Bayesian and robust insights in data analysis and classification of genomics and health data

Alexandra Posekany

TU Vienna University of Technology, Austria

06-07-2023

NOVAMATH
CENTER FOR MATHEMATICS
+ APPLICATIONS

REPÚBLICA
PORTUGUESA

1 Motivation

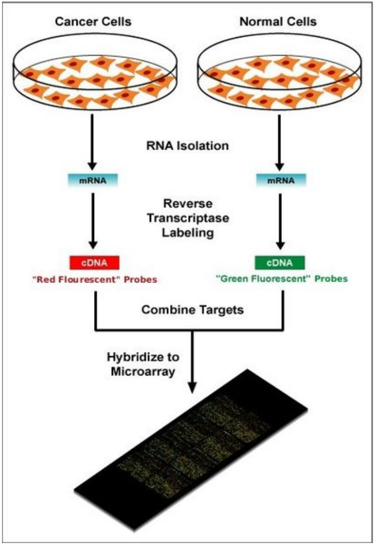
2 Bayesian Background

3 Robust Bayesian ANOVA Models

4 MCMC

5 Biological findings

Biological Motivation - Microarrays



Biological Motivation

Overdispersed data and **outliers** are common for genetic data as found by microarrays, . . .

- optimized and standardized protocols for analysis
- availability of multiple computational tools
- RNA-Seq files are considerably larger than microarray files

- How well-fitting are the Gaussian models (used e.g. by BioConductor package limma)?
- Does a more widely dispersed distribution as likelihood better fit the data?
- Which impact does the model choice have on the biological interpretation of results?
- Can identifying over-dispersed behaviour of certain genes/arrays/experiments provide us with a tool for quality control?

Statistical Motivation

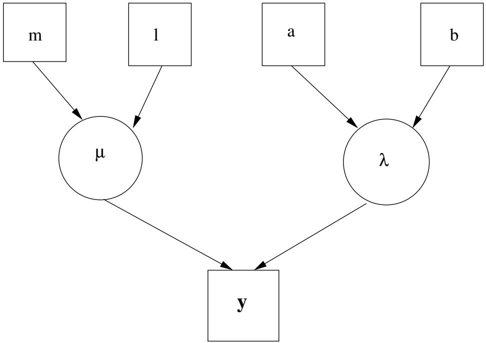
genetic data collected by microarrays, etc. pose several challenges for data analysis (cf. Huber et al.):

- 'noise' behaviour: overdispersion, non-Gaussianity;
  no clear knowledge of systematics of this behaviour
- high-dimensional data $n \ll p$
  very small sample size, many parameters to estimate
  algorithmic methods can handle this
  Bayesian approach 'pools' information in hierarchical priors
- BIG data
  computational analysis becomes an additional challenge
  cf. methods for making parallel computation feasible for MCMC which is inherently serial in nature

Bayesian hierarchical models & DAGs

Bayesian paradigm: consider parameters as random variables
⇒ add prior on parameter, additional latent parameters

Directed acyclic graph (DAG): visualisation of hierarchical model



S. Petrone: "Hierarchical models are a breach with classical frequentist statistical approaches. They can only be well-defined within the Bayesian paradigm."

Bayes factor and Savage-Dickey density ratio

Bayes Factor for testing hypothesis 1: $\theta \in \Theta_1$ against hypothesis 2: $\theta \in \Theta_2$

$$BF = \frac{\int_{\Theta_1} p(\theta|\text{Data})d\theta}{\int_{\Theta_2} p(\theta|\text{Data})d\theta}$$

problem: evaluating the marginal likelihoods corresponding to the hypotheses
for point hypotheses → Savage Dickey density ratio

$$SDR = \frac{p(\theta = \theta_0|\text{Data})}{p(\theta_0)}$$

'only' estimate marginal posterior density $p(\theta|\text{Data})$ at point θ_0

Bayesian Global robustness

for a set Γ of considered prior or likelihood functions, calculate the range of results $r(\Gamma)$

$$\begin{aligned} r(\Gamma) &= \|\overline{\psi}-\underline{\psi}\|, \\ &\overline{\psi}=\sup_{\pi\in\Gamma}\psi(\pi,f),\quad \underline{\psi}=\inf_{\pi\in\Gamma}\psi(\pi,f), \end{aligned} \tag{1}$$

where π represents the prior, f the likelihood function and $\psi(\pi,f)$ a decision of some kind, e. g. a point estimator from the posterior or some quantity of interest.

