Research Statement

Jie Peng

There are two major topics of my current research, one is statistical genetics/genomics based on high throughput data, and the other is functional data analysis based on sparsely observed noisy measurements over time. The concepts and tools from high-dimensional inference play important roles in both fields of my research. I will briefly describe my work in these two areas. Throughout, I will discuss some ongoing works and my immediate research plan. I will also mention several specialized softwares that I developed, which are important for the reproducibility of my research.

1 Statistical genetics/genomics

Many biological studies aim to identify molecules whose activities influence certain phenotypes of a living organism. It is of great scientific interests to study how these molecules regulate the phenotypes as well as their interactions with each other, and with external environmental factors. This knowledge facilitates our understanding of the underlying complex mechanisms defining these phenotypes.

My earlier works, including my doctoral dissertation, focus on statistical genetics, in particular, quantitative trait loci (QTL) mapping in humans. Most of these works aim at studying issues in pedigree-based linkage analysis, such as selective phenotyping through ascertainment ([1], [4]) and gene-covariate interactions ([3]). In these studies, we assume single-locus models and calculate efficient score statistics at each marker locus. Scan statistics are then derived by taking maximum over hundreds or thousands of markers (depending on the marker density). The significance level and power of the corresponding testing procedures are approximated using tools developed for stochastic processes ([3], [2]). Although a lot of valuable information has been obtained by such studies, these methods may not be very powerful for complex traits that are defined by many genes, with each gene having a small effect. Moreover, besides sequence level variants such as single nucleotide polymorphisms (SNPs), there are other biological molecules such as RNA and protein which may play more direct role in shaping these phenotypes. These aspects motivated me to study the functional associations of these biological molecules. Consequently, my recent works focus more on developing statistical tools for elucidating the underlying relationships among different types of biological molecules based on high throughput genomic data. In [5], we develop a method for expression-QTL (e-QTL) mapping in linkage studies. In particular, the model takes into account the complicated relationships among the expression levels. This greatly reduces false positive discoveries of trans-regulatory QTLs by excluding indirect regulations caused by interactions among expression levels. Partly motivated by this study, in [11], we propose a model to estimate the concentration network for p variables based on n samples under the n << p setting, commonly termed as the high-dimension-low-sample-size setting. This is equivalent to identifying the non-zero partial correlations for a
large number of variables based on a (relatively) small sample. This method is used to build regulatory networks for RNA expression levels in a breast cancer study.

Since different types of biological molecules provide complimentary information in the characterization of genetic regulatory relationships, it is natural to build genetic regulatory networks (GRNs) based on multiple types of genomic data. Also, according to the central dogma of molecular biology, DNA operates at a lower level than RNA for encryption of the information in cell regulations and in turn RNA operates at a lower level than protein. Thus the models should also take into account the biological principles which suggest directionality of these regulations. To model the relationships between two sets of variables, in [12], we propose a method for fitting multivariate regression models when both the number of predictors and the number of responses are larger than the sample size. We apply this method to a breast cancer data set consisting of CGH measurements (sequence level data) and expression data. This results in an estimated regulatory network of DNA copy number alterations on RNA expression levels. This method is also readily applicable to many other types of genomic studies involving two types of biological molecules, such as the e-QTL studies.

Roughly, all three works ([5], [11], [12]) fall under the regularized regression framework with high-dimensional predictors and responses. Suitable regularization schemes are employed and computationally efficient algorithms are developed accordingly. In [11], we also prove consistency of the proposed estimator in terms of model selection and parameter estimation under a regime of diverging model dimension (i.e., \( p \) goes to infinity with \( n \)). We are currently extending these ideas to develop statistical methods for building GRNs based on three levels of genomic data: namely, DNA level data such as SNPs or CGH measurements, RNA expressions and protein expressions. Although such data are still scarce in human studies, they are available for some model organisms such as yeast.

2 Functional data analysis

My research in functional data analysis focuses on modeling the underlying common stochastic mechanisms for a group of subjects (random curves). In particular, it mainly concerns developing suitable methods when individual curves are only measured on an irregular, and typically sparse set of time points, usually contaminated by measurement noise. Such data are often referred to as longitudinal data and are very typical in biological studies, sociological studies, functional genomics studies, among others. My research in this area has two major themes - one is based on the exploration of the phenomenological aspects of the data through functional principal components analysis (FPCA), and the other is on using differential equations to model the dynamical aspects of the data.

