New Measurement Error Data Structures

Raymond J. Carroll

Texas A&M University and University of Technology Sydney

August 15, 2016

Outline

- **Co-authors** — How many have a Carroll number of 1: Over 10 for sure!
- **A Tour** — Recent group papers
- **Chernobyl** — Radiation dose models
- **Time-varying exposures** — diet, physical activity, etc.
- **MIMIC Models** — Latent variable models
- **Semiparametric deconvolution** — Accounting for heteroscedasticity

Other MEM Structures

Continuous and Discrete Variables Measured with Error

- With Grace Yi, Yanyuan Ma and Donna Spiegelman
- Basically, after modeling the misclassification, suggests functional methods such as SIMEX

Multivariate Mixtures of Continuous and Excess Zero Data

- With the NCI folks, including Victor Kipnis
- The analysis of the Healthy Eating Index
- Includes surveillance and epidemiology

Why it is Silly to Focus on Single Foods or Single Physical Activity Measures

- The NYT had an article about the lack of clear results in nutritional epidemiology
- That is because the one food/nutrient at a time approach is so stupid
- Diet is multivariate: get over it!

Model Averaging and Selection With Variables Measured With Error

- With Xinyu Zhang, Haiying Wang and Yanyuan Ma

Percentage of Non-Consumers With Measurement Error

- With Anindya Bhadra, Victor Kipnis, and others
- Some people never consume alcoholic beverages, red meat, etc.
- How many?

Measurement Error Models with Interactions

- With Doug Midthune, Victor Kipnis and Larry Freedman

Calibration of Different Labs When There is Measurement Error

- With Mitch Gail, Molin Wang and others
- All about measurement error and the calibration of data from multiple labs to a common lab

Spatial Statistics and Measurement Error

- Papers with Brent Coull and Louise Ryan
- Coull: SIMEX
- Ryan: Semiparametric regression

Radiation Epidemiology

- Nevada Test Site (NTS) Thyroid Disease Study
- Hanford (Nuclear Site) Thyroid Disease Study
- The Chernobyl Accident
- Household Radon Exposure
- Classical Error Model in Linear Regression

Radiation Epidemiology

- With radioiodine fallout exposure to an individual is difficult to measure
- In many of the studies, exposure is retrospective
- Such studies include sophisticated biological models for the fallout coupled with retrospective assessment of locations, milk consumption, etc.
- In all of these studies, there are classical and Berkson error components

Original Models

- The original models were of multiplicative type.
- True dose is D^{tr} and measured dose is D^{mes}
- They posited a latent variable W and modeled

$$\begin{aligned} \log(D^{\mathrm{tr}}) &= W + U_{\mathrm{berk}}; \\ \log(D^{\mathrm{mes}}) &= W + U_{\mathrm{class}}, \end{aligned}$$

where $(W, U_{\mathrm{berk}}, U_{\mathrm{class}})$ are mutually independent.

Original Models

- Physical models estimated the total variance of $(U_{\text{berk}}, U_{\text{class}})$ but generally the individual variances were unknown and apportioned from the total using sensitivity analyses.

Chernobyl

- Chernobyl is different because measurements were taken repeatedly from children and adults after the accidents
- The data are better than previous studies
- Ukrainian statisticians and physical scientists have put enormous effort into understanding the uncertainty in the measured doses

Figure: In the Chernobyl Exclusion Zone. With Sergii Masiuk (a statistician), Alex Kukush (a well-known MEM person)and Ilya Likhtarov (the chief radiation biologist for Chernobyl since 1986, and how got thyroid cancer from his exposure the day after the "accident".

