Multivariate Statistics – Random Projections and Johnson-Lindenstrauss Lemma

Suppose again we have n sample points $x_1, \ldots, x_n \in \mathbb{R}^p$. The data-point $x_i \in \mathbb{R}^p$ can be thought of as the i-th row $X_{i\cdot}$ of an $n \times p$-dimensional data-matrix X.

The previous notes discussed how to embed the points in a lower-dimensional space as $z_1, \ldots, z_q$ with $q < p$.

The proposals so far do not provide an explicit "map" from $\mathbb{R}^p$ to $\mathbb{R}^q$. If we wanted to place a new data-point x_{n+1}, we have to recompute its position in the q-dimensional embedding space. The question is thus whether an explicit map exists from $\mathbb{R}^p$ to $\mathbb{R}^q$ that preserves the distances "well" in some sense?

The Johnson-Lindenstrauss lemma gives a positive answer to this question (under some conditions) and even provides an explicit solution, namely simple random projections. As a drawback, the embedding dimension will typically have to be larger than the 2 or 3 dimensions often used in MDS.

Random projections and the Johnson-Lindenstrauss Lemma. The Johnson-Lindenstrauss lemma states that a collection of n points can be embedded in a lower-dimensionsal space. Moreover, such an embedding can be achieved in practice by random projections in a certain sense.

Let $\Phi \in \mathbb{R}^{p\times q}$ be a matrix whose entries are sampled i.i.d. from a $\mathcal{N}(0,1)$-distribution. Instead of the matrix X we can then work with the "projected" data

$$X \in \mathbb{R}^{n\times p} \qquad \Rightarrow \qquad Z := X\Phi \in \mathbb{R}^{n\times q}. \tag{1}$$

Note that this is not a projection in the mathematical sense, just a mapping $\mathbb{R}^p \mapsto \mathbb{R}^q$ with $q < p$.

The question is then how the distances between the samples $x_1, \ldots, x_n \in \mathbb{R}^p$ (the rows of X) are preserved in the samples $z_1, \ldots, z_n \in \mathbb{R}^q$.

Lemma 1. Johnson-Lindenstrauss Lemma. *Let $\epsilon \in (0, 1/2)$. Let x_i, $i = 1, \ldots, n$ be a collection of n samples in $\mathbb{R}^p$. Set $q \geq 20\log(n)/\epsilon^2$. There exists a Lipschitz mapping $f : \mathbb{R}^p \mapsto \mathbb{R}^q$ such that for all $i, j \in \{1, \ldots, n\}$*

$$(1-\epsilon)\|x_i - x_j\|^2 \;\leq\; \|f(x_i) - f(x_j)\|^2 \;\leq\; (1+\epsilon)\|x_i - x_j\|^2.$$

The proof will be constructive showing that the random projection in (1) is a mapping f that fulfils the Lemma with positive probability. Since the probability of finding a suitable

function f with the random projection is greater than 0, a map f fulfilling the statement of the Lemma must exist.

We show in other words that

$$P_\Phi\Big((1-\epsilon)\|x_i - x_j\|^2 \;\leq\; \|\frac{1}{\sqrt{q}}(X\Phi)_i - \frac{1}{\sqrt{q}}(X\Phi)_j\|^2 \;\leq\; (1+\epsilon)\|x_i - x_j\|^2\Big) \geq \delta > 0$$
$$\text{with } \delta = 1 - 1/\sqrt{n},$$

where the probability is with respect to the multiplication with the random matrix Φ. A random projection thus fulfils the requirement with high probability. This has two implications:

(i) There definitely exists a map fulfilling the requirement (otherwise the probability of finding such a map with random projections would have to be zero).

