POINTS OF SIGNIFICANCE

The curse(s) of dimensionality

There is such a thing as too much of a good thing.

Naomi Altman and Martin Krzywinski

We generally think that more information is better than less. However, in the 'big data' era, the sheer number of variables that can be collected from a single sample can be problematic. This embarrassment of riches is called the 'curse of dimensionality'[1] (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity[1,2], multicollinearity[3], multiple testing[4] and overfitting[5]. These effects are amplified by poor data quality, which may increase with the number of variables.

Throughout, we use n to indicate the sample size from the population of interest and p to indicate the number of observed variables, some of which may have missing values for some samples. For example, we may have $n = 1{,}000$ subjects and $p = 200{,}000$ single-nucleotide polymorphisms (SNPs).

First, as the dimensionality p increases, the 'volume' that the samples may occupy grows rapidly. We can think of each of the p variables as an axis in a p-dimensional space. As p increases, any given neighborhood of the space is more likely to contain no data and thus be sparse. For example, in a sample of 1,000 points in a 2D normal distribution, only 6% fall within σ of $(x, y) = (1.5, 1.5)$ (scatter plot in Fig. 1). However, if we consider only one dimension at a time (histograms in Fig. 1), then 31% of the points fall within $\pm\sigma$ of $x = 1.5$ (for the projection on x), and similarly (30%) for y. For $p > 2$, the fraction of points within the p-dimensional sphere of radius σ decreases rapidly: 1.2% at $p = 3$ and 0.2% at $p = 4$.

This increase in sparsity is hard to escape. Even if we move the area of interest in Fig. 1 closer to the mean to have a fixed distance of 1.5 at any number of dimensions (by setting each of the coordinates to $1.5/\sqrt{p}$), the number of points within that area still drops to 14%, 7% and 3% at $p = 2$, 3 and 4, respectively.

Practically, the increase in sparsity makes it much more difficult to collect data that are representative of the population. Consider a simple case of classification or prediction of sample phenotype from genotype. Suppose there are $n = 1{,}000$ samples, each associated with five unlinked SNPs (A, B, C, D and E) that each appear with a minor allele frequency of 10%. If we simply tabulate according to SNP A, we will expect about 900 of the samples to have the major allele A and 100 to have the minor allele a. If we tabulate on two SNPs, A and B, we will expect only ten samples to exhibit both minor alleles with genotype ab. With SNPs A, B and C, we expect only one sample to have genotype abc, and with four or more SNPs, we expect empty cells in our table. We need a much larger sample size to observe samples with all the possible genotypes. As p increases, we may quickly find that there are no samples with similar values of a predictor.

Even with just five SNPs, our ability to predict and classify the samples is impeded because of the small number of subjects that have similar genotypes. In situations where there are many gene variants, this effect is exacerbated, and it may be very difficult to find affected subjects with similar genotypes and hence to predict or classify on the basis of genetic similarity.

If we treat the distance between points (e.g., Euclidian distance) as a measure of similarity, then we interpret greater distance as greater dissimilarity. As p increases, this dissimilarity increases because the mean distance between points increases as $\sqrt{p}$ (Fig. 2a). This effect is stark at high values of p. For example, at $p = 100$, the closest pair of points are farther from one another than the most distant two points are for about $p < 15$ (Fig. 2a, horizontal dashed line).

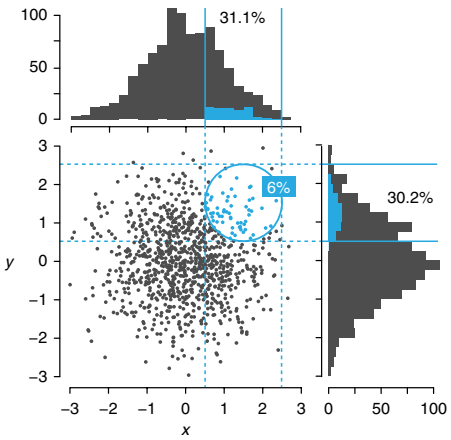

Fig. 1 | Data tend to be sparse in higher dimensions. Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. $\sigma = 1$, only 6% fall within σ of $(x, y) = (1.5, 1.5)$ (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins within blue solid lines) are within σ of 1.5. Blue bins in histograms correspond to the blue points.