Bayesian Likelihood robustness & model selection

notion of **likelihood robustness**: *global robustness* compares the range of estimates, varying within a given set of possible likelihoods
Shyamalkumar defines a *finite* set of possible likelihood functions, selects the "optimal", most robust model according to a some definition
notion of **Bayesian model selection**: select the a posteriori most probable model according to Bayes' rule

$$\mathbb{P}[Model|Data] = \int_{\Theta} p(Data|\theta)p(\theta|model)d\theta$$

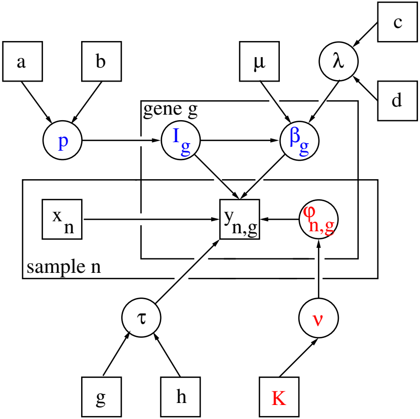
Robustification of the Error model for different methods

The general idea is to provide

1. a set of Gaussian and student's t distributions as error model or
2. a finite mixture model of Gaussian and student's t distributions components as error model for

- linear regression which usually has a Gaussian distribution of the errors ε_i
- Analysis of variance which usually has a Gaussian distribution of the errors ε_i
- linear discriminant analysis
- stochastic processes with Gaussian error etc.

Robust Bayesian ANOVA Model



$y_{n,g}$ ... measurement of gene g and experiment n

I_g ... differential expression indicator of gene g

β_g ... mean expressions vector of gene g

$\varphi_{n,g}$ ... rescaling factor of the t distribution

ν ... t distributions' degrees of freedom

Modelling differential expression

ANOVA model, implemented as a mixture over an indicator I_g

$$\begin{aligned} \beta_g | I_g &\sim I_g N_S(\mu, \tau^{-1} E_S) + (1 - I_g) N_S(\mu, \tau^{-1} \cdot \mathcal{I}_S) \\ I_g | p &\sim Bin(1, p) \end{aligned}$$

Hypothesis 1: no differential expression, i. e. all mean expressions $\beta_{g;s}$ are equal
Hypothesis 2: differential expression, i. e. at least two mean expressions $\beta_{g;s}, \beta_{g;s^*}$ differ

Model overview

$$
\begin{aligned}
y_{n,g} &\sim N\left(x_{n,g}^T\beta_g, (\varphi_{n,g}\tau_\varepsilon)^{-1}\right)\\
\beta_{g,0}|I_g = 0 &\sim N_1(\mu, (\lambda)^{-1})\\
\beta_g|I_g = 1 &\sim N_S(\mu, (\lambda)^{-1}E_S)\\
\lambda &\sim Ga(c, d)\\
\tau_\varepsilon|g, h &\sim Ga(g, h)\\
\varphi_{n,g}|\nu &\sim Ga(\frac{\nu}{2}, \frac{\nu}{2})\\
\nu &\sim U_{\mathfrak{N}}\\
I_g|p &\sim Bin(1, p)\\
p &\sim Be(a, b)
\end{aligned}
$$

Likelihood Robustification

Parameters $(y_{n,g}, \varphi_{n,g})$ follow a bivariate Normal-Gamma distribution

$$y_{n,g}, \varphi_{n,g} | \beta_g, \tau, \nu \sim NormalGamma(x_n^T \beta_g, \tau^{-1}, \frac{\nu}{2}, \frac{\nu}{2})$$

which is defined as

$$\begin{aligned} y_{n,g} | \beta_g, \tau, \varphi_{n,g} &\sim N(x_n^T \beta_g, (\varphi_{n,g} \tau)^{-1}) \\ \varphi_{n,g} | \nu &\sim Ga(\frac{\nu}{2}, \frac{\nu}{2}) \end{aligned}$$

