Astrostatistics: Thu 08 Feb 2017

https://github.com/CambridgeAstroStat/PartIII-Astrostatistics

- Examples Classes (sheets provided ~1 week prior)
    - Fri Feb 16, Fri Mar 2, Wed Mar 14 (2:30 pm, Room MR5)
    - One more + Revision Class in Easter Term
- Fitting Statistical Models to Astronomical Data
    - Frequentist —> Bayes, Overview of Bayes, examples
    - Bayesian Inference: Ivezic, Ch 5, F&B Ch 3, Gelman BDA
    - Hogg, D., 2012. "Data analysis recipes: Probability calculus for inference." https://arxiv.org/abs/1205.4446

Cool App: Bayesian Inference of the Dimensionality of Spacetime

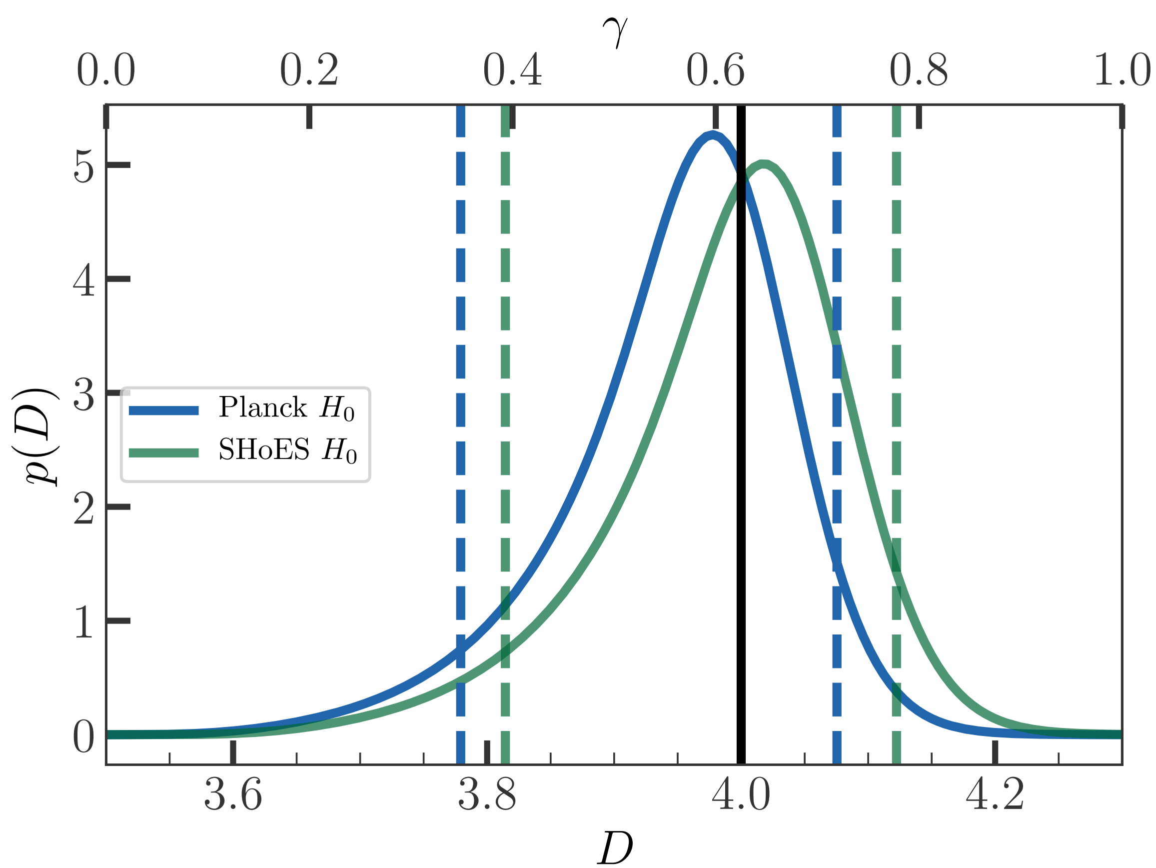


FIG. 1. Posterior probability distribution for the number of spacetime dimensions, D, using the GW distance posterior to GW170817 and the measured Hubble velocity to its host galaxy, NGC 4993, assuming the H_0 measurements from [21] (blue curve) and [22] (green curve). The dotted lines show the symmetric 90% credible intervals. The equivalent constraints on the damping factor, γ, are shown on the top axis. GW170817 constrains D to be very close to the GR value of $D = 4$ spacetime dimensions, denoted by the solid black line.

