HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 August 08.

Published in final edited form as:

Nat Methods. 2018 April ; 15(4): 233–234. doi:10.1038/nmeth.4642.

Statistics versus machine learning

Danilo Bzdok [Assistant Professor],
Department of Psychiatry, RWTH Aachen University, Germany, and a Visiting Professor at INRIA/Neurospin Saclay in France

Naomi Altman [Professor], and
Statistics at The Pennsylvania State University

Martin Krzywinski [staff scientist]
Canada's Michael Smith Genome Sciences Centre

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns. Two major goals in the study of biological systems are inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease. Prediction makes it possible to identify best courses of action (e.g., treatment choice) without requiring understanding of the underlying mechanisms. In a typical research project, both inference and prediction can be of value—we want to know how biological processes work and what will happen next. For example, we might want to infer which biological processes are associated with the dysregulation of a gene in a disease, as well as detect whether a subject has the disease and predict the best therapy.

Many methods from statistics and machine learning (ML) may, in principle, be used for both prediction and inference. However, statistical methods have a long-standing focus on inference, which is achieved through the creation and fitting of a project-specific probability model. The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a 'true' effect that is unlikely to result from noise. Furthermore, if enough data are available, we can explicitly verify assumptions (e.g., equal variance) and refine the specified model, if needed.

By contrast, ML concentrates on prediction by using general-purpose learning algorithms to find patterns in often rich and unwieldy data[1,2]. ML methods are particularly helpful when one is dealing with 'wide data', where the number of input variables exceeds the number of subjects, in contrast to 'long data', where the number of subjects is greater than that of input variables. ML makes minimal assumptions about the data-generating systems; they can be effective even when the data are gathered without a carefully controlled experimental design and in the presence of complicated nonlinear interactions. However, despite convincing

Competing Interests: The authors declare no competing interests.

prediction results, the lack of an explicit model can make ML solutions difficult to directly relate to existing biological knowledge.

Classical statistics and ML vary in computational tractability as the number of variables per subject increases. Classical statistical modeling was designed for data with a few dozen input variables and sample sizes that would be considered small to moderate today. In this scenario, the model fills in the unobserved aspects of the system. However, as the numbers of input variables and possible associations among them increase, the model that captures these relationships becomes more complex. Consequently, statistical inferences become less precise and the boundary between statistical and ML approaches becomes hazier.

To compare traditional statistics to ML approaches, we'll use a simulation of the expression of 40 genes in two phenotypes (−/+). Mean gene expression will differ between phenotypes, but we'll set up the simulation so that the mean difference for the first 30 genes is not related to phenotype. The last ten genes will be dysregulated, with systematic differences in mean expression between phenotypes. To achieve this, we assign each gene an average log expression that is the same for both phenotypes. The dysregulated genes (31–40, labeled A–J) have their mean expression perturbed in the + phenotype (Fig. 1a). Using these average expression values, we simulate an RNA-seq experiment in which the observed counts for each gene are sampled from a Poisson distribution with mean $\exp(x + \varepsilon)$, where x is the mean log expression, unique to the gene and phenotype, and $\varepsilon \sim N(0, 0.15)$ acts as biological variability that varies from subject to subject (Fig. 1b). For genes 1–30, which do not have differential expression, the z-scores are approximately $N(0, 1)$. For the dysregulated genes, which do have differential expression, the z-scores in one phenotype tend to be positive, and the z-scores in the other tend to be negative.

Our goal in the simulation is to identify which genes are associated with the abnormal phenotype. We'll formally test the null hypothesis that the mean expression differs by phenotype with a widely used generalized linear negative binomial model that allows for biological variability among subjects with the same phenotype. We'll perform a test for each gene and identify those that show statistically significant differences in mean expression, based on P values adjusted for multiple testing via the Benjamini–Hochberg method[3]. In an alternative Bayesian approach, we would compute the posterior probability of having differential expression specific to the phenotype.