(ii) We can construct the map by repeating the random projection multiple times. We will then get –with arbitrarily high probability if we allow the number of repetitions to increase– a random projection that fulfils the requirement among these multiple tries (although in practice a single try is often enough as the probability of a failure is bounded by $1/\sqrt{n}$ for a single try which is already very low for large n). The success probability is bounded from below for t repetitions (t new draws of the random matrix Φ) by

$$1 - (\frac{1}{\sqrt{n}})^t.$$

If we want to ensure that there is one random projection among the t repetitions for which the claim in the Lemma is true with probability at least $1 - \delta_0$, we thus need

$$t \geq \frac{\log(1/\delta_0)}{\log(\sqrt{n})}.$$

Before looking at the proof, note that the embedding dimension just depends like $\log(n)$ on the number of samples in the original high-dimensional space (and increases like ϵ^{-2} when the relative error $\epsilon \to 0$). Also note that the embedding dimension q does not depend on the original dimension p. To give an example: embedding $n = 1000$ sample points with a distortion factor of at most 2 ($\epsilon = 1/2$), one needs a dimension $d = 553$ even if the original space had a much larger dimension. MDS has, in contrast, in general no such performance guarantees (and typically a much lower embedding dimension q is used than in random projections).

Proof of the JL lemma. The proof uses the following lemma about the tails of a χ^2-distributed random variable.

Lemma 2. Tails of a χ^2-random variable. *Let $u \in \mathbb{R}^p$. Let $\Phi \in \mathbb{R}^{p\times q}$ be a matrix whose entries are sampled i.i.d. from a $\mathcal{N}(0,1)$-distribution. Then,*

$$P\Big((1-\epsilon)\|u\|^2 \;\leq\; \|\frac{1}{\sqrt{q}}\Phi^t u\|^2 \;\leq\; (1+\epsilon)\|u\|^2\Big) \geq 1 - 2\exp\big(-(\epsilon^2-\epsilon^3)\frac{q}{4}\big). \tag{2}$$

Using this tail bound (proved further below), we can prove the Johnson-Lindenstrauss lemma. Let $f(x_i) = \frac{1}{\sqrt{q}}\Phi^t x_i$. Note that there are $n\cdot(n-1)/2 \leq n^2/2$ pairs $(i,j) \in \{1,\ldots,n\}$. By the union bound

$$\begin{aligned}
&P\Big(\exists (i,j)\in\{1,\ldots,n\} \text{ such that the following inequalities are violated:}\\
&\qquad (1-\epsilon)\|x_i-x_j\|^2 \leq \|\frac{1}{\sqrt{q}}\Phi^t(x_i-x_j)\|^2 \leq (1+\epsilon)\|x_i-x_j\|^2\Big)\\
&\leq 2(n^2/2)\exp\big(-(\epsilon^2-\epsilon^3)\frac{q}{4}\big)\\
&\overset{q\geq 20\log(n)/\epsilon^2}{\leq} n^2\exp\big(-5(1-\epsilon)\log(n)\big)\\
&\overset{\epsilon<1/2}{\leq} \exp\big(-\log(n)/2\big)\\
&\leq 1/\sqrt{n}
\end{aligned}$$

Hence, the random projection (1) generates a mapping f that satisfies the statement in the Lemma with positive probability. A map f that fulfils the statement in the Lemma hence must exist.

Proof of the tail bounds. We still have to show the tail bounds as in (2). Note that

$$W := \frac{\|\Phi^t u\|^2}{\|u\|^2}$$

has a χ^2_q-distribution since each component $(\Phi^t u)_j/\|u\|$ has a $\mathcal{N}(0,1)$-distribution and all components are independent by construction of Φ. Using the union bound over both error types, it is thus enough to show that

$$\begin{aligned}
P\Big(W \geq (1+\epsilon)q\Big) &\leq \exp\Big(-(\epsilon^2-\epsilon^3)\frac{q}{4}\Big)\\
P\Big(W \leq (1-\epsilon)q\Big) &\leq \exp\Big(-(\epsilon^2-\epsilon^3)\frac{q}{4}\Big).
\end{aligned}$$

To prove this, let $V_1,\ldots,V_q$ be i.i.d. $\mathcal{N}(0,1)$ random variables. By Markov's inequality and

