RESEARCH STATEMENT

Eugene Katsevich

From my undergraduate work in computed tomography [1, 2] and cryo-electron microscopy [3, 4] to my graduate work in statistical genetics [5, 6], I have always had a keen interest in applying my quantitative skills to biomedical applications. Along the way, several technical challenges stemming from these applications have inspired me to think more generally about methodological and theoretical problems in statistics. In my PhD, I have investigated ways to extract patterns from complex data sets while providing replicability guarantees. In particular, I have designed multiple testing and variable selection methodologies that account for the *structure* of modern data sets and the *exploration* which often goes into analyzing them.

Current Work: Structure and Exploration in Multiple Testing

The data collected in a variety of fields is increasingly rich and complex, which creates novel opportunities and challenges for statisticians. New vast data sets collected in fields like finance, biology, astronomy, and social science promise insights about the molecular mechanisms of disease, the structure of the universe, and everything in between. On the other hand, the scale and complexity of these modern data sets makes it challenging to unlock these insights. One way of handling this complexity is to use the data itself to *generate* hypotheses, scanning for interesting patterns to examine more closely. While exploratory data analysis is a useful hypothesis-generating tool that has been in use for decades, it is increasingly important to additionally provide *replicability guarantees* for the hypotheses flagged for follow-up. Indeed, the look-elsewhere effect paired with the richness of modern data sets creates ample room for false discoveries.

I have used multiple testing, and in particular *false discovery rate* (FDR) control, as a formal way of encoding replicability. Consider the universe of hypotheses $\mathcal{H} = (H_1, \ldots, H_m)$ that can be encoded in a data set $\mathcal{D}$; the goal is to find a subset $\mathcal{R}^*$ of these that are supported by the data. Given p-values $\boldsymbol{p} = (p_1, \ldots, p_m)$ derived from $\mathcal{D}$, traditional methods to control the FDR—like the Benjamini-Hochberg (BH) procedure—consider sets of the form $\{j : p_j \leq t\}$ and choose a cutoff $t^* \in [0, 1]$ such that the FDR is bounded at a pre-defined level q. On the other hand, the hypotheses in $\mathcal{H}$ might have extra *structure* (e.g. spatial or graphical), which does not align with BH's "exchangeable" treatment of hypotheses. I have developed methods to search for rejection sets fitting into the context of this structure while preserving FDR guarantees. Moreover, traditional methods like BH leave no room for data scientists to *explore*, since modifying the contents of an FDR-controlling set post hoc is formally prohibited. I have worked on reconciling exploration with rigorous Type-I error guarantees, bridging the gap between the theory of statistical inference and the practice of data science.

Next, I outline three of my projects that exemplify how to integrate structure and exploration into multiple testing.

Controlled variable selection at multiple resolutions

Genome-wide association studies (GWAS), enabled by high-throughput genotyping technology, scan the genome for associations with a trait or disease of interest. Genetic measurements are made at the "high resolution" of single nucleotide polymorphisms (SNPs). However, SNPs have correlation patterns obeying the one-dimensional structure of the genome, making nearby SNPs hard to tell

apart. They are also structured into larger functional units such as genes. Therefore, SNPs are often analyzed in groups ("lower resolution") to facilitate statistical power and/or interpretability.

While traditional GWAS methodology relies on marginal testing, the recently developed knockoffs methodology [7] provides an attractive alternative. Knockoffs provide a means to test the more meaningful hypotheses of *conditional* independence between a response and a set of predictor variables and provide rigorous FDR control guarantees. Like BH, however, knockoffs do not accommodate for structure on the set of variables. For example, the set of significant SNPs returned by the knockoffs procedure cannot be grouped while retaining the FDR guarantee, since this operation can inflate the FDR.

This problem motivated me to develop the Multilayer Knockoff Filter (MKF) [5] with my advisor Chiara Sabatti. The MKF is a variable selection methodology that finds a high-resolution set of significant variables whose projections into multiple pre-specified lower resolutions control the FDR at each resolution at given target levels. For example, one might want to control the FDR at the SNP level *and* at the gene level. The MKF searches for rejection regions jointly across all resolutions (similar to the p-filter [8]) and leverages knockoff statistics to measure variable importance. The principal technical challenge lay in proving FDR control for multiple rejection sets (one per resolution) that are all coupled together in a complicated way. The key to the proof was to assume a worst-case scenario where the rejection thresholds at each resolution are chosen adversarially, and then bound this worst-case FDR by constructing an appropriate exponential supermartingale and applying the maximal inequality. Surprisingly, the price of this pessimistic analysis was an extra constant of only 1.93 in the FDR bound.