Chernobyl

- They have recently developed an entirely new model for the true doses
- It has additive classical and multiplicative Berkson doses

Model

- Let D be thyroid dose, Q thyroid radioactivity, and F a variable that depends on radioactivity transition and thyroid mass
- They write that $D^{\text{mes}} = F^{\text{mes}} Q^{\text{mes}}$
- Also, $D^{\text{tr}} = F^{\text{tr}} Q^{\text{tr}}$
- $F^{\text{tr}} = F^{\text{mes}} \delta_F$ (multiplicative Berkson)
- $Q^{\text{mes}} = Q^{\text{tr}} + \sigma(Q^{\text{mes}})\gamma$, $\sigma(Q^{\text{mes}})$ known (Classical, heteroscedastic)

Model

- The model was fit to the Chernobyl, but it is hard to get permission to publish the results in a statistics paper

$$\Pr(Y = 1|D^{\text{tr}}) = \frac{\lambda_0 + \text{EAR} \times D^{\text{tr}}}{1 + \lambda_0 + \text{EAR} \times D^{\text{tr}}}$$

- Here EAR is excess absolute risk
- As expected, the effect of ignoring the dose uncertainty is attenuation of the EAR

Reference

Masiuk, S., Shklyar, S., Kukush, A., Carroll, R. J., Kovgan, L. and Likhtarov, I. A. (2017). Estimation of radiation risk in presence of classical additive and Berkson multiplicative errors in exposure doses. *Biostatistics*,17, 422-436.

Time-Varying True Exposure

Time-Varying Truth

- Increasingly, in nutrition and physical activity, longitudinal data are becoming available
- The target exposures (usual dietary intake, usual physical activity) are increasingly recognized as varying with time
- This is especially true when the time frame goes across multiple years

Time-Varying Truth

- The recognition that over longer periods the **measured with error** target changes poses multiple interesting questions
- What is the target?
- What is the analysis of the measurement error properties?
- How do we relate the target exposure to a response?

Data

- I ignore covariates for simplicity
- Let i be the individual. Split time into time periods j.
- Split period j into separate periods k
- Within period j and sub-period k, the mean true values of exposure are T_{ijk}. Let these have a correlation structure.
- Let its average over the sub-periods be $\overline{T}_{ij}$

Model

- Let its average over the sub-periods be $\overline{T}_{ij}$
- Then, somewhat like a FDA model,

$$T_{ijk} = \overline{T}_{ij} + \zeta_{ijk}$$

- Here the ζ_{ijk} are white noise whose variance depends on time period j

Measurement Error

- The T_{ijk} and the true T_{ij} are not observed.
- At this point on, we have a measurement error model. For example, if we see some version of a biomarker, we observe M_{ijk}, where

$$M_{ijk} = T_{ijk} + \delta_{ijk}$$

- Here, the δ_{ijk} might white noise, or some other structure

Other Instruments

- In many applications, there will be other instruments correlated with true exposure
- Food Frequency Questionnaires in nutrition
- Physical Activity Questionnaires in PA research
- Such things can be accommodated in the usual ways

Applications

- In the Validation Studies Pooling Project, for % calories from Protein, time-varying model gave a much better fit than the usual non-time-varying model, by AIC
- Demonstrated that the attenuation is quite a bit worse (lower) than if the time-varying nature is ignored and 24HR recalls are used.

A Note to Collaborators

- This can all be put into the form of a new functional data analysis model

References

Freedman, L. S., Midthune, D., Dodd, K., Carroll, R. J. and Kipnis, V. (2015). A statistical model for measurement error that incorporates variation over time in the target measure, with application to nutritional epidemiology. *Statistics in Medicine*, 34, 3590-3605.

Kipnis, V: talk here in Banff

Rosner, et al. (2008), *Statistics in Medicine*

Keogh, et al. (2013), *Statistics in Medicine*

Latent Variable MIMIC Models

Latent Variable MIMIC Models

- There is a class of latent variables models called **MIMIC**
- Multiple Indicators (**Responses**), Multiple Causes (**Covariates**)

Latent Variable MIMIC Models

- These models are popular in the social science literature
- They require at least 3 responses
- They posit that, in addition to covariates, there is a latent construct, T, that drives the system and makes the responses independent

