Newest Developments and Urgent Issues in Measurement Error and Latent Variable Problems

Paul Gustafson (University of British Columbia)
Yanyuan Ma (Penn State University)
Liqun Wang (University of Mannitoba)
Grace Yi (University of Waterloo)
August 14-19, 2016

1 Overview of the Field, Development and Problems

Whenever measurements are taken and recorded, it is almost unavoidable that errors are involved, due to mechanical reason, human reason or mathematical modeling reason. In addition, scientific investigations often involve variables that cannot be measured directly. Vast majority of statistical analysis takes into account the measurement uncertainty in the response variable while ignores such uncertainty in the covariates. However, although this is acceptable practice for most prediction purposes, it can cause considerable bias in estimation and alter the scientific conclusion. Moreover, measurement error in covariates may mask the true relationship of interest and thus undermine the chance of scientific discoveries.

Measurement error models take into account the additional measurement uncertainty in the covariates by explicitly assuming the scientifically interesting variable is unobservable, while only an approximation of it is available. Methods are then needed to establish the relation between the response and the variables that are latent. Different from the usual missing data problems, in a measurement error model, the scientifically interesting variables are never observed. In other words, the missingness rate is 100%. This creates challenges as well as the need for new methods that are unique to the measurement error models.

Effect of measurement error and its correction. Generally speaking, ignoring measure error can cause bias in parameter estimation and lead to false conclusions in hypothesis tests. In the relatively simple linear or generalized linear models, the bias generally causes the coefficients to attenuate towards zero. However, as soon as the covariates enter the model in a nonlinear fashion, the bias can have very different impact. Various approaches have been proposed to correct for such bias, the most famous ones being the regression calibration and the simulation extrapolation methods. Both methods can decrease the estimation bias to different extents, while both methods rely on certain tuning components, respectively the working latent variable distribution and the extrapolation function. Thus, in relatively complex model settings, the selection of the tuning components becomes critical to the success of these methods and is worth careful further investigation.

Generalized linear models and corrected estimating equations. A popular class of statistical models in both biological studies and econometrics is the generalized linear

models. When measurement error is additive and normal, consistent estimating equations can be constructed through modifying the score functions from the error-free context. The effectiveness of the corrected estimating equation methods has prompted the need to extend the idea to more general settings, including non-normal error, models deviates from exponential family, and even the quantile regression models. While in certain special context, such extension has been successful, in most other contexts, more careful investigation is needed.

Semiparametric estimation approach. In a more general parametric regression context, viewing the latent variable distribution as a nonparametric component of the likelihood of the observations, the measurement error problem can be cast in a semiparametric framework. The semiparametric approach is a very powerful way of handling the measurement error models and has been successful in constructing estimators that were not obtainable previously. It is important to popularize this method in the measurement error area. In addition, the generalization of this approach to semiparametric regression models poses great challenge and calls for urgent further study.

Nonparametric measurement error problems. When the latent variable itself enters the regression model inside an unspecified smooth function, the measurement error problem becomes a very different one. Because the regression relation concerning the variable of interest is essentially completely left arbitrary, the treatment to the problem is very different from the parametric or semiparametric case. Currently, deconvolution and sieve methods are the two major approaches widely accepted, while both have their own limitations. Generally speaking, the deconvolution approach yields very slow convergence rate, and the properties of the sieve method is only studied in several limited cases. Thus, more in-depth study on the nonparametric measurement error problems is important and useful.

Related latent variable problems. Although measurement error stands alone as a unique problem because of its ubiquitous presence in practice, its statistical abstraction shares common structures with many other latent variable models such as mixed effect models in biostatistics, panel data models in social sciences, generalized linear latent variable models in psychology. It is also linked with missing data and censored observation problems. Thus, the research in measurement error problems is also directly linked to works in these different areas and exchange of ideas will be beneficial both ways.