Complete in Examples Sheet

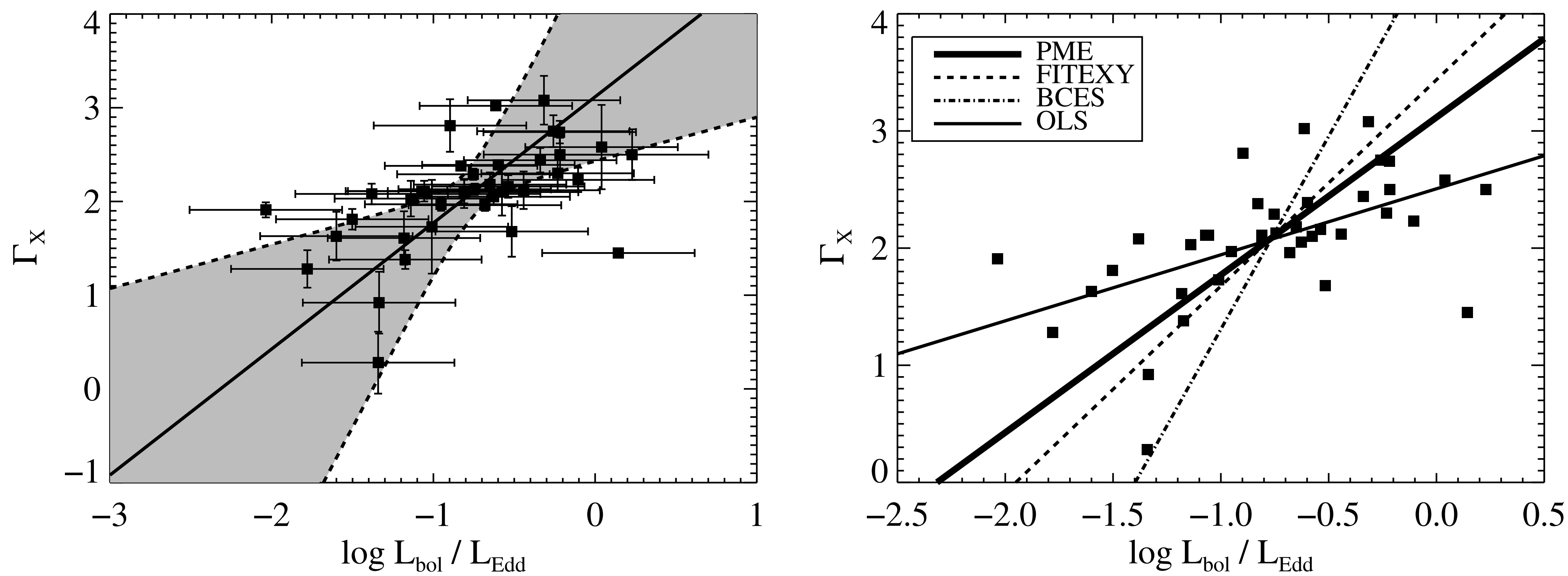


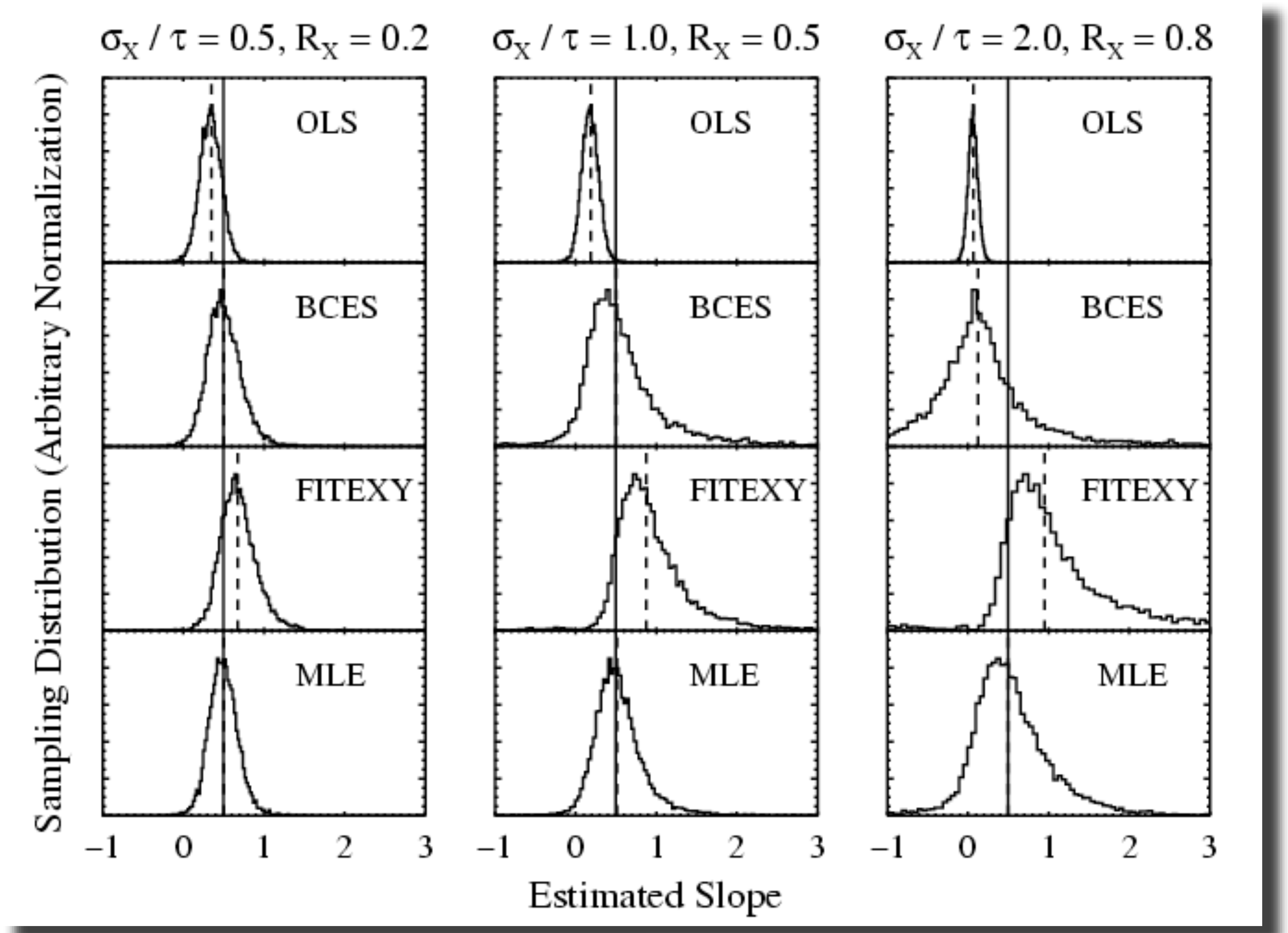
FIG. 10.—X-ray photon index Γ_X as a function of $\log L_{bol}/L_{Edd}$ for 39 $z \lesssim 0.8$ radio-quiet quasars. In both plots, the thick solid line shows the posterior median estimate (PME) of the regression line. In the left panel, the shaded region denotes the 95% ($2\,\sigma$) pointwise confidence intervals on the regression line. In the right panel, the thin solid line shows the OLS estimate, the dashed line shows the FITEXY estimate, and the dot-dashed line shows the BCES($Y|X$) estimate; the error bars have been omitted for clarity. A significant positive trend is implied by the data.

Modelling heteroskedastic, correlated measurement errors in both y and x, intrinsic scatter, nondetections, selection effects

