Distance in Big Dimensions

Jon Gillard

School of Mathematics, Cardiff University

GillardJW@Cardiff.ac.uk

Joint work with Emily O'Riordan and Anatoly Zhigljavsky (Cardiff)

Introduction

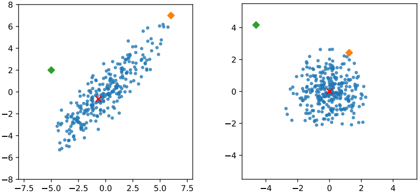
Context

- Many (all?) multivariate statistical techniques are reliant on computing distance.
- Whitening data can improve efficiency and accuracy.
- High-dim. data covariance matrix Σ often singular, or close.

'Gold-standard' (squared) Mahalanobis distance

Mahalanobis $\quad x^\top \Sigma^{-1} x$

Euclidean $\quad y^\top y \leftrightsquigarrow$ whitening $\leftrightsquigarrow$ transform $y = \Sigma^{-1/2} x$



This talk

Overview of two relatively new families of distances:

- Simplicial distances (by Pronzato, Wynn, Zhigljavsky)
- Minimal-variance distances

Finally, how we may (approximately) whiten data:

- Minimal-variance whitening

Simplicial Distances

see e.g. Pronzato, Wynn, Zhigljavsky, (2018), JMA

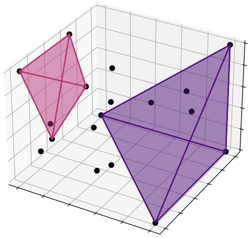


Figure: $k = 3$ dimensional simplices

The k-simplicial distance from x to X is the **average volume of all k-dimensional simplices formed by x and points in X**.

- k-dimensional simplicial distances are of the form $x^\top S_k x$
- 1-dimensional simplicial distance prop. to Euclidean
- d-dimensional/full-dimensional simplicial distance prop. to Mahalanobis
- The volumes can be exponentiated by a value δ.

Simplicial Distances: Computation

Methods of calculating the simplicial distances:

- Direct: average volume of all k-dimensional simplices between x and X,
- For exponent $\delta = 2$: fast polynomial method,
- Compute average volume of a γ% sample of all k-dimensional simplices.

Parameters

- k: simplex dimension
- δ: exponent
- γ: sampling percentage (if using sampling).

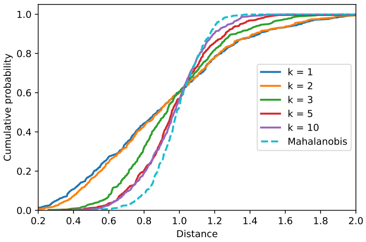


Figure: CDFs of simplicial distances with changing k

Simplicial Distances when $\delta = 2$ (squared volumes)

Covariance matrix Σ with eigenvalues $\Lambda = \{\lambda_1, \lambda_2, \dots \lambda_d\}$, $x^\top S_k x$

Elementary symmetric function in eigenvalues

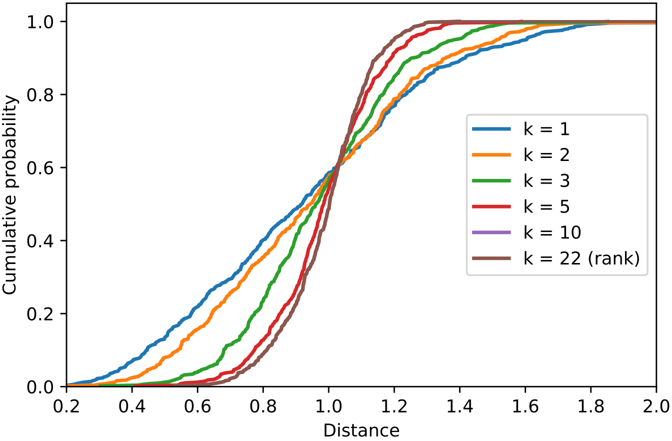
$$
\begin{aligned}
e_k(\Lambda) &= \sum_{1 \le i_1 < i_2 < \cdots < i_k \le d} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_k} \\
q_k(\Sigma) &= \sum_{i=0}^{k-1} (-1)^i e_{k-i-1}(\Lambda) \Sigma^i \\
S_k &= \frac{q_k(\Sigma)}{e_k(\Lambda)}
\end{aligned}
$$