Measurement error models in the era of complex and big data

There have been rapid developments of large scale and complex structured data in recent years. While a rich source of methods and models is becoming available in handling such data ignoring measurement errors, it is time that the measurement error aspects of these problems come under spotlight and the related methods brought up to date. Because of the latent nature of the covariates in measurement error models, the methodologies are much more complex than in the usual regression models, hence a unique challenge in the complex

and big data context is computation. Very scarse literature is available so far and great opportunities are awaiting.

2 Presentation Highlights

2.1 Monday, August 15

The workshop began on Monday, August 15. In the morning session, Professor Yanyuan Ma from the Penn State University introduced the first speaker Professor Raymond Carroll from the Texas A&M University, and Professor Hua Liang from the George Washington University introduced the second speaker Professor Donna Spiegelman from the Harvard University. In the first afternoon session, Professor Weixing Song from Kansas State University introduced Professor Xianzheng Huang from the University of South Carolina. The second afternoon session consisted of problem discussions. It is chaired by Professor Yair Goldberg from the University of Haifa in Israel, who introduced the two speakers, Professor Malka Gorfine from the University of Tel Aviv in Israel and professor Alicia Carriquiry from the Iowa State University. The discussion throughout the day was heated and exciting.

The workshop began with an excellent overview presentation by Professor Raymond Carroll, who presented a talk titled "New Measurement Error Data Structures". The presentation is a survey overview on some relatively recent measurement error data structures, and brief discussions of how might one analyze them. Problems include (a) mixtures of continuous and discrete variables both subject to error; (b) problems in which the target covariate varies over time and is measured with error; (c) classical additive and Berkson multiplicative errors in radiation exposure doses; (d) Calibration and seasonal adjustment for matched case-control studies; (e) Spatial regression with covariate measurement error; (f) using functional data analysis to assess measurement error in energy intake; (g) deconvolution with heteroscedastic measurement error.

Professor Donna Spiegelman presented a specific problem encountered in epidemiology. She talked about "Generalized methods-of-moments estimation and inference for the assessment of multiple imperfect measures of diet and physical activity in validation studies". In her talk, she explained why accurate and precise measurement of diet and physical activity (PA) in free-living populations is difficult, and hence as a result, key findings in nutritional epidemiology have been controversial, a notable example of this being the relation of dietary fat intake to breast cancer risk. There is a long literature in statistics on methods to adjust relative risk estimates for cancer and other chronic diseases for bias due to measurement error in long-term dietary intake and PA. To use the popular regression calibration method to correct for the bias, the de-attenuation factor needs to be estimated. Popular is used here in the sense that it is virtually the only method for correcting for bias due to

exposure measurement error that has been used in applications, and there are hundreds of published instances of this. In this talk, she developed semi-parametric generalized methods of moments estimators for the de-attenuation factor and other quantities of interest, in particular, the correlation of each surrogate measure with the unobserved truth and intra-class correlation coefficients characterizing the random within-person variation around each measurement. The method makes assumptions only about the first two moments of the multivariate distribution between the measures. The robust variance is derived to allow asymptotic inference. She considered a one-step method which is theoretically inefficient, as well as fully efficient methods that are iterative. For some variables of interest, such as total energy intake, protein density, and total PA, there may be unbiased gold standards (X) available and when they are available, they are used. When these are not available and even when so, she considered other objective (W) and subjective measures (Z), such as biased (concentration) biomarkers, self-report, accelerometer and pulse, as means of estimating the de-attenuation factor and other quantities of interest. Measurements denoted W are assumed to have errors uncorrelated with all other measurements, and those denoted Z are allowed to have correlated errors with one or more of the other measures. Harvards Womens Lifestyle Validation Study (WLVS) assessed diet and physical activity over a 1 year period among 777 women. Total physical activity was assessed by doubly labeled water, often considered to be the gold standard for energy expenditure, accelerometer, resting pulse, physical activity questionnaire (PAQ), and ACT24, an on-line PA assessment tool. Thus, dim(X)=1, dim(Z)=2, and dim(W)=2. Using all 5 of these measures, the deattenuation factor (kcal/ MET-hours) for total physical activity assessed by the PAQ was estimated to be 4.09 (95% CI 0.94, 7.24) and for ACT24 5.59 (1.43, 9.75). These de-attenuation factors are calibrating the units of the PAQ and ACT24 from MET-hours/day to kcal/day, as well as adjusting for bias due to measurement error. In addition, using all 5 measures, the respective correlations of PAQ and ACT24 with truth were 0.36 (0.30, 0.41) and 0.32 (0.26, 0.38), respectively, and correlations of the accelerometer and resting pulse with truth were 0.891 (0.887, 0.893) and -0.20 (-0.32, -0.71) respectively. Little gain in efficiency between the one-step and fully iterated estimators was evident in this example. User-friendly publicly available software is under development.