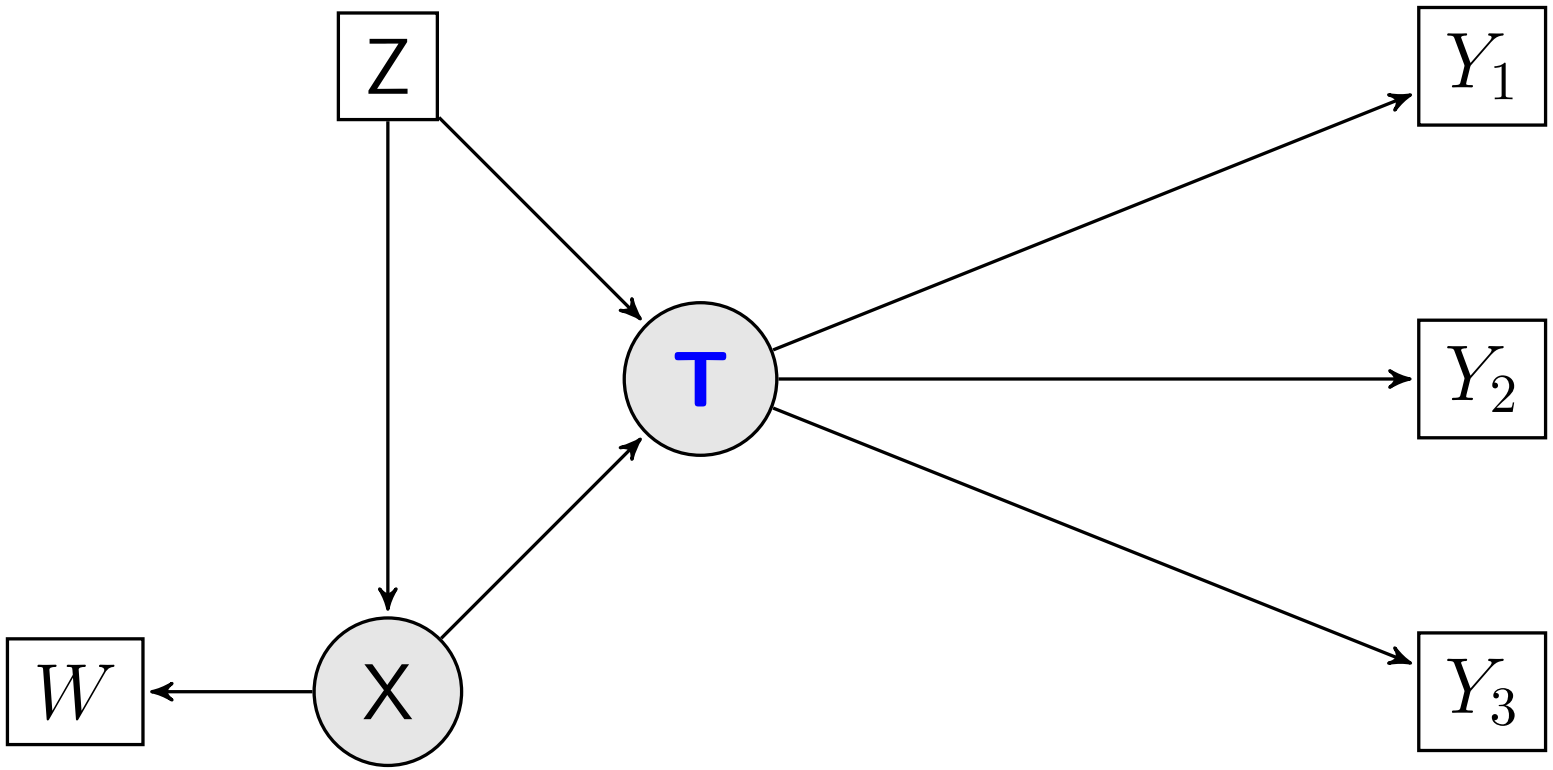

Figure: Schematic. Given Z = vector of error free covariates, X = the covariate measured with error, and **T** = the underlying latent construct, the outcomes Y_j are independent.

Example

- The latent construct is energy expenditure
- In our example, energy expenditure affects 3 responses affiliated with ATP hydrolysis
- There is a mismeasured variable Metabolic Rate.

