Methods for false discovery control

Statistical Learning

The seminal paper in which Benjamini & Hochberg introduced the concept of FDR has changed thinking about multiple testing quite radically, showing that FWER control is not only way to do of multiple testing, and stimulating the field of multiple testing enormously.

Compared to FWER control, the subject of FDR control is relatively young. Much method-development is still ongoing, and some important questions are still partially open. This holds especially for the complicated situation of dependent p-values that is so important for applications in genomics research. In this paper, we leave aside the extensive literature on FDR control for independent p-values, and focus only on results that are known or believed to be valid under fairly general forms of dependence. We follow the same structure as for FWER-based methods, discussing methods that are generally valid, methods valid under the assumption that Simes inequality holds, and methods based on permutations. For FDR, unlike for FWER, the Simes-based method is the oldest and best known one, so we start there.

Benjamini & Hochberg

The Benjamini & Hochberg procedure is a step-up procedure just like the Hochberg procedure, only with higher critical values. It compares each ordered p-value $p_{(i)}$ with the critical value $i\alpha/m$, finds the largest j such that $p_{(j)}$ is smaller than its corresponding critical value, and rejects the j hypotheses with the j smallest p-values.

It has been shown that this procedure controls FDR at level α under the technical assumption of *positive regression dependence on a subset*, which is essentially the same assumption that underlies Hochberg's method, namely that the Simes inequality holds for the subset of true hypotheses. In fact, control of FDR under these assumptions is even at level $\pi_0\alpha$, where $\pi_0 = m_0/m$.

The close connection between the Benjamini & Hochberg method and Simes inequality is immediately obvious from the fact that the Benjamini & Hochberg and Simes critical values are identical. Weak control of FWER, a necessary condition for FDR control, for the Benjamini & Hochberg method follows immediately from Simes' inequality.

The critical values of the Benjamini & Hochberg procedure are much larger than those of Hochberg or Hommel, and consequently many more rejections can be made. Clearly, without changing the assumptions, relaxing the criterion from FWER to FDR can make a huge difference in terms of power.

A gain in power of Benjamini & Hochberg's method relative to Hochberg's, and in general of FDR-based versus FWER-based methods is most pronounced when many false hypotheses are present. This can be understood by comparing the FDR and FWER criteria. In FDR, the more hypotheses are rejected, the higher the denomimator of the false discovery proportion Q, and the less stringent the error criterion for the next rejection becomes.

The Benjamini & Hochberg method, like Bonferroni, controls its error rate at level $\pi_0\alpha$, rather than at α.

This suggests the possibility an alternative, more powerful Benjamini & Hochberg procedure that uses critical values $i\alpha/(\hat{\pi}_0 m)$ rather than $i\alpha/m$ if a good estimate $\hat{\pi}_0$ of π_0 would be available. Such a procedure might have an FDR closer to the chosen level α, and would be even more powerful than the original procedure if many hypotheses were false. Such procedures are called *adaptive* procedures, and many have been proposed based on various estimates of π_0.

A problem with the adaptive approach, however, is that estimates of π_0 can have high variance, especially if p-values are strongly correlated. Naive plug-in procedures, in which this variance is not taken into account, will therefore generally not have FDR control, especially if $\pi_0 \approx 1$.

More sophisticated methods are needed that do take the estimation error of π_0 into account. One such procedure, by Benjamini, Krieger and Yekutieli, adjusts the α-level slightly from α to $\alpha^* = \alpha/(1+\alpha)$ to adjust for the additional variance from estimation of π_0. This procedure estimates π_0 by first performing an initial Benjamini & Hochberg procedure at the slightly reduced level α^*, estimating π_0 by $\hat{\pi}_0 = (m - R_0)/m$, where R_0 is the number of rejections obtained in this first step. In the second and final step, a subsequent Benjamini and Hochberg procedure is done at level $\alpha^*/\hat{\pi}_0$. Note that, unlike simpler plug-in procedures, this latter procedure is not guaranteed to give more rejections than the regular, non-adaptive Benjamini & Hochberg procedure, since $\alpha^*/\hat{\pi}_0$ may be smaller than α. This reflects the additional risk incurred in estimating π_0.

FDR control for the adaptive Benjamini, Krieger and Yekutieli procedure has only yet been proven under independence, although simulations suggest FDR control under positive dependence as well. In any case, adaptive procedures are expected to have increased power over the ordinary Benjamini & Hochberg procedure only of the proportion π_0 of true null hypotheses is substantially smaller than 1. If π_0 is near 1, the power of such procedures may actually be worse.

FDR control under general dependence

The equivalent to the Benjamini & Hochberg procedure that is valid even when the conditions for Simes' inequality does not hold is the procedure of Benjamini & Yekutieli. This procedure is linked to Hommel's variant of the Simes inequality in the same way that the procedure of Benjamini & Hochberg is linked with Simes inequality itself.