In the first talk of Monday afternoon, Professor Xianzheng Huang presented "Nonparametric Modal Regression in the Presence of Measurement Error". The main theme of the talk is in the context of regressing a response Y on a predictor X. She considered estimating the local modes of the distribution of Y given $X = x$ when the data for X are contaminated with measurement error. She proposed two nonparametric estimation methods. In one approach she related this problem to estimating the partial derivative of the joint density of $(X; Y)$ in the presence of measurement error; and the second approach is built upon estimating the partial derivative of the conditional density of Y given $X = x$ using error-prone

data. She studied the asymptotic properties of the mode estimator resulting from each method, and demonstrate their performance via simulation experiments.

In the discussion session on Monday, Professor Malka Gorfine talked about her partially finished work on "Nonparametric adjustment for risk-prediction model with mis-measured family history". In her problem, mis-measured time to event data used as a predictor in risk prediction models will lead to inaccurate predictions. This arises in the context of self-reported family history, a time to event predictor often measured with error, used in Mendelian risk prediction models. Using validation data, she proposed a method to adjust for this type of error. She estimated the measurement error process using a nonparametric smoothed Kaplan-Meier estimator, and use Monte Carlo integration to implement the adjustment. She applied her method to simulated data in the context of both Mendelian and multivariate survival prediction models. Simulations are evaluated using measures of mean squared error of prediction (MSEP), area under the response operating characteristics curve (ROC-AUC), and the ratio of observed to expected number of events. These results show that her method mitigates the effects of measurement error mainly by improving calibration and total accuracy, and only partially improving discrimination. Hence her question is "Can you improve the proposed methodology for improving ROC-AUC?" While the audience raised many questions, it was also pointed out that the performance she saw was actually already better than what others have observed. Hence a potential question arises—is it possible to quantify the theoretical bounds of the improvement and hence provide a quantitative evaluation on how good the performance is and provide a guide on whether or not more effort should be spent to further achieve improvement?

The second presenter of the problem session, Professor Alicia Carriquiry modified her original presentation title to "Bivariate kernel deconvolution density estimation: An application to vitamin D". She presented a unique type of problems that arises from analyzing vitamin and bone health data which calls for change point estimation in the presence of measurement errors. Her current approach is through multivariate deconvolution, which raises some concern due to the slow convergence rate of deconvolution in general and the challenging computational issues. In fact, for the purpose of estimating change point, which is only one aspect of the distribution instead of the whole distribution, using deconvolution may not be the most efficient approach. Discussions are on possible alternative methods, including semiparametric method and extended discussion on the issue was continued after the session.

2.2 Tuesday, August 16

On Tuesday, Professor Tanya Garcia from the Texas A&M University introduced Professor Eugene Huang from Emory University as the first speaker, and Professor Len Stefanski from

North Carolina University introduced Doctor Victor Kipnis from National Cancer Institute. After a brief tea break, Professor Liqun Wang from University of Manitoba introduced Professor Arthur Lewbel from Boston College and concluded the morning sessions. The first afternoon session consists two talks. The first is given by Professor Joan Hu from Simon Fraser University, she was chaired by Professor Fei Jiang from the University of Hongkong. The second is given by Professor Donglin Zeng from the University of North Carolina, chaired by Professor Yuanjia Wang from Columbia University. The problem session was chaired by Professor Malka Gorfine, who introduced Professor Yair Goldberg and Professor Haiying Wang from the University of New Hampshire.