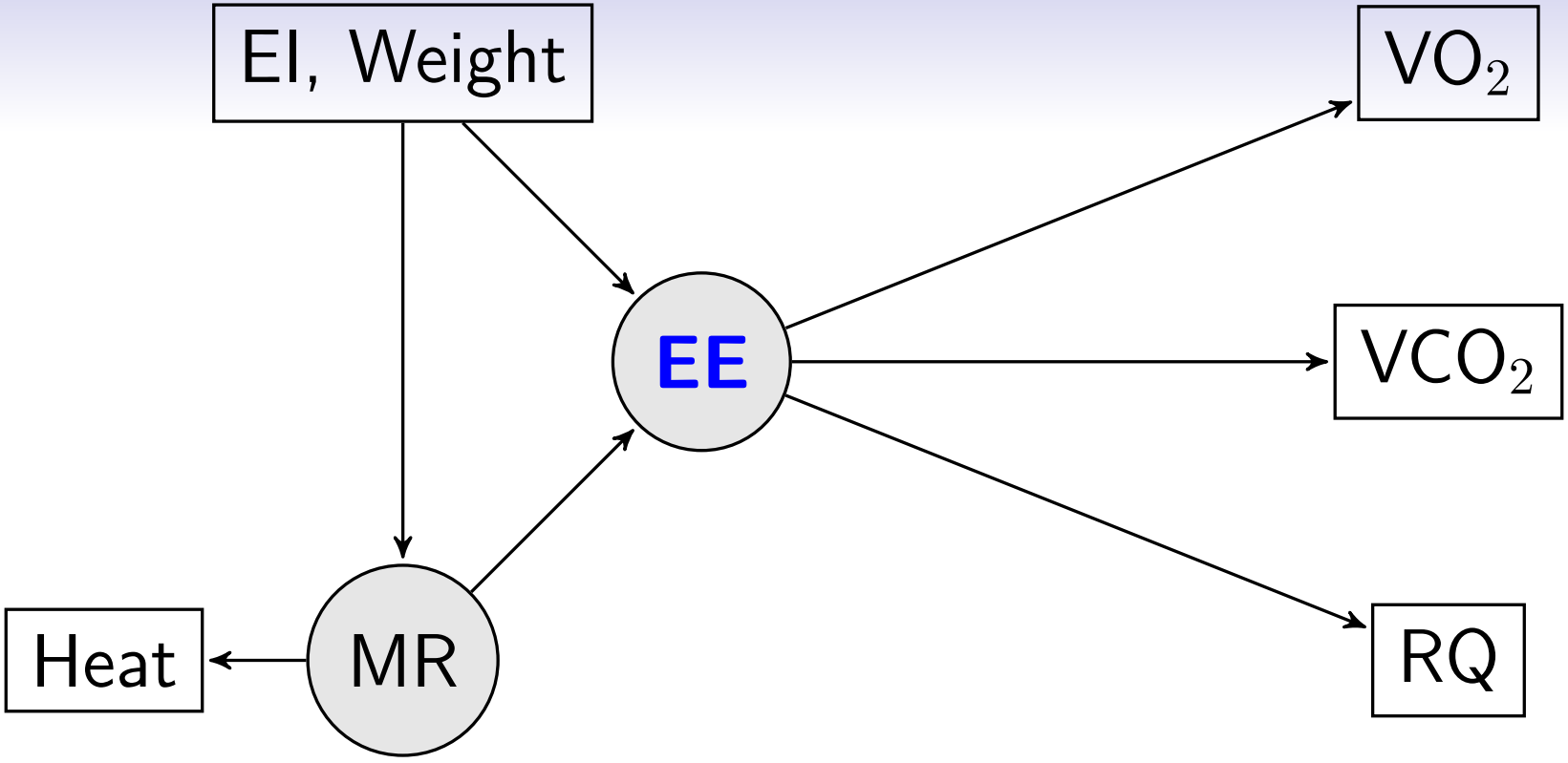

Figure: In our example, "EI" is energy intake, "MR" is metabolic rate and "Heat" is heat production. The latent construct is **Energy Expenditure**. There are 3 outcomes of energy expenditure VO_2, VCO_2) and RQ. Given energy intake, body weight and metabolic rate, the physical measures of energy expenditure at time t are all conditionally independent.