The eigenfunctions of the covariance kernel of a stochastic process (referred to as functional principal components) give a natural basis for representing functional data, and hence are very useful in problems related to model building and prediction. In [10] and [8], we adopt a restricted maximum likelihood (REML) approach for estimating the functional principal components. In [10], we develop efficient procedures for model fitting and model selection by explicitly utilizing the intrinsic geometry of the parameter space. In [8], we have established consistency and derived the rate of convergence of these REML estimators. Such results for REML estimators in the sparse functional data context are new in statistical literature. Our work also provides useful techniques for estimation problems in semi-parametric models that are based on minimization of a loss function (e.g. M-estimators) over a non-Euclidean manifold, which is a field with
comparatively few existing results. My other works in the area of FPCA include: (i) In [7], we develop a distance-based clustering method for sparsely observed functional data and apply it to classify on-line auction bidding behaviors; and (ii) In [6], we consider estimating functional principal components for correlated random curves. The results obtained in this paper are relevant for many practical circumstances where the sample curves are correlated, for example, time course gene expression data and spatio-temporal data.

Even though FPCA is a very powerful tool for exploring functional data, it could be restrictive in the sense that we assume the curves to be random realizations (either independent or correlated) of a stochastic process. However, if the underlying phenomena are governed by a nonlinear dynamical system, an FPCA approach could be inefficient and even inappropriate. Instead, an ordinary differential equation (ODE) model with subject-specific effects can easily incorporate nonlinearity and can result in more interpretable descriptions of the underlying phenomena. In [9], we propose a semi-parametric model for fitting autonomous nonlinear dynamical systems based on sparsely observed sample trajectories. Our work makes new contributions to the area of estimating continuous time dynamical systems in that, most existing works assume known functional forms of the dynamical system and/or require dense measurements. In our model, the estimation problem is an ill-posed inverse problem. Thus the analysis differs from standard semi-parametric function estimation problems and requires development of new analytical tools. We are currently working on extending this framework to model non-autonomous systems through a varying coefficient approach.

The above methods are also applicable to functional data measured on a dense grid. However, in order to deal with sparseness, an important common feature of all these methods is to efficiently combine information across subjects (under a semi-parametric framework). The techniques of both model fitting and theoretical analysis borrow strength from and contribute to the developments of high-dimensional inference.

3 High-dimensional inference

In this section, I will discuss the commonality of my works under the theme of high-dimensional inference. As mentioned earlier, my research in high-dimensional inference is mainly motivated by two areas: (i) genetics/genomics studies where high throughput data are used to derive genetic regulatory networks; and (ii) functional data analysis, where sparse and noisy measurements taken over a time (or spatial) domain are used to elucidate the underlying stochastic mechanisms. In spite of differences in motivations and applications, these high-dimensional inference problems bear some intrinsic similarities.

From a theoretical standpoint, we are dealing with models whose complexity grows with the amount of data available. For example, in [11], the (nominal) model dimension is diverging to infinity with the sample size, possibly at a faster rate. One of the key features of high-dimensional inference is that, typically, consistent estimation of the parameters of interest is possible only under appropriate restrictions on the parameter space. This is because large nominal dimension results in excessive overall variability of the estimates. Such restrictions are often imposed through some form of sparseness conditions, and the resulting (regularized) estimators are often solutions of a constrained optimization procedure. This is the case in many of the works mentioned above (e.g., [11], [8]). Consistent model selection under sparse settings is also an important aspect of some problems ([11]). A technique that has been commonly used for analyzing such estimators is to first reduce the problem to a neighborhood of the “optimal
parameter” in the restricted parameter space by performing appropriate expansions of the penalized loss function, and then quantifying the variability of the estimator through the use of suitable probabilistic bounds, including large deviation inequalities and results from random matrix theory. The bias is quantified through studying the complexity of the restricted parameter space. Apart from establishing consistency, refined asymptotics of the estimators have also been established in some cases (for example, derivation of an efficient score representation in [8]) through detailed analysis of the geometry of the parameter space.