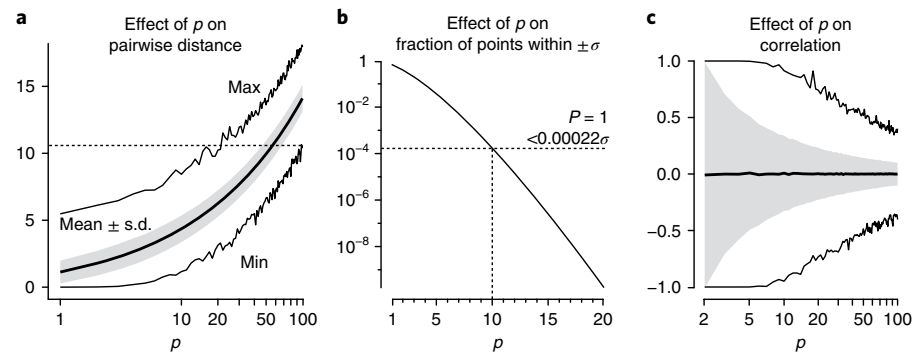

Fig. 2 | As the number of variables p increases, distances between points grow rapidly and correlations decrease. The plots show results for points sampled from a p-dimensional normal distribution with a mean of 0 and s.d. $\sigma = 1$ along each dimension. **a**, The average pairwise distance between two points increases as $\sqrt{p}$ and is about $2\sigma\sqrt{(p/2)}$ for large p. Shown are the minimum and maximum distances observed among 10,000 points (thin lines), as well as their mean and s.d. (thick black line and gray shaded region, respectively). **b**, The fraction of points within σ of the mean drops rapidly with increasing p. **c**, A decrease in the range of correlation between two random vectors with increasing dimensions. Lines and shading indicate the minimum, maximum, mean and s.d. as defined in **a**.

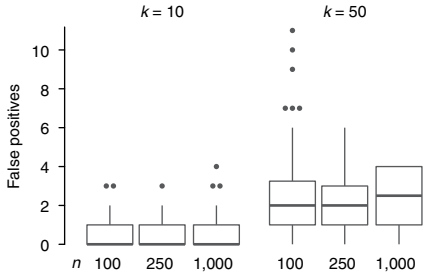

Fig. 3 | The number of false positives increases with each additional predictor. The box plots show the number of false positive regression-fit *P* values (tested at $\alpha = 0.05$) of 100 simulated multiple regression fits on various numbers of samples ($n = 100$, 250 and 1,000) in the presence of one true predictor and $k = 10$ and 50 extraneous uncorrelated predictors. Box plots show means (black center lines), 25th and 75th percentiles (box edges), and minimum and maximum values (whiskers). Outliers (dots) are jittered.

The assessment of whether a particular data value is an outlier accordingly also becomes difficult[2]. For example, 68% of normally distributed points in one dimension fall within σ of the mean, but this fraction drops off precipitously for large values of p (Fig. 2b). For example, at $p = 10$, the fraction of points within σ of the mean is only 0.017%, which is equivalent to the points within 0.00022σ in one dimension (Fig. 2b, dashed lines). Putting it another way, points within σ at $p = 10$ are as rare as points outside of 3.8σ at $p = 1$. These are not intuitive observations—proportions of how points are distributed at higher dimensions are not the same as in one dimension.

What if we use a different distance measure to express similarity between subjects, such as correlation? Unfortunately, we cannot escape the CoD—the range of correlations between points drops rapidly (Fig. 2c) with p. For example, at $p = 100$, among 10,000 random pairs of points we see no correlations greater than 0.5, and most are extremely tightly grouped around 0. We can understand this analogously to the five-SNP example above. As the number of variables increases, the number of subjects in each set of categories decreases and the correlation between any two subjects across variables also decreases.

