Journal of agricultural, biological, and environmental statistics, ISSN 1085-7117
Volume 20, number 4 (december 2015)

Abstract
Whole genome prediction (WGP) modeling and genome-wide association (GWA) analyses are big data issues in agricultural quantitative genetics. Both areas require meaningful input from the statistical scholarly community in order to further improve the accuracy of prediction of genetic merit and inference on putative causal variants as well as improving the computational efficiency of existing methods and algorithms. These concerns have become increasingly critical as new sequencing technologies will only exacerbate current model dimensionality problems. We focus primarily on mixed model and hierarchical Bayesian analyses which have been most commonly pursued by animal and plant breeders for WGP thus far. We draw attention to our observation that many such previous analyses have not carefully inferred upon hyperparameters defined at the top levels of the Bayesian model hierarchy, but simply arbitrarily specify their values. We also reassess previous discussions on WGP model dimensionality, believing that useful data augmentation schemes utilized in various Markov Chain Monte Carlo (MCMC) schemes have led to a general misunderstanding that heavy-tailed or variable selection-based WGP models may be highly parameterized relative to more standard mixed model representations. Computational efficiency is addressed with respect to MCMC and competitive, albeit approximate, alternatives. Furthermore, GWA analyses are reassessed, encouraging a greater reliance on shrinkage-based inferences based on critically chosen priors, instead of potentially nonreproducible fixed effects *P* value-based inference.

Abstract
Naturally and artificially selected populations usually exhibit some degree of stratification. In Genome-Wide Association Studies and in Whole-Genome Regressions (WGR) analyses, population stratification has been either ignored or dealt with as a potential confounder. However, systematic differences in allele frequency and in patterns of linkage disequilibrium can induce sub-population-specific effects. From this perspective, structure acts as an effect modifier rather than as a confounder. In this article, we extend WGR models commonly used in plant and animal breeding to allow for sub-population-specific effects. This is achieved by decomposing marker effects into main effects and interaction components that describe group-specific deviations. The model can be used both with variable selection and shrinkage methods and can be implemented using existing software for genomic selection. Using a wheat and a pig breeding data set, we compare parameter estimates and the prediction accuracy of the interaction WGR model with WGR analysis

ignoring population stratification (across-group analysis) and with a stratified (i.e., within-sub-population) WGR analysis. The interaction model renders trait-specific estimates of the average correlation of effects between sub-populations; we find that such correlation not only depends on the extent of genetic differentiation in allele frequencies between groups but also varies among traits. The evaluation of prediction accuracy shows a modest superiority of the interaction model relative to the other two approaches. This superiority is the result of better stability in performance of the interaction models across data sets and traits; indeed, in almost all cases, the interaction model was either the best performing model or it performed close to the best performing model.

An Integrated Approach to Empirical Bayesian Whole Genome Prediction Modeling P. 491-511

C. Chen, R. J. Tempelman

Abstract

Computational efficiency is an increasing concern for whole genome prediction (WGP) based on denser genetic marker panels such that algorithms other than Markov Chain Monte Carlo (MCMC) warrant greater consideration, particularly for hierarchical models that flexibly confer either heavy-tailed (e.g., BayesA) or stochastic search and variable selection (SSVS) instead of Gaussian specifications on marker effect distributions. The expectation maximization (EM) algorithm is one attractive alternative; however, recently proposed hierarchical model implementations of EM have not addressed formal estimation of underlying hyperparameters even though their specifications are known to impact WGP accuracy. Furthermore, EM can be sensitive to starting values. We develop and explore the properties of an empirical Bayes strategy by conditioning EM implementations of BayesA or SSVS WGP models on marginal modal estimation of variance components and other key hyperparameters. These empirical Bayes implementations are compared against their MCMC counterparts for estimation of hyperparameters and WGP accuracy, both within the context of a simulation study and application to a loblolly pine dataset. In all cases, starting values were deemed to be important for EM-based estimates. Starting values based on MCMC posterior means were preferable, whereas those based on setting all marker effects equal to zero generally led to inferior performance. Nevertheless, a recently proposed regularization procedure was useful in alleviating the impact of starting values in the EM implementation of the SSVS model, as was modifying the expectation step in the BayesA model to be based on relative variances rather than on relative precisions.

Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction P. 512-532

Sergio Pérez-Elizalde, Jaime Cuevas...

Abstract

One of the most widely used kernel functions in genomic-enabled prediction is the Gaussian kernel. Selection of the bandwidth parameter for kernel regression has generally been based on cross-validation. We propose a Bayesian method for estimating the bandwidth parameter h of a Gaussian kernel as the modal component of the joint posterior distribution of h and the form parameter φ We present a theory for the Bayesian selection of h in a Transformed Gaussian Kernel (TGK) model and its application in two plant breeding datasets (maize and wheat) that were already predicted using the kernel averaging (KA) model in the context of Reproducing Kernel Hilbert Spaces (RKHS KA). We also compared the prediction accuracy of the proposed method with a model that also uses a Gaussian kernel and estimates the bandwidth parameter using a restricted maximum likelihood method (GK REML). Results for the wheat dataset show that the predictive ability of TGK was at least as good as the predictive ability of model RKHS KA, with TGK showing a significantly smaller Predictive Mean Squared Error (PMSE) than the other two approaches. The TGK model was statistically a better predictor than methods GK REML and RKHS KA in terms of

mean PMSE and mean correlations in seven (out of 17) trait-environment combinations in the wheat dataset. Fewer differences were found between models for the maize data; the TGK model generally had similar or inferior prediction accuracy than GK REML and RKHS KA in various analyses. The superiority of GK REML over TGK based on mean PMSE was clear in seven maize traits.