The first talk on Tuesday was presented by Professor Eugene Huang. He presented "On heteroscedastic covariate measurement error in Cox regression". In his talk, he motivated the problem using many survival studies that have error-contaminated covariates, which may lack a gold standard of measurement. Furthermore, the error distribution can depend on the true covariates but the dependence structure is typically difficult to quantify; heteroscedasticity is a common manifestation. He suggest an additive measurement error model in this circumstance, and develop a functional modeling method for Cox regression when an instrumental variable is available. The estimated regression coefficients are consistent and asymptotically normal. Preliminary numerical studies, including simulations, were provided. The special feature of his talk is on decomposing the measurement error into a symmetric part and a heteroscedastic nonsymmetric part, which is very interesting, more flexible and surprisingly allows better results to be obtained than in simple settings.

In the second morning session, Doctor Victor Kipnis presented "Time-varying models for longitudinal data measured with error, with application to physical activity and sleep". Modern accelerometers provide interesting and objective longitudinal data on different characteristics of physical activity that may influence important health outcomes. Those characteristics may fluctuate over a short span of time due to life demands, and their dynamic nature at the individual level is often of principal interest. The current research is motivated by the problem of estimating the temporal effect of moderate and vigorous physical activity on sleep using accelerometry measurements. He analyzes weekly data from the BodyMedia study of 3650 women and 1009 men who wore accelerometers continuously for 12 consecutive weeks. On an appropriate scale, he proposed a joint multivariate linear mixed model when both the exposure and bivariate outcome (lying down minutes and sleep minutes) vary over time and are subject to measurement error. To accommodate the possibility that heterogeneities in person-specific trajectories in physical activity and sleep characteristics may be related, hes allow random effects in the corresponding parts of the model to be correlated. This correlation leads to important differences among the individual-level (or within-person), between-person, and population-level (or marginal) effects, as is exemplified by his data. His simulations also demonstrate that ignoring correlated random effects, as is

common in the mixed model approach to longitudinal data that are subject to measurement error, leads to substantial biases in estimated exposure effects.

Also in the second morning session, Professor Arthur Lewbel presented "Unobserved Preference Heterogeneity in Demand Using Generalized Random Coefficients". He modeled unobserved preference heterogeneity in demand systems as random Barten scales in utility functions. These Barten scales appear as random coefficients multiplying prices in demand functions. Consumer demands are nonlinear in prices and may have unknown functional structure. He therefore proved identification of additive Generalized Random Coefficients models, defined as additive nonparametric regressions where each regressor is multiplied by an unobserved random coefficient having an unknown distribution. Using Canadian data, he estimated energy demand functions with and without random coefficient Barten scales. He found that not accounting for this unobserved preference heterogeneity substantially biases estimated consumer-surplus costs of an energy tax.

In the first session of the afternoon, Professor Joan Hu presented "Application of Latent Class Models in a Cancer Survivorship Study". In her talk, Cancer survivors are often at risk of subsequent and ongoing health problems that are primarily treatment-related. She presented an analysis of the medical cost data associated with the longitudinal physician claims of a cancer survivor cohort and a sample from the general population under latent class models. It allows her to classify the survivors into two groups, the at-risk and not-at-risk groups, and to make comparisons between the survivor cohort and the general population.

The second presentation of the session is given by Professor Donglin Zeng, on "Threshold-Dependent Proportional Hazards Model to Assess Risk Factors for Incident Diabetes Defined by Plasma Glucose Levels". In his presentation, the Atherosclerosis Risk in Communities (ARIC) Study is a prospective study of risk factors for atherosclerosis being conducted in four U.S. communities. One important objective of this study is to assess the risk factors for diabetes. A participant is classified as diabetic when his or her fasting plasma glucose (FPG) value crosses a specified, fixed threshold. However, the exact time when the threshold is crossed is not observable when FPG values are subject to substantial measurement error. In this talk, he proposed a semiparametric regression model based on the generalized extreme-value distribution to model the longitudinal FPG values. His model is equivalent to modeling threshold-dependent time to diabetes via a Cox proportional hazards model, where the threshold-dependent event time is defined as the time of the FPG values crossing a given threshold. To account for measurement error in the FPG values, he estimated the model parameters using the nonparametric pseudo-likelihood approach and implement computation via the pseudo-EM algorithm. In analyzing the ARIC Study data, several factors were found to be significantly associated with diabetes.