I applied the MKF procedure to a targeted exome re-sequencing data set to study associations with HDL cholesterol. Cross-referencing the results with the literature on this well-studied trait, MKF reduced the number of false positive genes from 5 (out of 11 total) to just 1 (out of 6 total), at the cost of one extra false negative. I am currently in the process of applying MKF and related ideas to the exome-wide Finnish Metabolic Sequencing (FinMetSeq) and UK Biobank data sets, and integrating this methodology into the existing `knockoff` package in `R`. I have presented this methodology to genetics audiences at conferences (including this year at the American Society for Human Genetics), where it has been well received, including a best student poster award at a statistical genomics conference in 2017.

Controlling FDR while filtering discoveries

Grouping the elements of a rejection set is just one example of a *filtering* operation. Many other, more complicated, filtering operations are also common in data science, especially when hypotheses are structured. For example, International Classification of Diseases (ICD) codes, used in electronic health records and insurance claims, have a tree structure reflecting relationships between diseases (e.g. "pneumococcal meningitis" is more specific than "bacterial meningitis", so there is an edge from the latter to the former). In this and other applications involving hypotheses of varying degrees of specificity, filtering is common to reduce redundancy and improve interpretability of rejection sets. In the context of ICD codes, for instance, a rejection $i \in \mathcal{R}^*$ might be considered "redundant" if it has a descendant $j \in \mathcal{R}^*$. The *outer nodes filter* [9] might be employed to remove this redundancy, leaving a set of "distinct" discoveries $\mathcal{U}^* \subseteq \mathcal{R}^*$. Other filters can be more involved and take the form of software packages.

Like changing the resolution of discoveries, many filtering operations also run the risk of inflating the FDR. Therefore, applying an FDR procedure followed by a filter is in general a dangerous

operation. This presents a challenge, especially because a variety of filters may be applied in practice and it is infeasible to design a new FDR methodology for each filter. To address this challenge in full generality, I first formalized the concept of a filter as any mapping

$$\mathfrak{F} : (\mathcal{R}, \boldsymbol{p}) \mapsto \mathcal{U}, \text{ such that } \mathcal{U} \subseteq \mathcal{R}.$$

The outer nodes filter defined above might be one example of such an $\mathfrak{F}$. With this definition, a reasonable inferential goal is to control the *false filtered discovery rate*:

$$\mathrm{FDR}_{\mathfrak{F}} = \mathbb{E}\left[\mathrm{FDP}(\mathcal{U}^*)\right] = \mathbb{E}\left[\mathrm{FDP}(\mathfrak{F}(\mathcal{R}^*, \boldsymbol{p}))\right] \leq q.$$

In practice, it is often the case that the filter to be applied can be specified in advance. Therefore, in collaboration with Chiara Sabatti and Marina Bogomolov, I proposed Focused BH [6], an extension of the BH procedure that accounts for the effect of a pre-specified filter $\mathfrak{F}$ to control the above error rate. To prove FDR control, one must account for the interaction of the filter with the dependency structure of the p-values. In particular, I showed that Focused BH controls the FDR when the filter $\mathfrak{F}$ is *monotonic* filter and the p-values are PRDS (a kind of positive dependence).

I extensively tested Focused BH across a variety of simulation settings, including hypotheses with tree, DAG, and spatial structures. In the case of tree-structured hypothesis testing with the outer nodes filter, I demonstrated that Focused BH controls the FDR under weaker assumptions and is more powerful than Yekutieli's procedure [9] targeting the same error rate. In the case of spatially-structured GWAS hypotheses, I showed in simulations that Focused BH controls the FDR after a clumping filter, while the filter-blind BH suffers a substantial FDR inflation.