Genomic Prediction Models for Count Data P. 533-554

Osval A. Montesinos-López...

Abstract

Whole genome prediction models are useful tools for breeders when selecting candidate individuals early in life for rapid genetic gains. However, most prediction models developed so far assume that the response variable is continuous and that its empirical distribution can be approximated by a Gaussian model. A few models have been developed for ordered categorical phenotypes, but there is a lack of genomic prediction models for count data. There are well-established regression models for count data that cannot be used for genomic-enabled prediction because they were developed for a large sample size (n) and a small number of parameters (p); however, the rule in genomic-enabled prediction is that p is much larger than the sample size n. Here we propose a Bayesian mixed negative binomial (BMNB) regression model for counts, and we present the conditional distributions necessary to efficiently implement a Gibbs sampler. The proposed Bayesian inference can be implemented routinely. We evaluated the proposed BMNB model together with a Poisson model, a Normal model with untransformed response, and a Normal model with transformed response using a logarithm, and applied them to two real wheat datasets from the International Maize and Wheat Improvement Center. Based on the criteria used for assessing genomic prediction accuracy, results indicated that the BMNB model is a viable alternative for analyzing count data.

A Semi-parametric Bayesian Approach for Differential Expression Analysis of RNA-seq Data P. 555-576

Fangfang Liu, Chong Wang, Peng Liu

Abstract

RNA-sequencing (RNA-seq) technologies have revolutionized the way that agricultural biologists study gene expression as well as generated a tremendous amount of data waiting for analysis. Detecting differentially expressed genes is one of the fundamental steps in RNA-seq data analysis. In this paper, we model the count data from RNA-seq experiments with a Poisson–Gamma hierarchical model, or equivalently, a negative binomial model. We derive a semi-parametric Bayesian approach with a Dirichlet process as the prior model for the distribution of fold changes between the two treatment means. An inference strategy using Gibbs algorithm is developed for differential expression analysis. The results of several simulation studies show that our proposed method outperforms other methods including the popularly applied edgeR and DESeq methods. We also discuss an application of our method to a dataset that compares gene expression between bundle sheath and mesophyll cells in maize leaves. Supplementary materials accompanying this paper appear online.

Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates P. 577-597

Yet Nguyen, Dan Nettleton, Haibo Liu...

Abstract

A common challenge in analysis of transcriptomic data is to identify differentially expressed genes, i.e., genes whose mean transcript abundance levels differ across the levels of a factor of scientific interest. Transcript abundance

levels can be measured simultaneously for thousands of genes in multiple biological samples using RNA sequencing (RNA-seq) technology. Part of the variation in RNA-seq measures of transcript abundance may be associated with variation in continuous and/or categorical covariates measured for each experimental unit or RNA sample. Ignoring relevant covariates or modeling the effects of irrelevant covariates can be detrimental to identifying differentially expressed genes. We propose a backward selection strategy for selecting a set of covariates whose effects are accounted for when searching for differentially expressed genes. We illustrate our approach through the analysis of an RNA-seq study intended to identify genes differentially expressed between two lines of pigs divergently selected for residual feed intake. We use simulation to show the advantages of our backward selection procedure over alternative strategies that either ignore or adjust for all measured covariates.

Hierarchical Modeling and Differential Expression Analysis for RNA-seq Experiments with Inbred and Hybrid Genotypes P. 598-613

Andrew Lithio, Dan Nettleton

Abstract

The performance of inbred and hybrid genotypes is of interest in plant breeding and genetics. High-throughput sequencing of RNA (RNA-seq) has proven to be a useful tool in the study of the molecular genetic responses of inbreds and hybrids to environmental stresses. Commonly used experimental designs and sequencing methods lead to complex data structures that require careful attention in data analysis. We demonstrate an analysis of RNA-seq data from a split-plot design involving drought stress applied to two inbred genotypes and two hybrids formed by crosses between the inbreds. Our generalized linear modeling strategy incorporates random effects for whole-plot experimental units and uses negative binomial distributions to allow for overdispersion in count responses for split-plot experimental units. Variations in gene length and base content, as well as differences in sequencing intensity across experimental units, are also accounted for. Hierarchical modeling with thoughtful parameterization and prior specification allows for borrowing of information across genes to improve estimation of dispersion parameters, genotype effects, treatment effects, and interaction effects of primary interest.

Empirical Bayes Analysis of RNA-seq Data for Detection of Gene Expression Heterosis P. 614-628

Jarad Niemi, Eric Mittman, Will Landau...

Abstract

An important type of heterosis, known as hybrid vigor, refers to the enhancements in the phenotype of hybrid progeny relative to their inbred parents. Although hybrid vigor is extensively utilized in agriculture, its molecular basis is still largely unknown. In an effort to understand phenotypic heterosis at the molecular level, researchers are measuring transcript abundance levels of thousands of genes in parental inbred lines and their hybrid offspring using RNA sequencing (RNA-seq) technology. The resulting data allow researchers to search for evidence of gene expression heterosis as one potential molecular mechanism underlying heterosis of agriculturally important traits. The null hypotheses of greatest interest in testing for gene expression heterosis are composite null hypotheses that are difficult to test with standard statistical approaches for RNA-seq analysis. To address these shortcomings, we develop a hierarchical negative binomial model and draw inferences using a computationally tractable empirical Bayes approach to inference. We demonstrate improvements over alternative methods via a simulation study based on a maize experiment and then analyze that maize experiment with our newly proposed methodology.
Supplementary materials accompanying this paper appear on-line.