The last session on Tuesday is also a problem session. First, Professor Yair Goldberg

presented "Inference for mixed effect kernel machines". Kernel machines are widely used in analysis of complex and high-dimensional data due to their flexibility, their ability to incorporate nonlinearity, and their fast computational aspects. Applying kernel machine for data with dependencies is not trivial. Liu et. al, (2007) and Pearce and Wand (2009) show connection between kernel machines and mixed effect models and proposed mixed effect kernel-machine estimators. In this talk, he stated that he would like to develop powerful inference techniques for these estimators and study their properties.

The second presentation was given by Professor Haiying Wang on "Subset sampling with measurement errors". He raised an important problem in dealing with massive data. There, one approach is to use a subset of the full data so that available computing facilities can handle it. The key is to choose the most informative data points so that a small subsample preserves the major information contained in the full data. He has been working on projects from this aspect using two different approaches: random subsampling and deterministically selection. But all methods and algorithms available so far assume that the full data are observed precisely. If data (in covariates) are subject to measurement errors, both approaches will be affected and have to be adjusted for desirable performances. He asks whether this is a problem worth to discuss, and of course it is very much so. In fact, not only it is worth, it is also hard and much research activities are to be expected.

2.3 Wednesday, August 17

The Wednesday program is relatively short since it consists only of a half day presentations. The first speaker was Professor Weixing Song introduced by Professor Eugene Huang. After a brief tea break, Professor Ingrid Van Keilegom from Catholic University of Louvain la Neuve was introduced by Professor Aurore Delaigle from the University of Melbourne to give the second talk, and Professor Yingyao Hu from the Johns Hopkins University was introduced by Professor Donna Spiegelman to give the last talk.

In the first presentation. Professor Weixing Song talked about "Regression Calibration in Measurement Error Modeling". When a p-dimensional parameter θ is defined through the moment condition $Em(X, \theta) = 0$, a simple estimation procedure of θ is proposed by Hong and Tamer when X, a k-dimensional random vector, is contaminated with Laplace measurement error U, that is, he can only observe $Z = X + U$. However, the estimation procedure was designed particularly for the cases where the components of the measurement error vector U are independent. He introduced a general multivariate Laplace distribution, then extend the Hong-Tamer moment estimation procedure to a more general multivariate scenario. Moreover, the Hong-Tamer moment estimation procedure is based on the unconditional expectation $Em(X, \theta) = EH(Z, \theta)$ for some function H. Example shows this techniques does not work in some cases. He further discussed an estimation procedure based

on the condition expectation $E(m(X,\theta)|Z)$, which can be treated as an extension of the regression calibration technique. Large sample properties of the proposed estimation procedure will be investigated. Next, he tried to extend the above extended regression technique to nonparametric setup, particularly focusing on the normal and Laplace measurement error.

Professor Ingrid Van Keilegom presented "Frontier estimation in the presence of measurement error with unknown variance". She considered the problem of estimating a stochastic frontier, i.e. a frontier that is subject to (additive) measurement error. Contrary to other papers in the literature who work with unknown frontiers and normal noise variables, she considered the case where the variance of the noise is unknown. She showed that under weak model assumptions this variance is identifiable, and she proposed three ways to estimate this variance. The first proposal is given in Kneip, Simar and Van Keilegom (2015), who study the asymptotic theory and finite sample behavior in detail. The two other proposals are currently under investigation. Preliminary results will be given showing their excellent finite sample behavior. All three methods will first be studied in the univariate case (i.e. in the case where the boundary of the support of a univariate variable is of interest). The extension to (two- or more-dimensional) frontier models will be given in a second step.