High-probability FDP bounds after exploration

In the context of Multilayer Knockoff Filter and Focused BH, note that the operations applied to the rejection set were required to be specified *before* seeing the data (pre hoc). In certain cases, like the ones described above, this is a reasonable assumption. In other cases, it is important for data scientists to participate in the choice of a rejection set $\mathcal{R}^*$ *after* seeing the data (post hoc), leveraging their domain knowledge and intuition. Consider the following scenario: a data scientist applies BH at level $q = 0.05$, and it turns out that only two discoveries were made. She then tries BH at level $q = 0.1$, which yields ten discoveries. The extra eight discoveries seem promising, so all ten are reported as significant at FDR level 0.1. While in theory it is clear that this amounts to "data snooping" and invalidates the FDR guarantee, in practice the expectation that the target level q is set a priori is not always reasonable.

To allow data scientists to explore their data while retaining replicability guarantees, Aaditya Ramdas and I have proposed a *simultaneous selective inference* approach [10]. The user is presented with a data-dependent "menu" (path) of nested rejection sets

$$\varnothing = \mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_n \subseteq \mathcal{H}$$

with accompanying FDP bounds $\overline{\mathrm{FDP}}(\mathcal{R}_k)$, which under p-value independence are simultaneously valid across k with high probability:

$$\mathbb{P}\left[\mathrm{FDP}(\mathcal{R}_k) \leq \overline{\mathrm{FDP}}(\mathcal{R}_k) \text{ for all } k\right] \geq 1 - \alpha.$$

The user can inspect this menu and choose a rejection set whose content and FDP bound is to her liking, *the FDP bound of the chosen set retaining validity despite the user's data-dependent*

decision. This approach is related to exploratory multiple testing [11], which by simultaneously bounding the FDP of all subsets $\mathcal{R} \subseteq \mathcal{H}$ allows the scientist to choose from this exponentially large menu of options. While this all-subsets approach relies on closed testing, I obtain simultaneous FDP bounds across a path of rejection sets by studying $\text{FDP}(\mathcal{R}_k)$ as a stochastic process in k. At the cost of providing a more modest (though perhaps more focused) menu of options to the data scientist, I show in simulations that simultaneous selective inference can yield much tighter bounds on the FDP. Therefore, in terms of power and flexibility, simultaneous selective inference is a compromise between selective inference (guarantees for one rejection set) and simultaneous inference (guarantees for all possible rejection sets).

Short-Term Research Agenda: Correlations and Resampling

Much of my work has focused on handling complex structures in the context of multiple testing. Another kind of complexity in this context that has not been adequately addressed is the issue of correlation among p-values. It is the easiest to prove results in the unrealistically optimistic case of independence or in the unrealistically pessimistic case of arbitrary dependence. Aside from the somewhat mysterious PRDS condition, not too much has been done in the middle ground between these two extremes. Some of the most promising work in this direction has been based on resampling. I believe resampling is a powerful methodology to deal both with complex structures and dependency patterns, and I see several fruitful directions in which to expand these ideas.

Permutation-based false (filtered) discovery rate control

While permutation-based methods were actively developed in the context of the stringent family-wise error rate (FWER) [12], satisfactory methodology and theory for permutation-based FDR control is still lacking. I believe that work in this direction could lead to powerful multiple testing methodology that works under realistic and verifiable dependence assumptions and does not rely on conservative corrections. A possible place to start is a permutation-based version of Focused BH I proposed, which while currently lacking theory performed quite well in simulations, boosting power while retaining FDR control. Exploring this procedure and developing theory for it could lead to a promising way to account for p-value correlations as well as complicated filters.

Variable selection via resampling

Given a collection of m random variables $X_1, \ldots, X_m$ and a response variable Y, consider testing the conditional independence hypotheses $H_j : X_j \perp\!\!\!\perp Y | X_{-j}$ for $j = 1, \ldots, m$. This variable selection problem is known to be hard, especially in high dimensions. In the "model-X framework" where we have knowledge of the distribution of X, the *conditional randomization test* [13] (resampling X_j from its distribution conditional on X_{-j}) is a simple and elegant way to test conditional independence. One might wonder whether it would be possible to modify it to be more robust to its strong model-X assumption. The conditional permutation test [14] is a step in that direction, though is still relies fairly heavily on knowing the distribution of X and is even more computationally costly than the CRT. I would like to improve upon the CRT and CPT to develop a variable selection methodology that is lighter on computations and assumptions.

