COVER LETTER

Zhiguang (Caleb) Huo (zhh18@pitt.edu)

My research statement contains the following topics: my background, training and experiences; my research interest, highlight my research accomplishment and why my research is important; my view of the field and future research plan.

My background is a little bit complex. I had a Bachelor's and Master's degree in physics, which turned to provide me a solid foundation of mathematics and capability of thinking problems via modeling. During my Ph.D training, I have devoted myself in learning statistics/machine learning, which turned out to be useful tools for solving real problems. I was exposed to a lot of data because of UPMC. I had traveled and learn from multiple workshops.

My research interest lies in both statistical methodology and application. I had solid training in statistics, machine learning, computing. Biostatistics/bioinformatics is a fast moving field and new types of data are coming out frequently. It is exciting that I can develop data-driven methodology, which in turn will enhance the data analysis. I have explored and motivated by classes lectures/workshops. I want to focus the following aspect of research: meta-analysis, data integration, high-dimensional, Bayesian, optimization, computing, software development. Below I described my past, on going research and future research plan.

1 Statistical methodology

I am lucky that I had my PhD training at University of Pittsburgh, where I was exposed to tons of data from UPMC. This gave an opportunity to motivate Statistical methodology ideas. Pitt student are allowed to take class from both PITT/CMU. And therefore I am lucky to take class from varies top statistics/machine learning/optimization researchers. I learning the cutting edge tools and use them to solve the problem in my hand. Below I will describe my research interest based on my past research, on-going project and future plan.

1.1 Genomic data meta analysis

Large amount of genomic data are publicly available. Meta-analysis aims to combine multiple same-type genomic data to increase statistical power, accuracy and validated conclusion. I have been working in this area and developed several statistical approach to combine genomic data.

- Meta Sparse Kmeans is meta analysis of clustering analysis. The algorithm aims to define disease subtype from multiple genomics cohort. The algorithm will achieve feature selection and guarantee clustering pattern across cohorts simultaneously.
- BayesMP is meta analysis approach to detect differential expression genes and characterize differential meta-pattern. The algorithm uses Bayesian non-parametric approach to describe p value distribution. The meta-pattern is helpful to detect homogeneous and heterogeneous differential gene patterns across cohorts.
- AW computing a meta-analytic approach to combine p-values. By using importance sampling and spline, we obtain a fast computing for AW, which is weighted fisher.
- Involvement: metaDCN, metaPCA.

1.2 Genomic data integration

Genomic data integration combines multi-omics data (e.g. gene expression, CNV, genotyping, methylation, somatic mutation, miRNA) of the same cohort. We could gain statistical power and have a better understanding of the inner omics relationships.

- IS-Kmeans. Integrative sparse Kmeans. The purpose is to combine multi-omics data, achieve feature selection and incorporating prior group information simultaneously.
- Partially involved in Bayesian group lasso problem.

1.3 High dimensional data optimization and statistical computing

Introduction of high dimensional data. High dimensional data processing

- I have a solid training in optimization. I am quite familiar with optimization algorithms, especially lasso related problems. My research projects are highly involved with optimization (ref). The first example: The second example ADMM to solve the overlapping group lasso problem.
- github. Github is originally design for programmers. It is also very powerful for statistician. This is an easy way to host my code, program and packages. In the future, I will also make my teaching materials on gitbub.
- high-dimensional theory, this is equally important as optimization. I had pursue this direction by taking several related classes. I was attracted by the beauty of proof and its impact on the problem itself. This may be my future direction.

1.4 Bayesian inference and graphical model

Bayesian is very convenient to model complex data structure, get soft inference on result.

- Bayesian MP: this is Bayesian non-parametric.
- Single cell imputation, this is part of my thesis, I used to conditional random field to impute.
- Partially involved in Bayesian group lasso.

2 Bioinformatics application

As a biostatistician, an important job is to work with local biologist on their data. This is also a good resource for me to think and develop methodologies. This is a fast field and I have been exposed to multiple types of data. Though data are different and processing pipeline may differnet, but the enssence of statistical approach remain the same.

2.1 Data types

- Microarray, CNV. Problems: DE analysis, subtype discovery. ref prostate cancer. psycharitry
- RNAseq, mutation, fusion genes.
- single cell data (methylation, expression).
- In the future, brain imaging data...

2.2 Software development

Purpose is to facilite biologist to use statistical softwares.

- Meta Omcis R package.
- Meta Omcis suit, implementing using R shiny.