Measurement Error in the Example

- The crucial variable Metabolic Rate is not observed
- Instead, heat production of rats is measured repeatedly over a 24 hour period.
- This is used to estimate the average energy metabolism during a twenty-four hour period
- A classical error model is reasonable

$$HP(t) = MR(t) + U(t)$$

Measurement Error in the Example

- Classical model

$$HP(t) = MR(t) + U(t)$$

- Can cast this into the guise of functional data analysis with random functions $MR(\cdot)$ plus white noise $U(\cdot)$
- We used the **FPCA package in R** to estimate the measurement error variance

Measurement Error in the Example

- The attenuation was ≈ 0.70, roughly the same as in the old Framingham systolic blood pressure example in our book.
- Because of non-trivial correlations among the covariates, the coefficients for energy intake and the latent energy expenditure were numerically and statistically significantly different for the MEM corrected analysis

References

Tekwe, C. D., Carter, R. L., Cullings, H. M. and Carroll, R. J. (2014). Multiple indicators, multiple causes measurement error models. *Statistics in Medicine*, 33, 4469-4481.

Tekwe, C. D., Zoh, R. S., Bazer, F. W., Wu, G. and Carroll, R. J. (2017). Functional multiple indicators, multiple causes measurement error models. In revision.

Semiparametric Deconvolution

Nonparametric Density Deconvolution

- In the classical model

$$W_i = X_i + U_i$$

- The goal is to estimate the density, $f_X(\cdot)$ of X
- There is a substantial nonparametric literature when X_i and U_i are independent, and the distribution of U is known.
- Requires that U is homoscedastic

Nonparametric Density Deconvolution

- In the classical model

$$W_i = X_i + U_i$$

- The usual tool is called **deconvoluting kernel density estimators**, which are a type of measurement error corrected score
- Originally derived by Len Stefanski (I was a co-author)
- This means that the expectation of them given X is the same as a usual kernel density estimator

Nonparametric Density Deconvolution

- Aurore Delaigle has Matlab and R programs (written by one of my students) to implement these methods
- **Do not use the decon package in R**!!!!!
- There are papers that allow the measurement error distribution to be unknown but estimated by replications (Hall, Ma, Delaigle)
- There is also a paper (Hall, Delaigle) that makes shape constraints and such and does not need replications

Nonparametric Density Deconvolution

- There is also a small literature on deconvolution with heteroscedasticity

$$W_i = X_i + \sigma_i U_i$$

- As far as I know, it requires the σ_i to be **known** (or well-estimated) and independent of U_i

Nonparametric Regression and Density Deconvolution

- These methods have been extended to nonparametric regression

Semiparametric Regression and Density Deconvolution

- A model that encompasses all these others is

$$\begin{aligned} Y_i &= m(X_i) + \sigma_\epsilon(X_i)\epsilon_i; \\ W_{ij} &= X_i + \sigma_u(X_i)U_{ij}. \end{aligned}$$

- Staudenmayer, Ruppert and Buonaccorsi developed a method for the density equation when the U_{ij} are normally distributed
- Both the variance function and $f_X(\cdot)$ were modeled by Bsplines.

Semiparametric Regression and Density Deconvolution

- A model that encompasses all these is

$$\begin{aligned} Y_i &= m(X_i) + \sigma_\epsilon(X_i)\epsilon_i; \\ W_{ij} &= X_i + \sigma_u(X_i)U_{ij}. \end{aligned}$$

- We make no assumptions about distributions
- We have developed semiparametric Bayesian methods for this general problem
- We have R software
- The density part has been extended to multivariate X

Semiparametric Regression and Density Deconvolution

- We used Bsplines to estimate $\sigma_\epsilon(x)$ and $\sigma_u(x)$
- The density of X is modeled as either an infinite mixture of normals (Dirichlet process mixture models) or a finite mixture, both with the number of mixtures unknown
- While not completely nonparametric, the methods are extremely flexible, and in simulations out-perform all other methods