When the number of dimensions is larger than the number of samples ($p > n$), another effect that confounds analysis is perfect multicollinearity[2]. In this case, we can always express at least one of the variables as a linear combination of the others. For example, if we have $p = 3$ variables but only $n = 2$ subjects and we think of the subjects as two 3D vectors, then from linear algebra we know that these vectors define either a point (i.e., they have the same three values) or a line. In both cases, the three variables are related linearly. This is the case any time there are fewer samples than dimensions—the variables span a lower-dimensional subspace in which some of the dimensions become 'redundant' and expressible in terms of other dimensions, thus yielding perfect multiple correlation.

This kind of correlation among variables is problematic when they are used for prediction, because it means that interpretation of the prediction equation will be uncertain. For example, suppose that we have measured three metabolites, *X*, *Y* and *Z*, that affect the level of a hormone H. Also suppose that there is a fourth metabolite, *U*, that is perfectly multiply correlated with the other three—say, $U = X + 2Y + 3Z$. Because we can solve for the level of any three of the metabolites in terms of the fourth, we need to use only three of the metabolites to predict the hormone level. However, it is impossible to understand the impact of each of the metabolites individually. Any set of three of the four, or all four, can be used to predict the hormone level, and the data do not tell us which, if any, of the metabolites is important.

We have previously discussed the problems of multiple testing in the context of testing whether each variable has a significant effect on the response—for example, testing phenotypic association for each SNP[4]. For this type of problem, we have seen that standard *P*-value cutoffs of $P < 0.05$ (or $P < 0.01$) generate far too many false positives. Typically, we instead resort to controlling the family-wise error rate or the false discovery rate to give some assurance that most of our discoveries are true positives. Adjusting for multiple testing requires more stringent testing criteria, which vastly reduces the test's power, leading to high false negative rates—another CoD.

We might also consider multiple testing in the context of using many variables as predictors, such as for regression or classification. In this context, we might want, for example, to test the contribution of each SNP to the prediction equation. We illustrate this scenario in Fig. 3, which demonstrates multiple linear regression of an output variable *Y* in the presence of one true predictor (X_0) and *k* extraneous predictors ($X_1, \ldots, X_k$) for *n* samples. We sample all predictors from a normal distribution with a mean of 0 and an s.d. of 1 and set $Y = X_0$ in each case. Thus, we expect a regression coefficient for X_0 with a low *P* value and nonsignificant *P* values for the coefficients of $X_1, \ldots, X_k$ predictors, as only X_0 affects *Y*. Among the *k* regression coefficients, we expect αk false positives if we test at $P < \alpha$. Importantly, the number of false positives is not mitigated by an increase in sample size. For example, for $k = 50$ extraneous predictors, we do not see significantly fewer false positives at $n = 1{,}000$ than we do at $n = 100$.

Finally, overfitting is another CoD that occurs because the flexibility of prediction equations[5] is in part determined by the number of variables involved. With increased flexibility, prediction and classification rules adapt to both the patterns in the population and the random idiosyncrasies of the training sample.

In general, it is preferable to have more data rather than less when exploring scientific questions. However, the proliferation of data that may be unrelated to the question(s) of interest lead to the CoD, which hinders our ability to detect real relationships and patterns. Dimension-reduction methods such as variable selection and principal component analysis[6] can help to reduce dimensionality, but may themselves be affected by the CoD.

Although improved statistical and machine learning methods and larger sample sizes can partially mitigate the CoD, nothing replaces expertise. We need to carefully distinguish between exploratory studies in which a large number of variables possibly related to the process may be examined, and confirmatory studies in which a more focused set of variables with reduced dimensions is used for detailed scientific discovery. ❐

Naomi Altman[1] and Martin Krzywinski[2]*
[1]*Professor of Statistics at The Pennsylvania State University, University Park, PA, USA.*
[2]*Staff scientist at Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada.*
**e-mail: martink@bcgsc.ca*

Published online: 31 May 2018
https://doi.org/10.1038/s41592-018-0019-x

References

1. Bellman, R. E. *Adaptive Control Processes: A Guided Tour* (Princeton Univ. Press, Princeton, NJ, 1961).
2. Zimek, A., Schubert, E. & Kriegel, H.-P. *Stat. Anal. Data Min.* **5**, 363–387 (2012).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 1103–1104 (2015).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
5. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 703–704 (2016).
6. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **14**, 641–642 (2017).

Competing interests

The authors declare no competing interests.