Example, $k = 2$

$$
e_2(\Lambda) = \sum_{1 \le i < j \le d} \lambda_i \lambda_j, \qquad q_2(\Sigma) = \left(\sum_{i=1}^{d} \lambda_i \right) I - \Sigma
$$

Simplicial Distances: Degenerate data

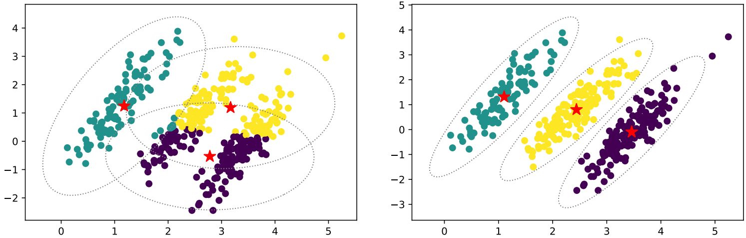
Simulated data with eigenvalues
$\Lambda = \{[100, 100] + [1] * 10 + [0.00001] * 10 + [0] * 28\}$



Simplicial Distances: Clustering

Simplicial distances:

- available when Mahalanobis isn't reliable/available
- can be shown to be more effective than Euclidean/Mahalanobis in simulated/real examples



(a) Euclidean distance

(b) Simplicial distance

Figure: Clustering (projected to two-dimensions)

Minimal-Variance Distances

see e.g. GO'RZ, (2022), JoSTaP

Form

$$x^\top A_k x \text{ where } A_k = \sum_{i=0}^{k-1} \theta_i \Sigma^i$$

Objective

$$\hat{\theta}_k := \mathrm{argmin}_\theta \, \mathrm{V}ar\left[x^\top A_k x\right] + constraint$$

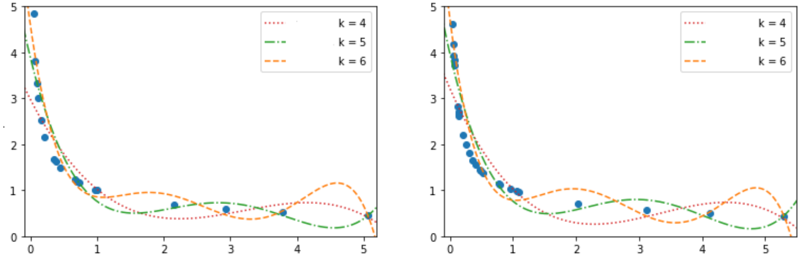
Constraint

Enforce some properties of the Mahalanobis distance
e.g. $\mathrm{trace}(A_k\Sigma) = d$

Solution with above constraint

$$\hat{\theta}_k \propto (V^\top V)^{-1}\left(\mathrm{trace}(\Sigma), \ldots, \mathrm{trace}(\Sigma^k)\right)^\top$$
$$V = (\lambda_j^{i+1})_{j=1,\ldots,d, i=0,\ldots,k-1}$$

Underlying principle: polynomial fit to 1/eigenvalues



(a) $d = 50$ (b) $d = 150$

Clustering example: adjusted rand and purity scores

Higher scores are better

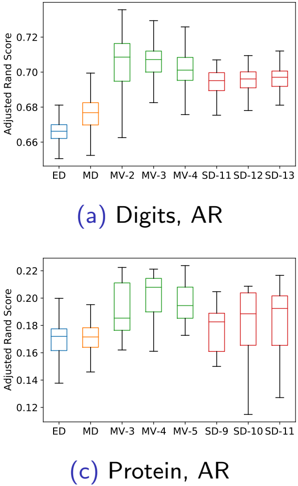


Figure: MV-k: Minimal-variance distance with parameter k, SD-k: Simplicial distance with parameter k.