It is a step-up procedure that compares each ordered p-value $p_{(i)}$ with the critical value $i\alpha/(m\sum_{j=1}^{m} 1/j)$, finds the largest j such that $p_{(j)}$ is smaller than its corresponding critical value, and rejects the j hypotheses with the j smallest p-values. Like Hommel's inequality relative to Simes, the Benjamini & Yekutieli procedure is more conservative than the Benjamini & Hochberg procedure by a factor $\sum_{j=1}^{m} 1/j$. Like Hommel's inequality, the Benjamini & Yekutieli procedure is valid under any dependency structure of the p-values.

The method of Benjamini & Hochberg is guaranteed to reject at least as much as Hochberg's procedure, which uses the same assumptions but controls FWER rather than FDR. The same does not hold for the method of Benjamini & Yekutieli relative to Holm's method, which is the standard method for FWER control under any dependency of the p-values.

FDR control by resampling

An FDR controlling method with the power, simplicity and reliability of the method of Westfall & Young has not yet been found. Research in this area is ongoing.

Use of FDR control

As we have seen from the example data, FDR control is usually much less conservative than FWER control. Control of FDR, since that criterion is concerned with the proportion of type I errors among the selected set, is more suitable for exploratory genomics experiments than FWER control. FDR control methods do a very good job in selecting a set of hypotheses that is promising, in the sense that we can expect a large proportion of the ensuing validation experiments to be successful. FDR has effectively become the standard for multiple testing in genomics. Nevertheless, FDR control has been criticised. It is helpful to review some of this criticism in order to understand the properties and the limitations of FDR control better.

FDR is the expected value of FDP, which is a variable quantity because the rejected set $\mathcal{R}$ is random. It has been pointed out, however, that the actual value of FDP can be quite variable. Sets $\mathcal{R}$ found by a method that controls FDR often have an FDP that is much larger than α, or one that is much smaller than α, and the realized FDP for a method controlling FDR at 0.05 can, for example, be greater than 0.29 more than

10% of the time. The variability of FDP around FDR is not taken into account in FDR control methods, and this variability is not quantified.

Unlike FWER control, FDR control is a statement about an average over the full set $\mathcal{R}$ only. The fact that a set $\mathcal{R}$ is rejected by a method with FDR control does not imply anything about hypothesis $H \in \mathcal{R}$ or about subsets $\mathcal{S} \subset \mathcal{R}$. The lack of this subsetting property has several consequences that have to be taken into account when working with the results of FDR controlling procedures.

One consequence that has frequently been mentioned is the possibility of 'cheating' with FDR. This cheating can be done as follows. If a researcher desires to reject some hypotheses using FDR, he or she can greatly increase the chances of doing so by testing these hypotheses together with a number of additional hypotheses which are known to be false, and against which he or she has good power. The additional guaranteed rejections alleviate the critical values for the hypotheses of interest, and make the rejection of these hypotheses more likely. The catch of this approach is that the resulting FDR statement is about the rejected set including the added hypotheses, and that no FDR statement may, in fact, be made about the subset of the rejected set that excludes the added hypotheses. The cheating as described above is blatant, of course, and would hardly occur in this way. More often, however, inadvertent cheating of the same type occurs, for example when a list of rejected hypotheses is considered but the obvious, and therefore uninteresting, results are discarded or ignored, when an individual rejected hypothesis is singled out, or when subsets of the rejected hypotheses are considered for biological reasons. If hypotheses are very heterogeneous (e.g. Gene Ontology terms rather than probes) it is difficult not to look at subsets when interpreting the results on an analysis.

A second consequence of FDR's lack of the subsets property relates to the interpretation of adjusted p-values. The adjusted p-value, being the largest α-level at which the hypothesis is rejected by the multiple testing procedure, is usually interpreted as a property of the hypothesis itself. For FWER-adjusted p-values this is warranted, as FWER control for the rejected set implies FWER control for each individual hypothesis within the set. FDR control is different, however: because it is a property of the whole set only, not of individual hypotheses within the set, the adjusted p-value similarly is a property of the whole rejected set, not of any individual hypothesis within that set. It is, therefore, hazardous to use such adjusted p-values as properties of individual hypotheses. To see this, consider a hypothesis with an FDR-adjusted p-value just below 0.05. If this hypothesis was the only hypothesis rejected at this α-level, we can be confident that it this hypothesis is not a type I error. If, on the other hand, 20 or more hypotheses were rejected at the same level, the same adjusted p-value can signify anything. In one extreme, it could be that the top 19 hypotheses are absolutely certain rejections, and that the 20th hypothesis, even though a type I error with high probability, was added just because the FDR level of 0.05 left enough room for an extra rejection. Clearly, interpreting FDR-adjusted p-values as properties of single hypotheses is problematic; an FDR-adjusted p-value is a property of a rejected set, not of an individual hypothesis.

FDR controlling methods are most useful in exploratory situations in which a promising set of hypotheses must be found, and when we are content to use the set of hypotheses with top p-values, or something close to that set. In such situations these methods are very powerful. If FDR control is used in a final reported analysis, the result of this analysis is the full rejected set, and researchers should be careful when making statements on subsets or individual hypotheses within that set.

References

- Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine 2014 33:1946-78