Marginal distribution of observations $y_{n,g}$ is a **t distribution**

$$y_{n,g} | \beta_g, \tau, \nu \sim t_\nu(x_{n,g}^T \beta_g, \tau^{-1})$$

Likelihood Robustification: Set of distributions

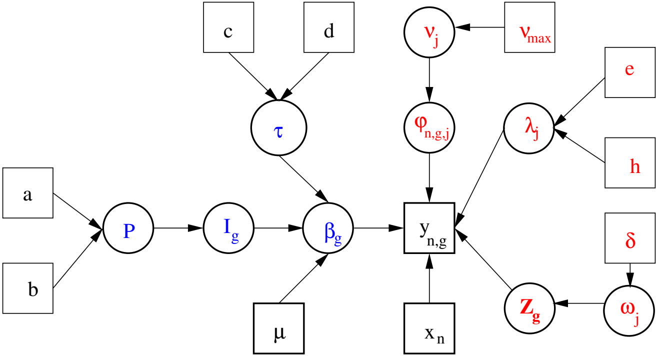
Following the notion of Bayesian likelihood robustness, we select the best fitting model from the **finite set of models**

$$\Gamma = \{\{t_\nu(\mu, \tau^{-1}), \nu \in \mathfrak{N} \setminus \{\nu_{max}\}\}, N(\mu, \tau^{-1})\}$$

Possible noise models include:

- *Cauchy* distribution ($\nu = 1$)
- *non-central t* distributions with various degrees of freedom
- *Normal* distribution ($\nu \to \infty$), represented by $\nu = \nu_{max}$

Mixture model in bioinformatics context



$Z_g \in \{1, \ldots, J\}$ ... mixture component label

componentwise precision λ_j, rescaling factor of the t distribution $\varphi_{n,g,j}$ and degrees of freedom ν_j

ANOVA model of Gauss-t-mixtures for microarrays

$$\begin{aligned}
y_{n,g} &\sim \sum_{j=1}^{J} \omega_j f(X\beta_g, \lambda_j, \nu_j) \\
y_{n,g}|Z_i = j &\sim N(X\beta_g, \frac{1}{\lambda_j \varphi_{n,g,j}}) \\
(N_1, \ldots, N_J) &\sim MN_{N;(\omega_1,\ldots,\omega_J)} \quad N = \sum_{j=1}^{J} N_j \\
\omega = (\omega_1, \ldots, \omega_J) &\sim Dir(\delta, \ldots, \delta) \\
\varphi_{n,g,j}|\nu_j &\sim Ga(\frac{\nu_j}{2}, \frac{\nu_j}{2}) \\
\lambda_j &\sim Ga(e, h) \\
\nu_j &\sim U_{\{[1,\nu_{max}],\infty\}}
\end{aligned}$$

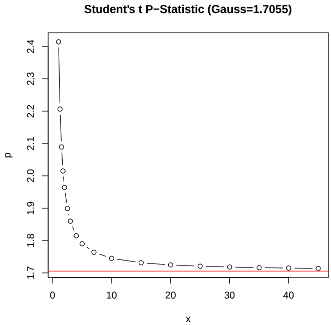
Our Markov Chain Monte Carlo Algorithm

Hybrid MCMC algorithm

- **Metropolis-Hastings**: update of ν and Z_g (partially collapsed)
- **Reversible-Jump MCMC**: (β_g, I_g), change between student's t and Gaussian distribution
- **Gibbs**: rest

Implementation for R available on `https://github.com/alexposekany/RobBayMA`

Measuring ‘peakedness’



Motivation: kurtosis does not exist for student’s t distributions with $\nu \leq 4$ ⇒ measure for non-Gaussianity with ‘peakedness’

$$Peak(g) = \sum_{i=1}^{N} \frac{quant(0.975; \nu_i^{(g)}) - quant(0.025; \nu_i^{(g)})}{quant(0.875; \nu_i^{(g)}) - quant(0.125; \nu_i^{(g)})}$$

Microarray quality control

Bayes factor for the ‘t-ness’ compared to Gaussianity

$$BF_{n,g} = \frac{\mathbb{P}[y_{n,g}|\nu = \hat{\nu}_g, \beta_g = \hat{\beta}_g, \lambda = \hat{\lambda}_g]\mathbb{P}[\nu = \hat{\nu}_g]}{\mathbb{P}[y_{n,g}|\nu = \infty, \beta_g = \hat{\beta}_g, \lambda = \hat{\lambda}_g]\mathbb{P}[\nu = \infty]}.$$