Outlier labelling example: adjusted rand scores

Dataset	Euclidean	Mahalanobis	Simplicial	Min-Var
Lympho	0.287	0.633	**0.814** ($k = 5$)	**0.814** ($k = 2$)
WBC	0.568	**0.620**	0.568 ($k = 1$)	0.568 ($k = 1$)
Glass	**0.066**	**0.066**	**0.066** ($k = 1$)	**0.066** ($k = 1$)
Vowels	0.142	0.569	0.569 ($k = 11$)	**0.611** ($k = 6$)
Cardio	0.554	0.547	**0.627** ($k = 10$)	**0.627** ($k = 4$)
Thyroid	0.123	0.510	0.491 ($k = 5$)	**0.558** ($k = 5$)
Musk	0.201	**1.000**	**1.000** ($k = 2$)	**1.000** ($k = 2$)
Satimage-2	0.825	0.652	**0.942** ($k = 7$)	**0.942** ($k = 3$)
Letter	-0.012	**0.268**	0.159 ($k = 10$)	0.258 ($k = 10$)
Speech	-0.000	**0.129**	0.016 ($k = 4$)	**0.129** ($k = 5$)
Pima	0.140	0.132	0.145 ($k = 2$)	**0.149** ($k = 2$)
Satellite	0.200	0.349	0.364 ($k = 8$)	**0.395** ($k = 8$)
Shuttle	0.864	0.951	**0.953** ($k = 6$)	0.947 ($k = 6$)
BreastW	**0.863**	0.830	**0.863** ($k = 1$)	**0.863** ($k = 1$)
Arrhythmia	0.333	**0.953**	0.402 ($k = 9$)	0.420 ($k = 9$)
Ionosphere	0.178	**0.743**	0.723 ($k = 7$)	**0.743** ($k = 4$)
MNIST	0.333	0.512	0.418 ($k = 10$)	**0.547** ($k = 8$)
Optdigits	-0.021	-0.028	0.135 ($k = 21$)	**0.207** ($k = 3$)
Cover	-0.010	0.077	0.384 ($k = 5$)	**0.507** ($k = 4$)
Mammography	0.247	0.355	0.347 ($k = 5$)	**0.367** ($k = 5$)
Annthyroid	0.035	**0.318**	0.297 ($k = 5$)	0.305 ($k = 4$)
Pendigits	0.173	0.053	0.372 ($k = 3$)	**0.398** ($k = 2$)
Wine	0.875	0.755	0.875 ($k = 1$)	**1.000** ($k = 4$)

Data obtained from the Outlier Detection DataSets Source (ODDS). Exercise: find labelled outliers.

Minimal-Variance Whitening

Idea

Try to 'approximately' whiten data using $y = A_k^{-1/2} x$

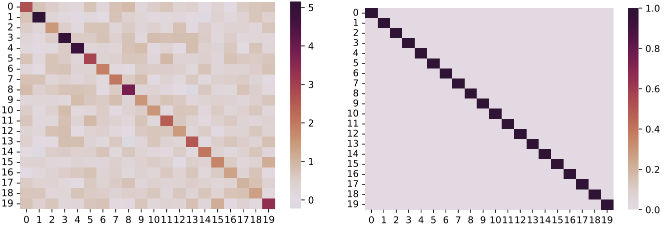
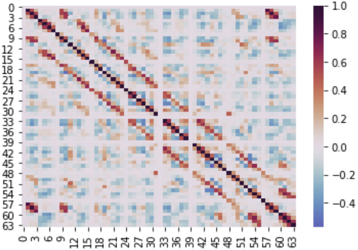
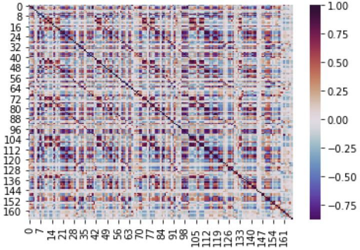