$0 < \lambda < 1/2$

$$\begin{aligned} P\Big(W \geq (1+\epsilon)q\Big) =& P\Big(\sum_{i=1}^{q} V_i^2 \geq (1+\epsilon)q\Big) \\ =& P\Big(e^{\lambda \sum_{i=1}^{q} V_i^2} \geq e^{\lambda(1+\epsilon)q}\Big) \\ \leq& \frac{E\Big(e^{\lambda \sum_{i=1}^{q} V_i^2}\Big)}{e^{\lambda(1+\epsilon)q}} \\ =& \frac{E\Big(e^{\lambda V_1^2}\Big)^q}{e^{\lambda(1+\epsilon)q}} \\ =& e^{-\lambda(1+\epsilon)q}\Big(\frac{1}{1-2\lambda}\Big)^{q/2}, \end{aligned}$$

since the expectation is just the moment-generating function of a χ_1^2-distributed random variable. Choosing

$$\lambda = \frac{\epsilon}{2(1+\epsilon)},$$

which minimizes the above expression, we have

$$\begin{aligned} P\Big(W \geq (1+\epsilon)q\Big) \leq& \Big((1+\epsilon)e^{-\epsilon}\Big)^{q/2} \\ \leq& \exp\Big(-(\epsilon^2-\epsilon^3)\frac{q}{4}\Big), \end{aligned}$$

having used the bound $1+\epsilon \leq \exp(\epsilon - (\epsilon^2-\epsilon^3)/2)$. The second bound follows analogously.

Curse of dimensionality. The JL lemma shows that the embedding dimension q can be chosen independently of the original dimension p and the latter can be arbitrarily high. The catch is that for large values of p the Euclidean metric becomes less meaningful in the following sense. Let $X = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$ be a p-dimensional random variable with normalized components in the sense that $E((X^{(k)})^2) = 1$ for all $k \in \{1, \ldots, p\}$. If you look at the distance between two independent draws of X, say X_i and X_j, then

$$\|X_i - X_j\|_2^2 = \sum_k (X_i^{(k)})^2 + \sum_k (X_j^{(k)})^2 - 2 \cdot \sum_k X_i^{(k)} X_j^{(k)}.$$

The latter contribution is of order $\sqrt{p}$, whereas the first two are of order p. Assume for simplicity that $X \sim \mathcal{N}(0, 1_{p\times p})$ independently for all samples. Then,

$$\begin{aligned} (X_i^{(k)} - X_j^{(k)})^2 \sim \mathcal{N}(0,2) \quad \text{for all } k \in \{1,\ldots,p\} \text{ and hence} \\ \frac{1}{2}\sum_k (X_i^{(k)} - X_j^{(k)})^2 \sim \chi_p^2 \end{aligned}$$

If

$$\frac{\log(n)}{p} \to 0,$$

by using a union bound and the previous tail bound for a χ^2-distribution we get that, for $p \to \infty$,

$$\max_{(i,j)\in\{1,\ldots,n\};i\neq j} \left| \frac{1}{2p}\|X_i - X_j\|_2^2 - 1 \right| \xrightarrow{p} 0.$$

Requiring the condition $\log(n)/p \to 0$, which comes from the union bound, means requiring that the number of samples grows sub-exponentially fast with the dimension p.

The above convergence of the maximum implies that among n samples $X_i, \ldots, X_n$, all samples have (with high probability) approximately the same distance to each other (and the concept of nearest neighbour, for example, makes less sense). The same consideration is true if we normalize, that is we scale all X_i such that they have unit norm and are on the sphere in p dimensions.

However, the case considered above is special as all variables are drawn iid: there is no dependence between the variables and the eigenvalues of the population covariance are thus all identically large. If the eigenvalues of the population covariance of X are sparse (a few large and many zero or small – as opposed to the case of identically large values above for iid sampling) then the data really live on a lower-dimensional manifold in p dimensions and the distances can be meaningful again, as the so-called "ambient" dimension p is just hiding that the data live on a lower-dimensional manifold. In the most extreme case, we have just $s \ll p$ non-zero eigenvalues for the population covariance of X and $p - s$ dimensions are then superfluous in the sense that the variance along these dimensions is zero. These $p - s$ dimensions (the null-space of the covariance matrix of X) will not influence the result of a random projection, that is the outcome of the random projection is the same as if we had known the true s-dimensional linear manifold on which the data are distributed.