The last talk of the day was given by Professor Yingyao Hu on "Microeconomic Models with Latent Variables: Econometric Methods and Empirical Applications". He reviewed recent developments in nonparametric identification of measurement error models and their applications in applied microeconomics, in particular, in empirical industrial organization and labor economics. Measurement error models describe mappings from a latent distribution to an observed distribution. The identification and estimation of measurement error models focus on how to obtain the latent distribution and the measurement error distribution from the observed distribution. Such a framework is suitable for many microeconomic models with latent variables, such as models with unobserved heterogeneity or unobserved state variables and panel data models with fixed effects. Recent developments in measurement error models allow very flexible specification of the latent distribution and the measurement error distribution. These developments greatly broaden economic applications of measurement error models. This paper provides an accessible introduction of these technical results to empirical researchers so as to expand applications of measurement error models.

2.4 Thursday, August 18

Thursday's program was very full. The first presentation was given by Professor Samiran Sinha from Texas A&M University, introduced by Professor Xianzheng Huang. It is followed by a brief tea break and two more talks, Professor Aurore Delaigle, who was introduced by Professor Ingrid Van Keleigom and Professor Qihua Wang from Chinese Academy of Sciences, who was introduced by Professor Donglin Zeng. In the afternoon, the first session

was reserved to be student presentation session. The first talk was given by Di Shu from the University of Waterloo, introduced by her PhD supervisor Professor Grace Yi, and the second talk was by Tanja Hoegg from the University of British Columbia, introduced by her supervisor Professor Paul Gustafson. The afternoon also contained a problem discussion session, presented by Professors Malka Gorfine and Ingrid Van Keilegom. The program concluded with a bonus talk by Professor Zhongyi Zhu from Fudan University.

Professor Samiran Sinha presented "Analysis of proportional odds models with censoring and errors-in-covariates". He proposed a consistent method for estimating both the finite and infinite dimensional parameters of the proportional odds model when a covariate is subject to measurement error and time-to-events are subject to right censoring. The proposed method does not rely on the distributional assumption of the true covariate which is not observed in the data. In addition, the proposed estimator does not require the measurement error to be normally distributed or to have any other specific distribution, and he does not attempt to assess the error distribution. Instead, he constructed martingale based estimators through inversion, using only the moment properties of the error distribution, estimable from multiple erroneous measurements of the true covariate. The theoretical properties of the estimators are established and the finite sample performance is demonstrated via simulations. He illustrated the usefulness of the method by analyzing a dataset from a clinical study on AIDS.

Professor Aurore Delaigle talked about "Nonparametric covariate-adjusted regression". She considered nonparametric estimation of a regression curve when the data are observed with multiplicative distortion which depends on an observed confounding variable. She suggested several estimators, ranging from a relatively simple one that relies on restrictive assumptions usually made in the literature, to a sophisticated piecewise approach that involves reconstructing a smooth curve from an estimator of a constant multiple of its absolute value, and which can be applied in much more general scenarios. She showed that, although her nonparametric estimators are constructed from predictors of the unobserved undistorted data, they have the same first order asymptotic properties as the standard estimators that could be computed if the undistorted data were available. She illustrated the good numerical performance of her methods on both simulated and real datasets.

Professor Qihua Wang presented "LPRE criterion based estimating equation approaches for the error-in-covariables multiplicative regression models". In this talk, he proposed two estimating equation based methods to estimate the regression parameter vector in the multiplicative regression model when a subset of covariates are subject to measurement error but replicate measurements of their surrogates are available. Both methods allow the number of replicate measurements to vary between subjects. No parametric assumption is imposed on the measurement error term and the true covariates which are not observed in the data set. Under some regularity conditions, the asymptotic normality is established for

both methods. Some simulation studies are conducted to assess the performances of the proposed methods. Real data analysis is used to illustrate his methods.

In the afternoon student presentation session, first, Di Shu presented "IPTW estimation in marginal structural models with error-prone time-varying confounders". The inverse-probability-of-treatment weighted (IPTW) method is a useful approach for estimation of causal parameters pertaining to marginal structural models. This method requires that the measurements of the associated variables are precisely collected. In practice, however, measurement error arises commonly. In this talk, she discussed how measurement error in time-varying confounders can bias the IPTW estimators for causal effects. To adjust for the measurement error effects, she developed several methods to consistently estimate causal parameters. Numerical studies are conducted to assess the performance of her methods.