Accelerating resampling-based procedures

One of the biggest drawbacks of resampling-based procedures is their computational cost. Especially if permutation p-values are subjected to multiple testing corrections, they must be quite accurate, and therefore require more computation. I would like to work on significantly reducing the computational burden of resampling-based procedures. Progress in this direction would help the statistical advantages of these procedures outweigh their computational disadvantages.

Long-Term Research Agenda: Precision Medicine

The sequencing of the human genome in 2001 created high expectations for our ability to understand the biological mechanisms of human disease, paving the way for better and more personalized prognosis, diagnosis, and treatment. Nearly two decades later, the promise of "precision medicine" is still far from reality. My career goal is to move us closer to realizing that promise by tackling quantitative challenges relevant to biology and medicine. The following are two examples of research in this direction.

Functional genomics and causality

The emerging field of functional genomics picks up from where association studies leave off: trying to explain the mechanism by which a given genetic variant leads to a disease. While association analysis looks for correlations, functional genomics looks for causal explanations. I would love to contribute to the rapid development of the fields of functional genomics and causal inference to advance our mechanistic understanding of human diseases. Causal inference is related to what I have already done since replicability is the hallmark of a causal effect. Moreover, my work on variable selection is adjacent to causality in that it deals with testing independence while conditioning away confounders. Testing such conditional hypotheses helps disentangle direct effects of predictors on outcomes from indirect effects through other predictors. I am interested in exploring these connections and learning more about causal inference.

Inference from electronic health records

In addition to genomics, another increasingly large source of biomedical data is electronic health records (EHR). Especially when linked with genomic data, EHR data contain much promise for medical insights. Given the heterogeneiety and pervasive missingness in EHR data, however, unlocking these insights presents significant statistical challenges. Moreover, EHR data are highly structured and thus amenable to analysis with some of the techniques I have already developed. I am interested in learning more about EHRs and applying my experience with the analysis of structured data to this rich data source.

Conclusion

The unifying theme of my work is identifying and addressing statistical challenges arising from biomedical applications. I believe that the best way for me to make a difference in biomedicine is to collaborate closely with scientists in this domain, pairing my statistical expertise with their domain knowledge. I am excited to learn from these scientific collaborators as well as from my colleagues in math statistics, and I hope that standing at the intersection of these two domains will allow me to achieve my goal of advancing precision medicine.

References

[1] **E. Katsevich**, A. Katsevich, and G. Wang. Stability of the interior problem for polynomial region of interest. *Inverse Problems*, 28(6), 2012.

[2] B. Shi, **E. Katsevich**, B. Chiang, A. Katsevich, and A. Zamyatin. Image registration for motion estimation in cardiac CT. In *SPIE Medical Imaging*, San Diego, California, February 2014.

[3] **E. Katsevich**, A. Katsevich, A. Singer. Covariance matrix estimation for the cryo-EM heterogeneity problem. *SIAM Journal on Imaging Sciences*, 8(1):126–185, 2015.

[4] J. Anden, **E. Katsevich**, and A. Singer. Covariance estimation using conjugate gradient for 3D classification in cryo-EM. In *IEEE Int Symp Biomed Imaging*, New York, New York, April 2015.

[5] **E. Katsevich** and C. Sabatti. Multilayer Knockoff Filter: Controlled variable selection at multiple resolutions. *Annals of Applied Statistics*, to appear, 2018.

[6] **E. Katsevich**, C. Sabatti, and M. Bogomolov. Controlling FDR while highlighting distinct discoveries. In preparation, 2018+.

[7] R. F. Barber and E. J. Candes. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 2015.

[8] R. F. Barber and A. Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statical Society, Series B*, 2017.

[9] D. Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the Americal Statistical Association*, 2008.

[10] **E. Katsevich** and A. Ramdas. Towards "simultaneous selective inference:" post-hoc bounds on the false discovery proportion. *Annals of Statistics*, in revision, 2018+.

[11] J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 2011.

[12] P. H. Westfall and S. S. Young. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, 1993.

[13] E. J. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statical Society, Series B*, 2018.

[14] T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test. Preprint, 2018.