\theta(d) = \partial l_\theta(d)/\partial\theta^T = S_\theta(d)$. The theorem in (14) also requires the identification of a function $\Delta(X): \Re^2 \to \Re^d$ (where d is the dimension of θ) such that $E\Delta(X) = 0$ and $E|\Delta(X)|^2 < \infty$ and such that the function $\pi_1 l_\theta(x) = (\delta_{x_1} - P) \times Pl_\theta$ is stochastically differentiable at zero. Here $\pi_1 l_\theta(x) = \delta_{x_1} Pl_\theta\{d(x_1, X)\} - P^2 l_\theta\{d(X_1, X_2)\} = El_\theta\{d(x_1, X)\} - El_\theta\{d(Xb_1, X_2)\}$. One can identify the function $\Delta(X) = \int_{\Re^2} S_\theta\{d(X, x_2)\} f_X(x_2) dx_2$, so that

$$E\Delta(X) = \int_{\Re^2}\int_{\Re^2} S_\theta\{d(x_1, x_2)\} f_\theta(x_2) f_\theta(x_1) dx_1 dx_2 = ES_\theta\{d(X_1, X_2)\} = 0.$$

From Theorem 5.5.7 we then conclude that $n^{\frac{1}{2}}\hat{\theta}_3 \xrightarrow{d} \mathbf{Z}$ as n tends to infinity, where $\mathbf{Z}$ is $N(0, \Delta)$ and $\Delta = 4A^{-1}\{\text{cov}\Delta(X)\}A^{-1}$, and we have seen above that $A = \left[-El''_0(D)\right]$. Finally we have

$$
\begin{aligned}
\mathrm{cov}\{\Delta(X)\} &= E_X \Delta(X)\Delta(X)^T \\
&= \int_{\Re^2} \left[\int_{\Re^2} S_\theta\{d(x_1,x_2)\} f_\theta(x_2) dx_2\right]\left[\int_{\Re^2} S_\theta\{d(x_1,x_2)\} f_\theta(x_2) dx_2\right]^T f_\theta(x_1) dx_1 \\
&= \int_{\Re^2} \left[\int_{\Re^2} S_\theta\{d(x_1,x_2)\} f_\theta(x_2) dx_2\right]\left[\int_{\Re^2} S_\theta\{d(x_1,x_3)\} f_\theta(x_3) dx_3\right]^T f_\theta(x_1) dx_1 \\
&= \int_{\Re^2} \left[\int_{\Re^2}\int_{\Re^2} S_\theta\{d(x_1,x_2)\} f_\theta(x_2) S_\theta^T\{d(x_1,x_3)\} f_\theta(x_3) dx_2 dx_3\right] f_\theta(x_1) dx_1 \\
&= \int_{\Re^2}\int_{\Re^2}\int_{\Re^2} \left[S_\theta\{d(x_1,x_2)\} S_\theta^T\{d(x_1,x_3)\}\right] f_\theta(x_1) f_\theta(x_2) f_\theta(x_3) dx_1 dx_2 dx_3 \\
&= E\left[S_\theta\{d(X_1,X_2)\} S_\theta^T\{d(X_1,X_3)\}\right],
\end{aligned}
$$

which matches the expression in [4].

A variance estimator in the nonparametric setting

Let $F_D(d) = E1(d(X_1,X_2) \le d)$ be the cumulative distribution function of the interpoint distance between two randomly selected points X_1 and X_2 generated from some distribution F_X. A consistent estimator for the interpoint distance cumulative distribution function $F_D(d)$ at d is $F_n(d) = (n(n-1)/2)^{-1} \sum_{i<j} 1(d(X_i,X_j) \le d)$. As discussed in (2) and in (5), if one considers a grid of points $\{d_1,...,d_K\}$ along the distance axis, then the vector $\sqrt{n}\{F_n(d_1) - F_D(d_1), \ ..., F_n(d_K) - F_D(d_K)\}$ converges in distribution as n tends to infinity to a zero-mean multivariate normal random variable, with variance-covariance matrix $\Sigma = \{\sigma_{a,b}\}$, with $\sigma_{a,b} = \mathrm{cov}\left[1\{d(X_1,X_2) \le d_a, d(X_1,X_3) \le d_b\}\right]$, $a,b = 1,...,K$.