Test for equal distribution of 'overdispersed' values based on multinomial assumption which leads to a Dirichlet distribution of the relative amount of overdispersed values on the arrays

$$\begin{aligned} H_0: \quad \pi &= (1/N, \dots, 1/N) \\ H_A: \quad \pi &\neq (1/N, \dots, 1/N) \end{aligned}$$

$$SDR = \frac{p(\pi = (1/N, \dots, 1/N)|(\alpha_1^*, \dots, \alpha_N^*))}{p(\pi = (1/N, \dots, 1/N)|(\alpha_0, \dots, \alpha_0))}$$

empirical Bayes prior: $\alpha_0 = 2.5 \cdot N$

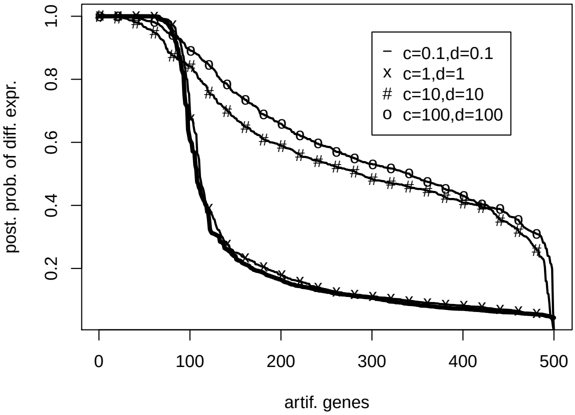
Calibrating the algorithms

Tested aspects of the algorithm:

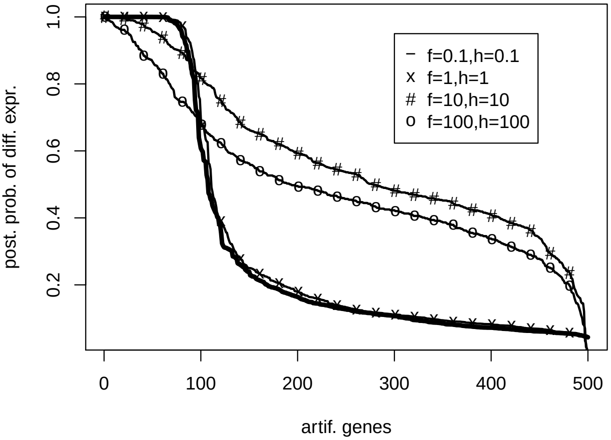
- **sensitivity analysis**

  Choice of hyper parameters influences inference results
  → how? in which interval is inference robust?
  locally robust in reasonably large neighbourhood of chosen parameters

Sensitivity Analysis



Sensitivity analysis



Calibrating the algorithms

Tested aspects of the algorithm:

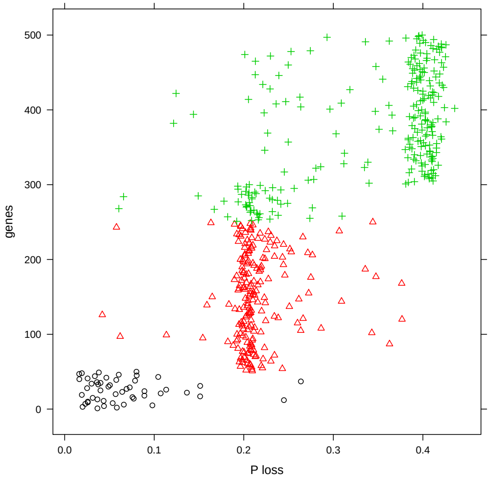
- **sensitivity analysis**

  Choice of hyper parameters influences inference results
  → how? in which interval is inference robust?
  locally robust in reasonably large neighbourhood of chosen parameters

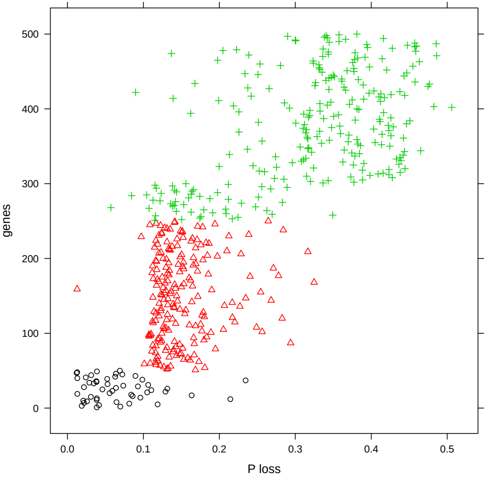
- **can algorithm recognise noise model?**

