

STATISTICS SEMINAR

Zhining Ou

Preliminary Ph.D. Exam

Wednesday, November 11, 2015

Call Hall 204, 1:00-2:00 pm



Efficient Heteroskedastic Extensions for Genomic Prediction and Genome-wide Association Studies

Whole-genome prediction (WGP) models that use single-nucleotide polymorphism marker information to predict genetic merit of animals and plants typically assume homogeneous residual variance. However, variability is often heterogeneous across agricultural production systems and may subsequently bias WGP based inferences. This study extends classical WGP models based on normality, heavy-tailed specifications and variable selection to explicitly account for environmentally-driven residual heteroskedasticity under a hierarchical Bayesian mixed-models framework. WGP models assuming homogeneous or heterogeneous residual variances were fitted to training data generated under simulation scenarios reflecting a gradient of increasing heteroskedasticity. Model fit was based on pseudo-Bayes factors and also on prediction accuracy of genomic breeding values computed on a validation data subset one generation removed from the simulated training dataset. Homogeneous versus heterogeneous residual variance WGP models were also fitted to two quantitative traits, namely 45-minute post-mortem carcass temperature and loin muscle pH, recorded in a swine resource population dataset pre-screened for high and mild residual heteroskedasticity, respectively. Fit of competing WGP models was compared using pseudo-Bayes factors. Predictive ability, defined as the correlation between predicted and observed phenotypes in validation sets of a 5-fold cross-validation was also computed. Heteroskedastic error WGP models showed improved model fit and enhanced prediction accuracy compared to homoscedastic error WGP models although the magnitude of the improvement was small (less than two percentage points net gain in prediction accuracy). Nevertheless, accounting for residual heteroskedasticity did improve accuracy of selection, especially on individuals of extreme genetic merit. In this research proposal, I present ideas to extend the assessment of heteroscedasticity to a GWAS for identification of genomic regions putatively associated with a trait of interest. I also propose to enhance computational efficiency of heteroskedastic WGP models by incorporating an EM-based model approximate to fixed and random effect components of the residual variances.