From a statistical modeling viewpoint, as already indicated, regularization is a key ingredient of all these problems in order to efficiently utilize the data as well as to obtain meaningful estimates of model parameters. The choice of suitable regularization schemes depends heavily on the type of underlying structures we envision, and it should also take into account both model interpretability and computational tractability. In particular, regularization schemes should be calibrated in such a way that, the limited resources are optimally used to bring out the most important features of the problem (which of course depend on specific goals of the study). In [5], [11], [12], regularizaton schemes are designed to: (i) explicitly impose sparsity constraints since genetic regulatory relationships are believed to be parsimonious; (ii) efficiently utilize special structures of the underlying models such as the existence of well-connected nodes (hubs). These not only improve efficiency, but also enhance the interpretability of the resulting models and the chance of identifying the most important model components (in these cases, the hub nodes). In [10], a reduced rank formulation (which amounts to a sparsity constraint) of the covariance kernel is used to fit functional principal components. This framework avoids over-parametrization and helps in estimating the leading eigenfunctions. It can also be viewed as a form of regularization of the likelihood surface. In [9], we aim to estimate the gradient function of a dynamical system, and impose smoothness constraints on it through a basis representation approach. This is more efficient than approaches based on estimation of the trajectories and their derivatives through FPCA where smoothness is imposed directly on sample curves. This is because the trajectories of a nonlinear ODE (i.e., the flow of the dynamics) form a nonlinear manifold and consequently FPCA (or more generally a linear basis expansion) of the sample curves would require many basis functions.

From a computational viewpoint, the fitting techniques for these models often involve optimization over complex spaces. This is the case when either the observations or the parameters reside in a high-dimensional space or a non-Euclidean manifold. Consequently, developing and implementing suitable computational tools are critical aspects of these studies. In [11], [12], a fast iterative algorithm, active-shooting, is invented and applied to fit the models, and is shown to increase computational efficiency of a commonly used existing method shooting (a.k.a. Pathwise Coordinate Optimization) by many folds for high-dimensional sparse models. In [10], an optimization tool developed for the Stiefel manifold is applied to FPCA. This method takes into account the intrinsic geometry of the parameter space and consequently results in more efficient estimates than methods that ignore the geometrical structure. In [9], nonlinear optimization techniques and numerical ODE solvers are combined to fit the model. A common technique adopted by all these methods involves suitably grouping the unknown parameters and performing iterative updates by cycling through different groups. Finally, these methods are more generally applicable beyond the scope of these papers. For example, active-shooting can be readily extended to solve many other penalized optimization problems. The development and analysis of optimization tools on special manifolds have proven to be important components of my research for various other problems involving geometric
Determination of the amount of regularization is another important issue for all these methods. This is often referred to as model tuning, or under some context, simply as model selection. There are two major challenges in terms of model tuning when dealing with high-dimensional data. One is the choice of a suitable criterion function which meets the goal of the study, and the other is the computational feasibility of the tuning procedures. In [6] and [10], we use leave-one-curve-out cross-validation (CV) scores based on empirical predictive Kullback-Leibler loss for model selection. In literature, CV scores based on $\ell_2$ prediction error loss are often used in FPCA for sparse data. A first order expansion of the difference between the averaged loss function under the true and estimated parameters shows that Kullback-Leibler loss correctly scales the difference, while the prediction error loss does not ([6]). In [12], we propose a tuning procedure called cv.vote based on the results of V-fold cross validation. It is based on the belief that edges that are consistently selected by many cross validation folds are more likely to be in the true model than the edges that are selected by only a few cross validation folds. Compared to CV, this procedure has been shown to be more effective in controlling the number of false positives, while only slightly increasing the number of false negatives. Since our goal here is to reconstruct the regulatory network rather than to build a good prediction model, the cv.vote procedure is more appropriate than CV. We are currently working on extending this idea and studying its properties. To address the issue of computational feasibility, in [10], [6], [9], we derive approximations to the leave-one-curve-out CV scores. The idea is to approximate the leave-one-out estimates by the estimates based on the full data via Taylor expansions and then plug in these approximations into the formula for the CV score. The resulting approximated CV scores require almost no additional computation and have been shown to be effective in terms of selecting the correct dimension of the models.

Finally, reproducibility of research is being increasingly emphasized in the scientific community. For high-dimensional inference, this is particularly important because of the complexity of the models and the corresponding procedures. For all my recent methodological works, I have been developing publicly available softwares. In particular, R packages space, remMap, fpca, dynamics are developed for the methods proposed in [11], [12], [10] and [9], respectively. All these packages have been well-documented with examples and are regularly maintained.

4 Ongoing works and research plan

I have already mentioned some ongoing works that are direct continuations of my existing research. In this section, I will briefly describe my immediate research plan. I will also mention some ongoing works which are motivated in part by the tools and formulations in my earlier works, but branch out into new fields with different sets of questions.