  Data simulated from t or normal distributions
  estimates of degrees of freedom are accurate and precise

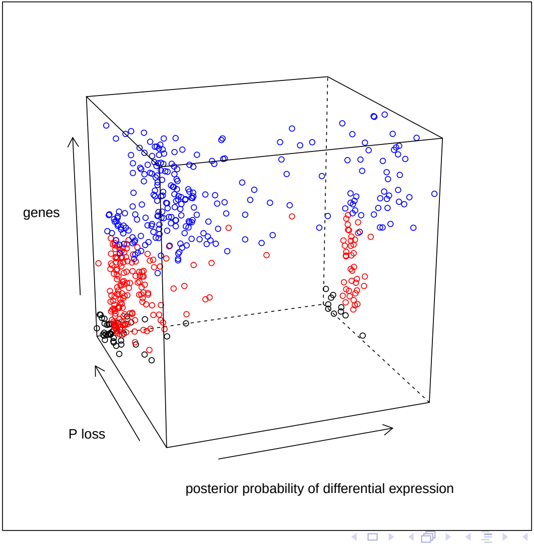
Test Results for normal, t_4, t_1



Test Results for normal, t_{10}, t_1



Recognising the distributions



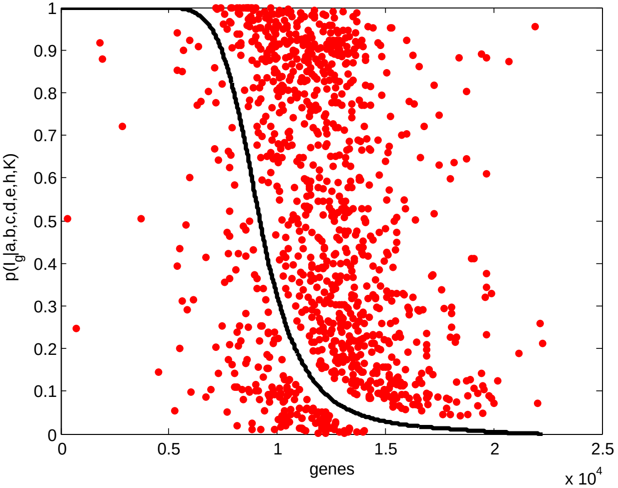
Results for Microarrays

- All 14 considered data sets prefer a **t distribution with low degrees of freedom**, with posterior *mean degrees of freedom of* ∼ 1 - 4 (independent of the used normalisation method). → compare with Hardin and Wilson (2009) and Novak et al. (2006)
- the lists of genes and Gene Ontology terms generated by t and Gaussian models **differ significantly**
- in mixtures with few interpretable components (2-3) no Gaussian components show up at all, discarding the idea that 85% of the data are Gaussian (Novak et al. (2006))