Since $\sigma_{a,b} = E\left[1\{d(X_1,X_2) \le d_a, d(X_1,X_3) \le d_b\}\right] - E1\{d(X_1,X_2) \le d_a\} E1\{d(X_1,X_3) \le d_b\}$, the same U-statistics results used to construct the consistent estimator for the variance-covariance matrix of $\hat{\theta}_3$ can be applied to this nonparametric case. The covariance matrix Σ can be estimated consistently by the terms

$$
\begin{aligned}
\hat{\sigma}_{a,b} = 4 \Bigg(& \frac{1}{\binom{n}{3}} \sum_{1 \le i < j < k \le n} h(X_i, X_j, X_k; d_a, d_b) \\
& - \left[\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} 1\{d(X_i,X_j) \le d_a\}\right]\left[\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} 1\{d(X_i,X_j) \le d_b\}\right] \Bigg),
\end{aligned}
$$

where $h(X_i,X_j,X_k;d_a,d_b) = 6^{-1} \sum_\rho \left[1\{d(X_{\rho_1},X_{\rho_2}) \le d_a, d(X_{\rho_1},X_{\rho_3}) \le d_b\}\right]$ is the symmetrised kernel computed over the collection $\rho = \{(\rho_1,\rho_2,\rho_3)\}$ of the six permutations of the indices (i,j,k). In the calculation of this estimator, for efficiency the triple sum should be implemented as a single loop by making use of (fast) matrix multiplications for the inner sums.

Proof of equality [6]

We now prove the equality

$$
\sum_\rho \sum_{j=1}^{n/2} S_\theta\{d(X_{\rho_{(2j-1)}}, X_{\rho_{(2j)}})\} = n(n-2)! \sum_{i<j} S_\theta\{d(X_i,X_j)\}
$$

given earlier. Note that for each permutation $\rho = \{\rho_1,...,\rho_n\}$ of the integers $\{1,...,n\}$, the left hand side contains the

$n(n-2)!$ permutations, because it can appear in any of the $n/2$ positions $(1,2),(3,4),...,(n-1,n)$, and for each of these, there are $(n-2)$ ways of rearranging the other elements in the permutation. As $S_\theta\{D_{(1,2)}\} = S_\theta\{D_{(2,1)}\}$, there are a total of $(2n/2)(n-2)! = n(n-2)!$ permutations that produce the term $S_\theta\{D_{(1,2)}\} = S_\theta\{D_{(2,1)}\}$. This proves the result. Note also that since there are $n(n-2)!$ permutations that produce the term $S_\theta\{D_{(i,j)}\}$ for any pair (i,j), and because there are $n(n-1)/2$ such pairs, from the right hand side of [2] we anticipate a total of $(n(n-1)/2)^{-1}\, n(n-2)! = (n!n)/2$ terms in the double sum on the left hand side. That this is the case is clear from the observation that each of the $n!$ permutations produces $n/2$ terms, so that the total number of terms in the left hand side is indeed $(n!n)/2$.

Calculation of $E\left\{S_{\sigma^2}(D_{(1,2)})S_{\sigma^2}(D_{(1,3)})\right\}$ for the bivariate normal

First, note that

$$\begin{aligned} ED^2_{(1,2)}D^2_{(1,3)} &= E\left\{(X_{11}-X_{21})^2+(X_{12}-X_{22})^2\right\}\left\{(X_{11}-X_{31})^2+(X_{12}-X_{32})^2\right\} \\ &= E\Big\{(X_{11}-X_{21})^2(X_{11}-X_{31})^2+(X_{11}-X_{21})^2(X_{12}-X_{32})^2+ \\ &+(X_{12}-X_{22})^2(X_{11}-X_{31})^2+(X_{12}-X_{22})^2(X_{12}-X_{32})^2\Big\}, \end{aligned}$$

with $E(X_{11}-X_{21})^2(X_{12}-X_{32})^2 = E(X_{12}-X_{22})^2(X_{11}-X_{31})^2 = 4\sigma^4$ and

$$\begin{aligned} &E\left\{(X_{11}-X_{21})^2(X_{11}-X_{31})^2\right\} = E\left\{\left(X_{11}^2+X_{21}^2-2X_{11}X_{21}\right)\left(X_{11}^2+X_{31}^2-2X_{11}X_{31}\right)\right\} \\ &= E\Big\{X_{11}^4+X_{11}^2X_{31}^2-2X_{11}^3X_{31}+X_{11}^2X_{21}^2+X_{21}^2X_{31}^2-2X_{11}X_{21}^2X_{31}-2X_{11}^3X_{21}- \\ &-2X_{11}X_{21}X_{31}^2+4X_{11}^2X_{21}X_{31}\Big\} = 3\sigma^4+\sigma^4+\sigma^4+\sigma^4 = 6\sigma^4. \end{aligned}$$