Figure 2a shows the P values of the tests between phenotypes as a function of the log fold change in gene expression. The ten dysregulated genes are highlighted in red; our inference flagged nine out of the ten (except F, with the smallest log fold change) as significant with adjusted $P < 0.05$. We could use the fold change as a measure of effect size, with a confidence interval or highest posterior region used to indicate the uncertainty in the estimate. In a realistic setting, genes identified by the analysis would then be validated experimentally or compared with data from other sources such as proposed gene networks or annotations.

To ask a similar biological question using ML, we would typically try several algorithms evaluated by cross-validation on independent test subjects, or bootstrap methods with ‘out-

of-sample' evaluation[4] to select one with good prediction accuracy. Let's use a random forest (RF) classifier[5] that will simultaneously consider all genes and grow multiple decision trees to predict the phenotype without assuming a probabilistic model for the read counts. The result of this RF classification with 100 trees is shown in Figure 2b, where the *P* values from the classical inference are plotted as a function of feature (gene) importance. This score quantifies a given gene's contribution to the average classification improvement[5] within a partition when the tree is split selecting that gene. Many ML algorithms have analogous measures that allow some quantification of the contribution of each input variable to the classification. In our simulation, eight of ten genes with the largest importance measures were from the dysregulated set. Not in the top ten were genes D and F, which had the smallest fold changes (Fig. 2a).

If we perform the simulation 1,000 times and count the number of dysregulated genes correctly identified by both approaches—on the basis of either classical null-hypothesis rejection with an adjusted *P* value cutoff or predictive pattern generalization with RF and top-ten feature importance ranking—then we find that the two methods yield similar results. The average number of dysregulated genes identified is 7.4/10 for inference and 7.7/10 for RF (Fig. 2c). Both methods have a median of 8/10, but we find more instances of simulations for which only 2–5 dysregulated genes were identified with inference. This is because the way we've designed the selection process is different for the two approaches: inference selects by an adjusted *P* value cutoff so that the number of selected genes varies, whereas in the RF we select the top ten genes. We could have applied a cutoff to the importance score, but the scores do not have an objective scale on which to base the threshold.

We've used pre-existing knowledge about RNA-seq data to design a statistical model of the process and draw inference to estimate unknown parameters in the model from the data. In our simulation, the model encapsulates the relationship between the mean number of reads (the parameter) for each gene for each phenotype and the observed read counts for each subject. The output of the statistical analysis is a test statistic for a specific hypothesis and confidence bounds of the parameter (mean fold change, in this example). In our example, the model's parameters relate explicitly to aspects of gene expression—the counts of molecules produced at a certain rate in a cell can be directly interpreted.

To apply ML, we don't need to know any of the details about RNA-seq measurements; all that matters is which genes are more useful for phenotype discrimination based on gene expression. Such generalization greatly helps when we have a large number of variables, such as in a typical RNA-seq experiment that may have hundreds to hundreds of thousands of features (e.g., transcripts) but a much smaller sample size.

Now consider a more complex experiment in which each subject contributes multiple observations from different tissues. Even if we only conduct a formal statistical test that compares the two phenotypes for each tissue, the multiple testing problem is greatly complicated. The increase in data complexity may make classical statistical inference less tractable. Instead we could use an ML approach such as clustering of genes or tissues or both to extract the main patterns in the data, classify subjects, and make inferences about the

biological processes that give rise to the phenotype. To simplify the analysis, we could perform a dimension reduction such as averaging the measurements over the ten subjects with each phenotype for each gene and each tissue.

The boundary between statistical inference and ML is subject to debate[1]—some methods fall squarely into one or the other domain, but many are used in both. For example, the bootstrap[6] method can be used for statistical inference but also serves as the basis for ensemble methods, such as the RF algorithm. Statistics requires us to choose a model that incorporates our knowledge of the system, and ML requires us to choose a predictive algorithm by relying on its empirical capabilities. Justification for an inference model typically rests on whether we feel it adequately captures the essence of the system. The choice of pattern-learning algorithms often depends on measures of past performance in similar scenarios. Inference and ML are complementary in pointing us to biologically meaningful conclusions.