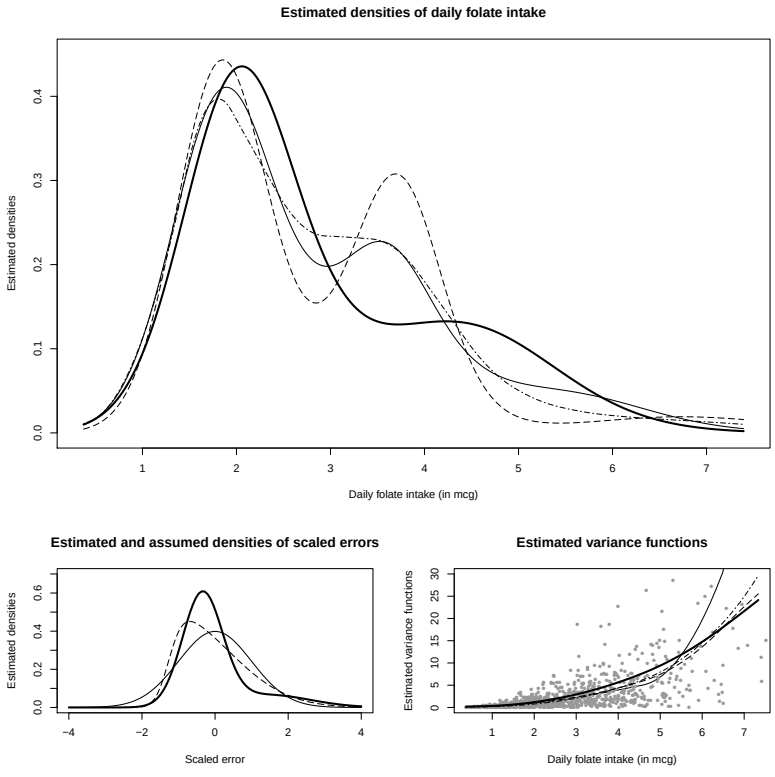

Figure: Density deconvolution for daily folate intakes. Thick solid lines uses the DPPM. Others are based on assuming distributions for the errors.

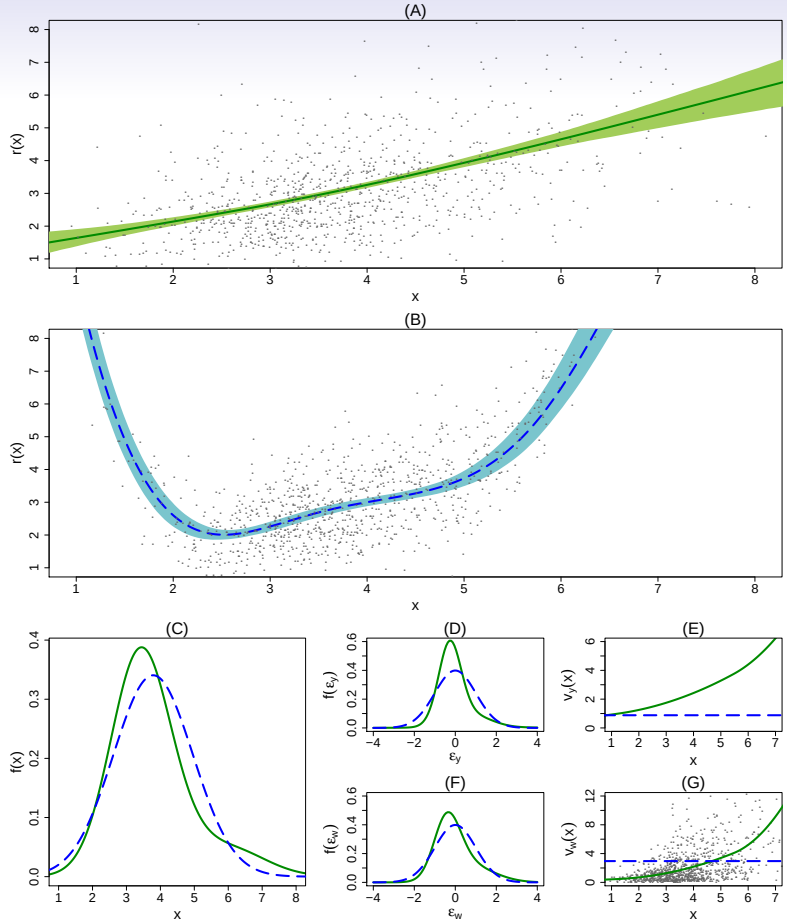

Figure: Regression of FFQ sodium on long-term sodium intake. In all panels the solid lines represent the estimates obtained by our method and the dashed lines represent the estimates obtained by the method of Berry, et al. (2002). (A) The regression function estimated by our method; (B) the regression function estimated by the BCR method; (C) covariate density; (D) regression error density (E) variance function of ϵ; (F) density of measurement errors. (G) variance function of the measurement errors.