Thus $ED^2_{(1,2)}D^2_{(1,3)} = 6\sigma^4+4\sigma^4+4\sigma^4+6\sigma^4 = 20\sigma^4$. Finally we have

$$\begin{aligned} E\left\{S_{\sigma^2}(D_{(1,2)})S_{\sigma^2}(D_{(1,3)})\right\} &= E\left\{\left(-\frac{1}{\sigma^2}+\frac{D^2_{(1,2)}}{4\sigma^4}\right)\left(-\frac{1}{\sigma^2}+\frac{D^2_{(1,3)}}{4\sigma^4}\right)\right\} \\ &= \frac{1}{\sigma^4}-\frac{1}{4\sigma^6}ED^2_{(1,2)}-\frac{1}{4\sigma^6}ED^2_{(1,3)}+\frac{1}{16\sigma^8}ED^2_{(1,2)}D^2_{(1,3)} = \frac{1}{\sigma^4}-\frac{4\sigma^2}{4\sigma^6}-\frac{4\sigma^2}{4\sigma^6}+\frac{1}{16\sigma^8}20\sigma^4 \\ &= \frac{1}{4\sigma^4}. \end{aligned}$$

Acknowledgments

This work was supported in part by NIH (NIAID) grant AI28076, National Library of Medicine grant LM07677-01, and by Massachusetts Department of Public Health/Centers for Disease Control and Prevention grant 52253337HAR.

References

1. Borel E. Traité du Calcul des Probabilités et de ses Applications, I. Paris: Gauthier-Villars, 1925.
2. Bartlett MS. The spectral analysis of two-dimensional point processes. Biometrika 1964; 51:299-311.
3. Silverman BW. Limit theorems for dissociated random variables. Advances in Applied Probability 1976; 8:806-819.
4. Sheng T.K. The distance between two random points in plane regions. Advances in Applied Probability 1985; 17:748-773.
5. Bonetti M, Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection. Stat Med 2005; 24(5):753-773.
6. M. Kulldorff, T. Tango, and P.J. Park. Power comparisons for disease clustering tests. Computational Statistics and Data Analysis 2003; 42:665-684.

7. Ozonoff A, Bonetti M, Forsberg L, Pagano M. Power comparisons for an improved disease clustering test. Computational Statistics and Data Analysis 2005; 48 (4):679-684.
8. Forsberg White L, Bonetti M, Pagano M. The choice of the number of bins for the M statistic. Computational Statistics and Data Analysis, In press, 2009.
9. Kowalski J, Pagano M, DeGruttola V. A nonparametric test of gene region heterogeneity associated with phenotype. Journal of the American Statistical Association 2002;97:398-408.
10. Bonetti M, Forsberg L, Ozonoff A, Pagano M. The distribution of interpoint distances. Frontiers in Applied Mathematics, SIAM, 2003.
11. Ozonoff A, Forsberg L, Bonetti M, Pagano M. A bivariate method for spatio-temporal syndromic surveillance. Morbidity and Mortality Weekly Report 2004; 53 (Suppl):61-66.
12. van der Vaart AW. Asymptotic Statistics. Cambridge University Press, 1998.
13. Newey WK, McFadden D. Large sample estimation and hypothesis testing, pages 2113-2241. Elsevier/ North-Holland, 1994.
14. de la Peña V, Giné E. Decoupling. Springer-Verlag, 1999.
15. Hoeffding W. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 1948; 19:293-325.
16. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin J, Gesteland PH, Treadwell T, Koski E, Hutwagner L, Buckeridge DL, Aller R, Grannis S. Syndromic surveillance: a guide informed by the early experience. Journal of the American Medical Informatics Association 2004; 11(2):141-150.
17. Olson KL, Bonetti M, Pagano M, Mandl KD. Syndromic surveillance: a population-adjusted stable geospatial baseline for outbreak detection in syndromic surveillance. Abstract, Morbidity and Mortality Weekly Report 2004; 53 (Suppl):256.
18. Olson KL, Bonetti M, Pagano M, Mandl KD. Real time spatial cluster detection using interpoint distances among precise patient locations. BMC Med Inform Decis Mak 2005; 5:19.
19. Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr Emerg Care 2004; 20(6):355-360.
20. Abbott A. Sequence analysis: New method for old ideas. Annual Review of Sociology 1995; 21:93-113.
21. Billari F, Piccarreta R. Analyzing demographic life courses through sequence analysis. Mathematical Population Studies 2005; 12.
22. Abbott A, Tsay A. Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociological methods and Research 2000; 29:3-33.