EDITORIAL

Guest Editors' Introduction to the Special Issue on "Statistical Genomics and Transcriptomics in Agriculture"

Dan NETTLETON and Robert J. TEMPELMAN

Agriculture has been a key motivator of statistical methods development since the very beginning of the modern era of statistical practice. R. A. Fisher's foundational work, for example, was clearly shaped by the challenges in agriculture faced during his days at Rothamsted Experimental Station. Modern technological developments have led to many new types of agricultural data, and with these voluminous data come interesting challenges for those who develop and apply statistical methods. This invited special issue of the *Journal of Agricultural, Biological, and Environmental Statistics* examines two new types of agricultural data and associated challenges for data analysis that have arisen over the past 15 years with enhanced methods for genotyping and for sequencing DNA molecules. These areas are described as genomics and transcriptomics in the title of this issue for the sake of brevity, but the work presented here is more focused on two main problem areas.

The first five articles in the special issue address genomics problems where researchers use thousands of DNA variations to predict the genetic merit of plants and animals for agriculturally important traits. In statistical terms, the primary challenge is prediction with many more correlated predictor variables than independent observations. The last four articles of the special issue consider transcriptomics problems where researchers must make inferences about thousands of response variables with data from few experimental units. Both types of problems have led researchers to develop hierarchical models and Bayesian or empirical Bayesian methods for analysis that differ substantially from the approaches Fisher pioneered nearly a century ago.

The special issue opens with a review paper by Tempelman that introduces past work on whole genome prediction (WGP) and genome wide association (GWA) analyses. WGP and

Dan Nettleton is Distinguished Professor (✉) Department of Statistics, Iowa State University, Ames, IA, USA (E-mail: dnett@iastate.edu). Robert J. Tempelman is Professor Department of Animal Science, Michigan State University, East Lansing, MI, USA (E-mail: tempelma@msu.edu).

© 2015 International Biometric Society

Journal of Agricultural, Biological, and Environmental Statistics, Volume 20, Number 4, Pages 439–441

DOI: 10.1007/s13253-015-0235-0

GWA analyses are research areas of major importance in agricultural quantitative genetics. These techniques are expected to drive enhanced agricultural production needed to feed the world's growing population in the coming years. Both WGP and GWA analyses involve solving problems in high-dimensional regression where the number of observations is far less than the number of explanatory variables. While prediction is the goal in WGP, inference about associations between explanatory variables and the response is the focus of GWA analysis. Tempelman explains the hierarchical models and Bayesian inference strategies that have been popular among plant and animal breeders and points out several areas where additional research is needed.

de los Campos et al. consider the challenge of WGP in the face of population stratification, where observations come from populations that are a union of unknown subpopulations. They show how to discover the existence of subpopulations and cluster observation into subpopulations based on genomic information. Given the assignment of observations to clusters, they allow for the effects of explanatory variables to differ across subpopulations. They illustrate their methods with analyses of pig and wheat datasets and find that their approach tends to provide good prediction accuracy across a variety of realistic circumstances.

Chen and Tempelman propose empirical Bayes alternatives to the computationally intensive Markov chain Monte Carlo (MCMC) approaches commonly used by the plant and animal breeding community to carry out WGP. They investigate the impact of starting values for empirical Bayesian methods based on use of the expectation-maximization (EM) algorithm. They find that performance can be quite sensitive to starting values, and that starting values commonly used in practice may lead to poor performance. They discuss a recently proposed regularization technique that can reduce the impact of starting values for one popular and useful WGP method.

Pérez-Elizalde et al. discuss kernel regression methods for WGP. They focus, in particular, on bandwidth estimation when a Gaussian kernel function is used to define correlations among observations based on similarity of genomic information. In contrast to computationally intractable cross-validation approaches, they develop a Bayesian strategy for estimating the bandwidth parameter via posterior maximization. They demonstrate their approach through the analyses of wheat and maize data and compare the performance of their approach with other strategies that have been popular in practice.

Montesinos-López et al. study WGP for count responses. They develop a fully Bayesian approach for generating predictions with negative binomial response models. They compare their approach to analogous methods that assume counts follow Poisson or Gaussian distributions, and also study the commonly used tactic of applying WGP methods for Gaussian data after log-transforming counts. No one method emerges as a clear favorite when prediction accuracy is tested on two wheat datasets via cross-validation techniques and multiple evaluation criteria.

The final four articles in the special issue present methods for detecting differentially expressed (DE) genes using RNA sequencing (RNA-seq) data. The RNA-seq datasets considered in the special issue involve a matrix of counts with thousands of rows (corresponding to genes) and between 6 and 32 columns (corresponding to experimental or observational units). Each count in the matrix serves as a quantitative measure of a gene's transcript abundance within a particular experimental or observational unit. Genes whose transcript

abundance levels are significantly associated with an explanatory variable of interest are considered to be DE. All four papers use negative binomial response models for RNA-seq count data and develop methods for sharing information across genes to improve detection of DE genes.

Liu, Wang, and Liu study the problem of identifying genes DE across two treatments. They develop a nonparametric Bayes approach in which the distribution of treatment effects is modeled as a draw from a Dirichlet process. They illustrate their approach by discovering genes DE across two maize cell types. A simulation study highlights the advantage of their approach over popular alternative methods.

Nguyen et al. consider the problem of identifying DE genes while accounting for the effects of multiple covariates recorded for each experimental or observational unit. Their approach involves testing for the significance of a factor of primary scientific interest in gene-specific generalized linear models that include a selected set of covariates. The selected set of covariates is obtained via a backward selection procedure that eliminates covariates that appear irrelevant when considering results from the analysis of all genes. Simulations show that the proposed approach can identify relevant covariates and identify DE genes while controlling false discovery rate at desired levels. The work is illustrated by searching for genes DE between two lines of pigs that have undergone divergent selection for an important feed efficiency trait.

Lithio and Nettleton present work on modeling and analysis for RNA-seq experiments with complex designs involving inbred and hybrid plant genotypes. Their approach is based on gene-specific generalized linear mixed-effects models that account for split-plot structures in the experimental design. They assume key gene-specific parameters are draws from distributions whose hyperparameters are estimated via an empirical Bayes approach. Given hyperparameter estimates, integrated nested Laplace approximation is used to make posterior inferences about tests of interest that involve genotypes, environmental stresses, and their interactions.

Niemi et al. also analyze RNA-seq data from inbred and hybrid plant genotypes. They focus on identifying genes that show expression levels in hybrids that are more extreme than their inbred parents. Such genes may play a role in the molecular genetic mechanisms underlying hybrid vigor, the phenomenon in which hybrid genotypes outperform their inbred parents with respect to important phenotypic traits. Following estimation of hyperparameters in their hierarchical model, Niemi et al. implement parallel MCMC techniques to obtain posterior inferences for each gene. The composite null hypotheses of interest in their problem present difficulties for most existing methods of RNA-seq analysis. A simulation study shows that the proposed approach is better able to distinguish between null genes and genes of interest than several competing methods.

We believe the contributions in this special issue highlight a variety of interesting challenges in the analysis of agricultural genomic and transcriptomic data. Even though we have been able to showcase only a small part of a much larger body of work on genomics and transcriptomics, we hope the ideas presented here will stimulate future research in these important and growing areas.