Figure: Heatmaps of a covariance matrix before and after whitening

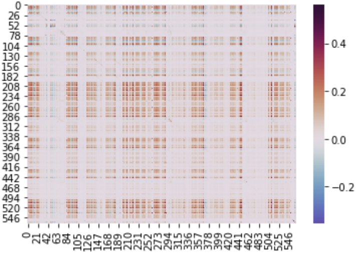
Minimal-Variance Whitening: Before



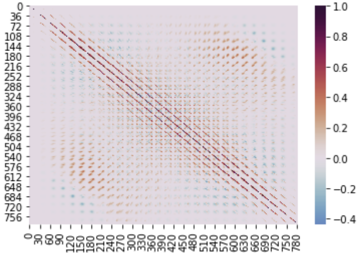
(a) Digits



(b) Musk

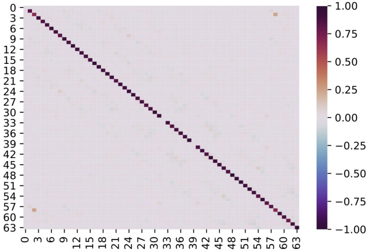


(c) HAR

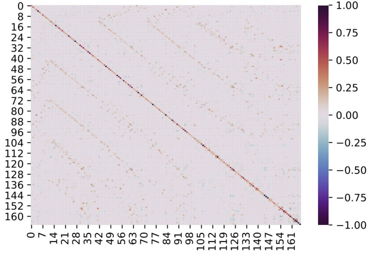


(d) MNIST

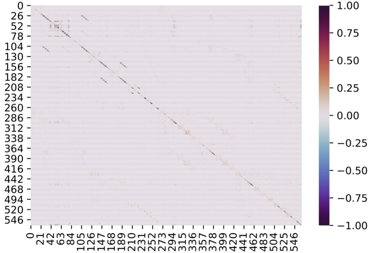
Minimal-Variance Whitening: After



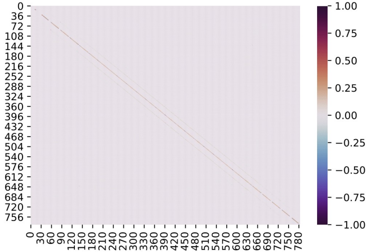
(a) Digits, $k = 7$



(b) Musk, $k = 5$



(c) HAR, $k = 5$



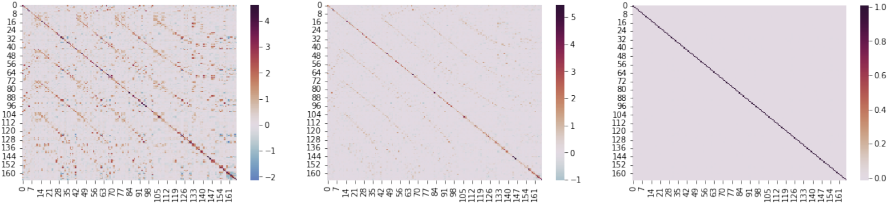
(d) MNIST, $k = 5$

Iterative Minimal-Variance Whitening

Idea

Apply small-k minimal-variance whitening repeatedly e.g

$$y = \dots \left(\tilde{A}_{k_2}^{-1/2}\left(A_{k_1}^{-1/2}x\right)\right)$$



(a) Iter 1 (b) Iter 2 (c) Iter 10

Figure: Iteratively MV-whitened covariance matrix of Musk data, $k = 2$

Thank you for listening

Jonathan Gillard, Emily O'Riordan, and Anatoly Zhigljavsky.
Polynomial whitening for high-dimensional data.
Computational Statistics, 2022.

Jonathan Gillard, Emily O'Riordan, and Anatoly Zhigljavsky.
Simplicial and minimal-variance distances in multivariate data analysis.
Journal of Statistical Theory and Practice, 16(1), 2022.

Luc Pronzato and Anatoly Zhigljavsky.
Measures minimizing regularized dispersion.
Journal of Scientific Computing, 78(3):1550–1570, 2018.

Luc Pronzato, Henry Wynn, and Anatoly Zhigljavsky.
Simplicial variances, potentials and Mahalanobis distances.
Journal of Multivariate Analysis, 168:276–289, 2018.