Second, Tanja Hoegg presented "Bayesian analysis of matched case-control data subject to outcome misclassification, with application to database studies of multiple sclerosis". Health administrative databases collected by the Canadian provincial governments are often used as a cost-effective data source for multiple sclerosis (MS) research at the population level. Due to a high misdiagnosis rate in MS, identification of study subjects from administrative data results in high numbers of false positives and thus requires statistical techniques allowing for imperfect outcome variables. Motivated by an ongoing Canada-wide matched case-control study examining healthcare utilization between MS cases and healthy controls, she investigated the impact of outcome misclassification on association measures under this sampling scheme. Further, she aimed to develop a Bayesian model for the analysis of associations between a binary exposure and a misclassified outcome variable. Emphasis will be placed on allowing for non-constant misclassification probabilities among the subjects as a way to incorporate information contributing to the certainty of an individuals true outcome, such as the total count of MS-related physician contacts.

In the problem discussion session, Professor Malka Gorfine discussed "Heritability estimation based on unrelated individuals and GWAS data". The popular Genome-wide Complex Trait Analysis (GCTA) software uses the random-effects models for estimating the narrow-sense heritability based on GWAS data of unrelated individuals when the causal loci are latent. Many methods have since extended this approach to various situations. However, since the proportion of causal loci among the variants is typically very small and GCTA uses all variants to calculate the similarities among individuals, the estimation of heritability may be unstable, resulting in a large variance of the estimates. Moreover, if the causal SNPs are not genotyped, GCTA sometimes greatly underestimates the true narrow-sense heritability. She presented a novel narrow-sense heritability estimator, named HERRA, using well-developed ultra-high dimensional machine-learning methods. However, improving HERRA is required as it depends on several complexity parameters.

Professor Ingrid Van Keilegom further presented "Identification and estimation of mea-

surement error models with unknown distribution" in the same session. She is very interested in the identification and estimation of models in the presence of measurement error with unknown distribution or just unknown variance. Some recent papers have shown that this problem is sometimes identifiable, and it appears that this is a very promising area. The talk she gave tried to give a partial answer in the context of frontier models, and she called for insight, ideas, and suggestions to further solve theoretical and numerical issues.

To conclude the day, Professor Zhongyi Zhu presented "Simultaneous Mean and Covariance Estimation of Partially Linear Models for Longitudinal Data with Missing Responses and Covariate Measurement Error". Missing responses and covariate measurement error are very commonly seen in practice. New estimating equations are developed to simultaneously estimate the mean and covariance under a partially linear model for longitudinal data with missing responses and covariate measurement error. Specifically, a novel approach is proposed to handle measurement error by using independent replicate measurements. Compared with existing methods, the proposed method requires fewer assumptions. For example, it does not require to specify the distribution of the mismeasured covariate or the measurement error, and does not need a parametric model to estimate the probability of being observed or to impute the missing responses. Additionally, the proposed estimating equations are easy to implement in most popular statistical softwares by applying existing algorithms for standard generalized estimating equations. The asymptotic properties of the proposed estimators are established under regularity conditions, and simulation studies demonstrate desired properties. Finally, the proposed method is applied to data from the Lifestyle Education for Activity and Nutrition (LEAN) study. This data analysis confirms the effectiveness of the intervention in producing weight loss at month nine.

2.5 Friday, August 19

On the last day, Professor Tanya Garcia was introduced by professor Samiran Sinha and Professor Magne Thoresen from the University of Oslo was introduced by Professor Yanyuan Ma. The last session of the day as well as the workshop was a problem session presented by Professor Wenqing He from the University of West Ontario.