As mentioned earlier, I am currently working on developing methods for estimating GRNs based on three levels of biological molecules. The key idea is to design models and regularization schemes such that biological principles can be appropriately incorporated, for example, to encourage the prevalence of certain types of regulatory relationships over others. We also plan to extend these models to relate the estimated GRNs to phenotypes such as cancer disease status and to incorporate network-environment interactions. As more components are included into the model, model tuning becomes increasingly challenging since more tuning parameters are now needed. We are working on developing tuning procedures for regularized regression problems that aim at selecting biologically sensible models and at the same time are computationally
feasible.

I am also collaborating with biologists in elucidating gene networks that regulate the development in tomato under different light conditions. Specific aims of this project include: 1) use high-throughput Solexa sequencing methods to identify DNA polymorphisms and transcribed mRNAs in cultivated and wild tomato. 2) Create super-high density maps in a set of interspecific near isogenic line (NIL) population to facilitate QTL cloning. 3) Define transcriptional pathways for leaf development and photomorphogenesis by expression profiling in the NIL sets. We are currently designing simulation studies to determine the required coverage of Solexa sequencing in order to achieve satisfactory power for simultaneously detecting DNA polymorphisms and differentially expressed genes. We plan to use an eQTL mapping approach to identify genetic components of variation between two closely related tomato species. Previously developed methods in [5], [11],[12] will also be extended and applied to build GRNs relating to leaf development and to identify network-environment interactions.

In the area of functional data analysis, I am currently working on extending non-parametric estimation procedures such as kernel smoothing, as well as existing computational tools, to models involving either data or parameters lying on a special manifold. The goal is to establish a general framework for parameter estimation, model selection and statistical inference. As illustrated in [10] and [8], the explicit use of geometry (such as tangent space and exponential map) has many advantages. Particularly, this results in more accurate estimators and helps in deriving efficient model selection procedures. Also, it helps in understanding the problem, as many asymptotical properties of these estimators, such as efficient score representation and asymptotic normality, need to be defined through the tangent space of the manifold.

This research is partly motivated by the analysis of diffusion tensor imaging (DTI) data. DTI is a high-resolution imaging technology used to measure the diffusion-driven displacement of water molecules that probe tissue structures in brain. The directionality of the displacement in each voxel is captured through a 3 by 3 positive definite matrix termed “diffusion tensor”. One important goal of DTI studies is the mapping of white matter fiber tracts. The challenge is to use noise-corrupted diffusion weighted images (DWI) observed at hundreds of thousands of voxels to extract information about the fiber tracts. Since the diffusion images/tensors are spatially dependent, this can be viewed as a nonparametric function estimation problem with geometric constraints on the tensors. We are working on generalizing non-parametric smoothing techniques to the manifold of 3 by 3 positive definite matrices. The novelty of our approach is in performing smoothing in the tangent space, as well as addressing the problems of estimating the function and its gradient field in a unified fashion. The results obtained so far indicate a significant advantage of the smoothing methods that are subject to geometric constraints. We are currently pursuing an extension of this approach in which diffusion tensor information is directly extracted from raw, voxel-based DWI data at a high angular resolution in the presence of possibly multiple fiber orientations.

I am also working on the interface between statistical genomics and functional data analysis. In particular, we are aiming at utilizing time-course expression data to elucidate genetic network structures. Since genetic regulatory relationships are dynamic in nature, time-course data provide complimentary information about GRNs which can not be captured by measurements taken at a single time point. By assuming a simple linear dynamics involving vector autoregressive (VAR) processes, the idea proposed in [12] can be readily extended to identify network hubs and in turn identify important network modules. We then plan to apply functional data analysis tools, including extending the method proposed in [9] to study the (common) underlying
dynamics for genes falling into the same network modules.

Grants Support

My research is currently supported by three grants.

- The research on statistical analysis for models involving Riemannian manifolds is partly supported by the grant DMS-0806128 from National Science Foundation, for which I am the principal investigator.

- The research on utilizing multiple types of high throughput data to elucidate GRNs is partly supported by the grant 1R01GM082802-01A1 from the National Institute of General Medical Sciences, for which I serve as the principal investigator for the subcontract at UC Davis.

- The research on elucidating tomato developmental networks is partly supported by the grant DBI-0820854 from National Science Foundation, for which I am a co-PI.

References