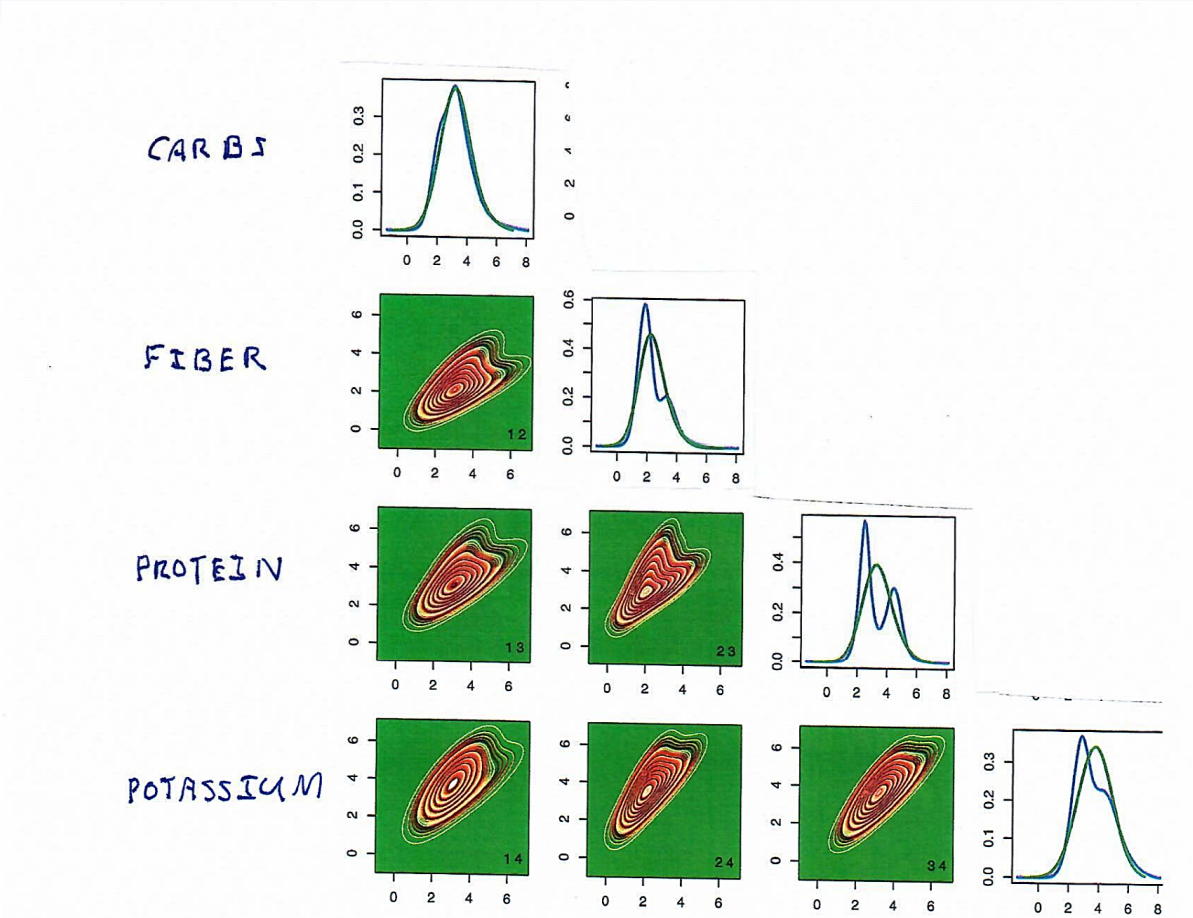

Figure: Densities of usual intake in the EATS data. Off-diagonals are the bivariate densities.

My References

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 165-184.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184-1186.

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D. and Carroll, R. J. (2014). *Journal of Computational and Graphical Statistics*, 25, 1101-1125.

Sarkar, A., Mallick, B. K. and Carroll, R. J. (2014). *Biometrics*, 70, 823-834.

Sarkar, A., Pati, D., Mallick, B. K. and Carroll, R. J. (2017). Bayesian semiparametric multivariate density deconvolution. In revision