Professor Tanya Garcia presented "Simultaneous treatment of unspecified heteroskedastic model error distribution and mismeasured covariates for restricted moment models". Her presentation is concerned with the consistent and efficient estimation of parameters in general regression models with mismeasured covariates. She assumed the distributions of the model error and covariates are completely unspecified, and that the measurement error distribution is a general parametric distribution with unknown variance-covariance. In this general setting, she constructed root-n consistent, asymptotically normal and locally efficient estimators based on the semiparametric efficient score. Constructing the consistent

estimator does not involve estimating the unknown distributions, nor modeling the potential model error heteroskedasticity. Instead, a consistent estimator is formed under possibly incorrect working models for the model error distribution, the error-prone covariate distribution, or both. A simulation study demonstrates that her method is robust and performs well for different incorrect working models, and various homoscedastic and heteroskedastic regression models with error-prone covariates. The usefulness of the method is further illustrated in a real data example.

Professor Magne Thoresen posed a question on "Ultrahigh dimensional variable screening with measurement error". Statisticians frequently encounter regression problems where the number of potential predictors/covariates exceeds the number of observations. Penalized regression methods may in such situations be a good solution. However, these days statisticians also have to deal with situations where the number of potential predictors is too high even for these methods, hence some initial variable filtering is necessary. There is a rather large literature on this, but it is unclear how one should deal with measurement error in such situations. He briefly presented one example, where the variable selection is further complicated by a number of issues, like non-linearities and missing data.

In the last problem session, professor Wenqing He talked about "Measurement error problems in image co-registration: a prostate cancer investigation". He is involved in a prostate cancer imaging project. The prostate of each patient will take several types of images, such as MRI, PET MRI, CT and U/S (recently Na-MRI is added), and then the prostate is removed from the body. Additional MRI and CT will be taken again for the prostate after it is outside the body. The prostate is sliced and the pathologists will contour every slice for tumor. The goal is to construct a predictive model to predict cancer for each unit of the prostate using multiple imaging data available (for prostate inside the body).

One of the main problems here is what is called co-registration procedure: the in vivo MRI and in vitro MRI need to be aligned with CT and the pathological results. There are several type of misclassifications, error prone measurements. He is working on how to incorporate the measurement error model in both co-registration and the prediction procedure. Note that the in vivo MRI and in vitro MRI data can not be exactly aligned, since the prostate may be shrunk when it is removed from the body, etc.

3 Outcome of the Meeting

Due to unexpected health reasons and visa problems, 40 participants attended the workshop among the 42. Among these 40 participants, there are 11 Canadians, 16 females, and 17 graduate students or junior researchers. The workshop also attracted well-known researchers. The workshop is a great success. Participants had extremely positive experiences with the workshop, and they were generally very satisfied with facilities, meals, and accom-

modation at the Banff Center. The workshop had a great impact on the graduate students and junior researchers with respect to their future career plannings. They find themselves greatly benefited. Senior researchers also find this an excellent place for strengthening collaboration and communication. The organizers have received many very positive comments, "This is the best workshop I have been!", "This is a wonderful workshop!", "I enjoyed it so much that it finished too quickly!". The workshop provides an excellent opportunity for leading and young researchers in the field to discuss recent developments, emerging issues, and future directions in the measurement error problems. One of the major goals of the workshop is to strengthen collaboration and communication among different research groups. We have successfully achieved this goal. There were many interesting and active discussions throughout the 5-day workshop. Senior researchers offered their visions, suggestions and guidance, and junior researchers learned many latest developments and exciting future research opportunities. Overall, the workshop is timely and provides a great platform for collaborative research and interactions between methodological and applied researchers and between biostatisticians and economists. We find that BIRS is an ideal place for such communications, and we could not achieve the same results in a "usual" scientific meeting. Finally, workshop participants have expressed great appreciation to BIRS and Banff Center staff members for the outstanding local arrangements and service. In particular, the workshop organizers would like to express their sincere appreciation to the BIRS Station Manager Linda Jarigina-Sahoo, the scientific programme coordinator Chee Chow, the Station Facilitator Bojan Cosic and Technology Manager Brent Kearney for their extremely professional help. We know that a large amount of work is involved in the organization and local arrangements. The wonderful BIRS staff team has made the workshop a huge success!