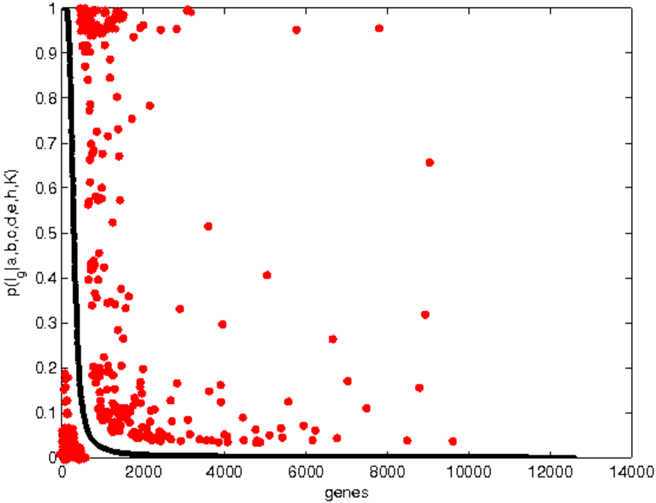
Table of Results

GEO ID	arrays	mean dfs ($\overline{\nu}$)			diff./cmn.	diff./cmn.
		vsn	loess	quant.	genes	GO terms
GDS3216	12	5	2	1	150/1176	78/111
GDS3225	4	6	1	1	290/832	21/161
GDS1404	10	14	1	1	136/1776	14/11
GDS1686	9	4	3	3	174/136	96/11
CAMDA 08	24	4	1	1	304/400	67/26
GDS1375	70	3	1	1	3561/6861	316/160
GDS810	31	4	1	1	135/72	51/9
GDS2960	101	4	3	3	166/318	2/51
GDS3221	24	4	3	3	119/180	52/108
GDS3162	10	4	1	1	446/797	66/112
GDS1555	8	4	1	1	183/131	110/24
GDS2946	15	5	2	2	157/146	306/14
GDS972	44	5	1	1	163/369	71/94

Comparing Results



Comparing Results



Results for array quality control

	Scenario 1		Scenario 2	
40%- 50%	χ^2 test	SDR	χ^2 test	SDR
$t_4 - t_{10}$; 5%	0.15	2.64	7.2e-11	7.4e-07
$t_4 - t_{10}$; 10%	0.13	2.07	4.0e-33	1.6e-19
$t_4 - t_{10}$; 15%	0.47	12.56	3.4e-19	2.1e-12
$t_4 - t_1$; 5%	0.69	29.26	1.5e-14	3.8e-09
$t_4 - t_1$; 10%	0.22	4.03	2.8e-29	9.5e-18
$t_4 - t_1$; 15%	0.40	9.19	1.6e-13	3.3e-09
$t_7 - t_1$; 5%	0.33	8.27	1.2e-14	4.3e-09
$t_7 - t_1$; 10%	0.72	33.44	1.3e-38	2.0e-22
$t_7 - t_1$; 15%	**0.88**	**58.45**	**1.5e-20**	**2.9e-13**
$t_{10} - t_1$; 5%	0.53	15.83	1.1e-03	5.5e-02
$t_{10} - t_1$; 10%	**0.07**	**0.90**	1.7e-04	7.4e-03
$t_{10} - t_1$; 15%	0.18	2.85	**4.3e-03**	**1.1e-01**

Scenario 1: equally distributed extreme BFs among all arrays
Scenario 2: 75% of the genes containing the most extreme 10% of BFs are accumulated on a single array, 25% are distributed among the other arrays

error model unchanged for each gene, only the label of the array changes

Conclusions

In many fields Gaussian error models are still standard lakelihood choices for computational conveniencs which can lead to erroneous results and conclusions

replace Gaussian methods by

- **non-parametric**
  sample size and computational burden are limiting factors for Bayesian (and non-Bayesian) non-parametric methods or
- leaving number of mixture components small and fixed allows for differentiation w.r.t. **peakedness** (heavy tails) and thus a **'quality interpretation'**.

future work:

- stand alone noise model and implementation compatible with sampling schemes and packages
- generalising to skewed normal and t-distributions for covering the **skewness** with the error model and not the heavy-tailed component and adjusting for bias

Thank you!

NOVAMATH
CENTER FOR MATHEMATICS + APPLICATIONS

REPÚBLICA PORTUGUESA

References

Posekany A. **Outlier detection in Bioinformatics with Mixtures of Gaussian and heavy-tailed distributions**. *Data Science - Analytics and Applications* 2021
RobBayMA: `https://github.com/alexposekany/RobBayMA` 2021
Posekany A, Felsenstein K and Sykacek P. **Biological assessment of robust noise models in microarray data analysis**. *Bioinformatics* 2011
Rao M, et al. **Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies**. *Front. Genet.* 2019 Huber W, Heydabreck A, Vingron M. **Error models for microarray intensities**. Bioconductor Project Working Papers, 2004
Gottardo R, et al. **Bayesian robust inference for differential gene expression in microarrays with multiple samples**. *Biometrics* 2006
Novak J, et al. **Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution**. *Biol. Direct* 2006
Sun Z, Kuczek T, Zhu Y. **Statstical Calibration of QRT-PCR, Microarray and RNA-SEQ gene expression data with measurement error models** . *The Annals of Applied Statistics* 2014