Research statement

Vadim Zipunnikov

I am interested both in real data applications, especially in brain imaging and wearable computing, and in advancing statistical theory and methodology. I strongly believe that work on applications is incredibly exciting and leads to a more informed selection of problems for methodological research. Deep understanding of statistical principles leads to better applied work. My research interests include ultra high-dimensional data, methods development for functional and imaging data analysis motivated by real data applications, multilevel models, and nonlinear time series. My current research agenda is motivated by large (terabyte-sized) longitudinal brain imaging studies. This new structure of observational studies led to numerous statistical and computational challenges that I have addressed during my post-doctoral studies at Johns Hopkins University. My diverse statistical and applied interests have arisen from unique opportunities during my doctoral studies at Cornell University and post-doctoral studies at Johns Hopkins University. The deep methodological and applied insights I have gained at both institutions have complemented my mathematical background obtained during my undergraduate studies in Moscow. I have found and embraced areas of research where hard mathematical and statistical concepts are used to resolve real life problems. In a technology-savvy world, the role of Statistics is constantly being re-defined. It is my intention to be at the core of these changes and help guide them using sound and principled methodological thinking. Below I provide specific areas of research I have worked on.

Research at Cornell University. During my doctoral studies I have been working on several open problems related to generalized linear mixed models and their applications. My research resulted in three papers described below.

In the *"Counting Tables using the Double Saddlepoint Approximation"* paper, I proposed a double saddlepoint approximation to calculate the number of two- and multi-way contingency tables with counts satisfying certain linear constraints. This has multiple applications including exact tests for categorical data. The suggested higher-order approximation was far superior to other previously proposed analytical approximations as well as highly accurate in a range of examples, including some for which analytical approximations were previously unavailable.

In the *"Closed form GLM cumulants and GLMM fitting with SQUAR-EM-LA$_2$ algorithm"* paper, I found closed form formula for the standardized cumulants of generalized linear models. This reduced the complexity of necessary calculations from $O(p^6)$ to $O(p^2)$ operations which allowed efficient construction of second-order saddlepoint approximations to the pdf of sufficient statistics. I adapted the result to obtain a closed form expression for the second-order Laplace approximation for a GLMM likelihood and developed a highly efficient accelerated EM procedure. Extensive simulation studies showed the excellent accuracy and speed of the approach even in the cases when the random effect was very high-dimensional.

In the *"Sparse Spherical-Radial EM algorithms for GLMM fitting"* paper, I suggested a computational procedure that exploited a partial linearity of a GLMM likelihood with respect to a random effect. I adapted spherical-radial representation of likelihood integrals and approximated them with quadrature points on the high-dimensional unit sphere that belonged to a linear hull of a dimension significantly lower than that of the sphere. This resulted in a sparse coverage of the sphere that both delivered a greater accuracy of the approximation and dramatically reduced the computational burden. The approach significantly outperformed competing fitting procedures including those based on Quasi-Monte Carlo.

Research at Johns Hopkins University. My research at Johns Hopkins University is concerned with multilevel and longitudinal imaging studies and their parsimonious modeling. My research program addresses various challenges posed by the volume and dimensionality of these type of data. I have worked with ultra-high dimensional brain images including RAVENS maps (brain morphology) and Diffusion Tensor imaging (water diffusion). Learning about these new

types of data and problems has been extremely rewarding. In particular, they have opened an entirely new area for my methodological research.

The motivation for my first project came from the analysis of longitudinal brain volumes (RAVENS maps). The data arose from a population-based study of lead exposure and its relationship to brain structure and function. In the *"Multilevel Functional Principal Component Analysis for High Dimensional data"* paper I developed a framework that combines powerful data compression techniques and statistical inference to decompose the observed multilevel data in population- and visit-specific means and subject-specific within and between level variability. The proposed algorithms are linear in the dimensionality of the data and do not require loading the data matrix into memory, an impossible task in many modern imaging applications.

Motivated by a study of multiple sclerosis patients, my second project has focused on modeling longitudinal studies of high-dimensional data, such as brain images, collected at multiple visits over irregular periods of time. In *"Longitudinal High Dimensional Data Analysis"* paper I generalized models designed for longitudinal data analysis to the case when the observed data are massively multivariate. The developed HD-LFPCA framework decomposes the longitudinal functional/imaging data into a subject specific, longitudinal subject specific, and subject-visit specific components. The proposed methods are computationally linear in the dimension of the functions or images. Essentially, the methods now allow modeling data that thought to be prohibitively large for statistically principled analysis.

In *"Functional principal component model for high-dimensional brain imaging"* paper the primary interest was characterizing and quantifying the parcellation of brain atlases. I showed a novel connection between Singular Value Decomposition and an FPCA model. Based on that, I developed efficient computing techniques that were applied to a morphometric data set including seven hundred RAVENS images. This allowed us to identify principal directions of morphometric variation in a population of organo-lead exposed workers.

All the methods developed by me at Hopkins are designed around the high efficiency/low resources philosophy, which led to implementations that take minutes on a regular laptop to analyze gigabytes of data.

Current and future research projects. I am currently leading several research projects. In the first project, I consider two classes of algorithms for HD-LFPCA. The first approach, streaming, recursively updates principal components every time a new part of data becomes available. For instance, if a new group of patients is scanned streaming avoids refitting a high dimensional model. The second approach estimates principal components via an accelerated EM method. In a related project, I am advancing the HD-LFPCA framework to perform bootstrap and cross-validation procedures for high dimensional data in a highly efficient way.

A general method for simultaneous dimensionality reduction of large populations of massive images was proposed in *"Population value decomposition, a framework for the analysis of images"*. The methodology was motivated by and applied to the Sleep Heart Health Study, the largest community-based cohort study of sleep containing more than 85 billion observations. Together with Genevera Allen of Rice University, we are now working on statistical foundations for PVD using matrix-variate distributions. We are particularly interested in exploring a lack of invariance of PVD to permutation of matrix elements and propose a few procedures offering a smaller reconstruction error. To introduce a sparse analogue of PVD, we investigate different penalization schemes for the objective function.

My interest in Statistical methods is guided by new problems that I work on in collaborative scientific teams. In addition to the brain imaging studies, I have recently become involved in wearable computing. More precisely, I am interested in design of experiments and activity prediction using three-axis acceleration data. Another area of interest is complex time series obtained from EEG monitoring during sleep. In general I am interested in simple methods based on sound statistical knowledge that provide fast inferential solutions to the most complex data.