References

1. Bzdok D. Front Neurosci. 2017; 11:543. [PubMed: 29056896]
2. Bzdok D, Krzywinski M, Altman N. Nat Methods. 2017; 14:1119–1120. [PubMed: 29664466]
3. Krzywinski M, Altman N. Nat Methods. 2014; 11:355–356.
4. Lever J, Krzywinski M, Altman N. Nat Methods. 2016; 13:603–604.
5. Altman N, Krzywinski M. Nat Methods. 2017; 14:933–934.
6. Kulesa A, Krzywinski M, Blainey P, Altman N. Nat Methods. 2015; 12:477–478. [PubMed: 26221652]

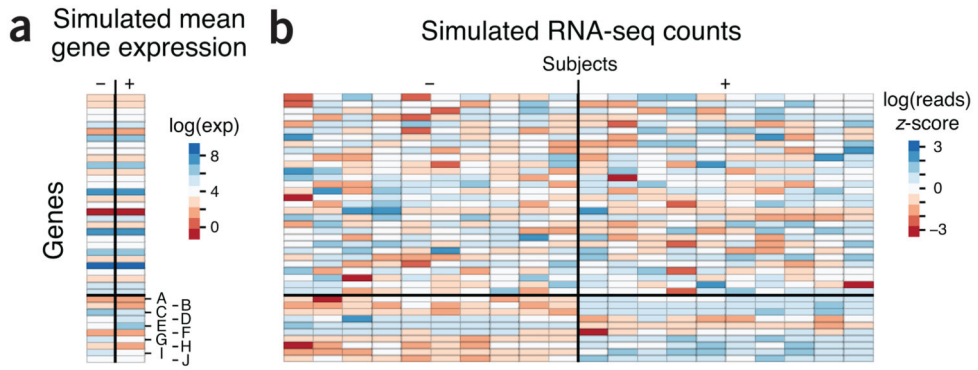

Figure 1.

Simulated expression and RNA-seq read counts for 40 genes in which the last 10 genes (A–J) are differentially expressed across two phenotypes (−/+). Simulated quantities and heat maps are log-scaled. (**a**) Simulated log mean expression levels for the genes generated by sampling from the normal distribution with mean 4 and s.d. 2. In the + phenotype the differential expression of genes A–J was created by the addition of a standard normal to each mean expression in the – phenotype. (**b**) The simulated RNA-seq read counts for ten subjects in each phenotype generated from an overdispersed Poisson distribution based on mean expression in a with biological variation. The heat map shows *z*-scores of the read counts normalized across all 20 subjects for a given gene.

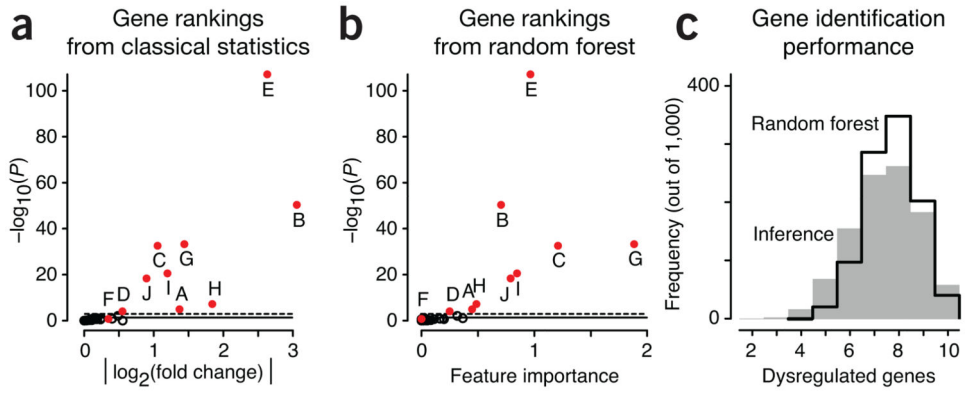

Figure 2.

Analysis of gene ranking by classical inference and ML. (**a**) Unadjusted log-scaled *P* values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (**b**) Log-scaled *P* values from **a** as a function of gene importance from random forest classification. In **a** and **b**, red circles identify the ten differentially expressed genes from Figure 1; the remaining genes are indicated by open circles. (**c**) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).