B. Kelly et al. 2007, “Some Aspects of Measurement Error in Linear Regression of Astronomical Data.” ApJ, 665, 1489

Complete in Examples Sheet

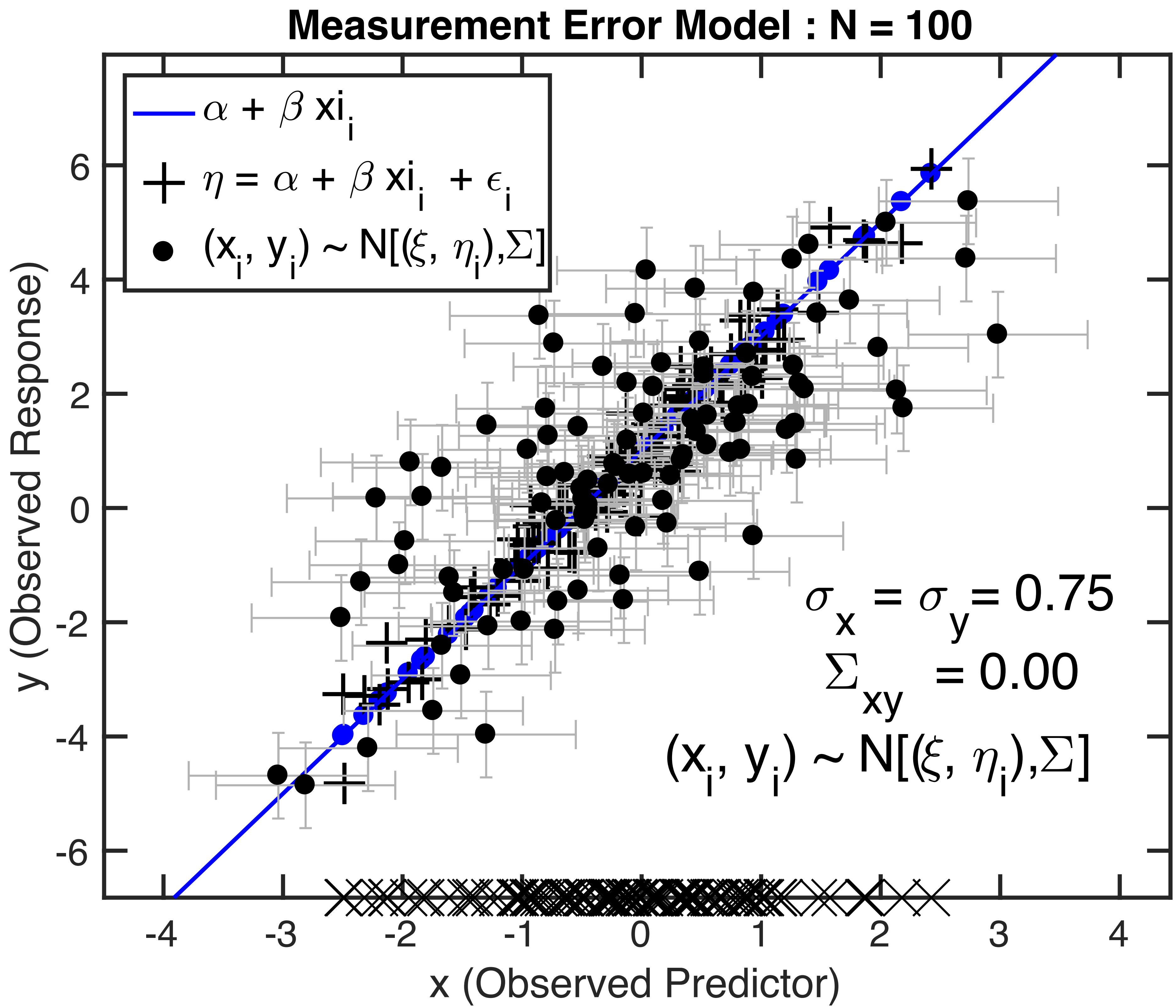
Simulation Study: Slope



Dashed lines mark the median value of the estimator, solid lines mark the true value of the slope. Each simulated data set had 50 data points, and y-measurement errors of $\sigma_y \sim \sigma$.

http://astrostatistics.psu.edu/su07/kelley_measerr07.pdf

Probabilistic Generative Modelling



Probabilistic Generative Model

1. Population Distribution $\xi \sim N(\mu|\tau^2)$

Population Parameters: $\boldsymbol{\psi} = (\mu, \tau)$

2. Regression: $\eta_i | \xi_i \sim N(\alpha + \beta x_i, \sigma^2)$

Regression Parameters: $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$

Latent (true) Variables: (ξ_i, η_i)

3. Measurement Error: $[x_i, y_i] | \xi_i, \eta_i \sim N([\xi_i, \eta_i], \Sigma])$

Observed Data: (x_i, y_i)

Formulating Likelihood Function: Marginalising (integrating out) latent variables

"Complete Data Likelihood" (one datum)

$$P(x_i, y_i, \xi_i, \eta_i | \boldsymbol{\theta}, \boldsymbol{\psi}) = P(x_i, y_i | \xi_i, \eta_i) P(\eta_i | \xi_i, \boldsymbol{\theta}) P(\xi_i | \boldsymbol{\psi})$$

Measurement Error

Regression

Population Distribution

"Observed Data Likelihood" (one datum): integrate out latent variables

$$P(x_i, y_i | \boldsymbol{\theta}, \boldsymbol{\psi}) = \int \int P(x_i, y_i | \xi_i, \eta_i) P(\eta_i | \xi_i, \boldsymbol{\theta}) P(\xi_i | \boldsymbol{\psi}) \, d\xi_i \, d\eta$$

Observed Data Likelihood (all data):

$$P(\boldsymbol{x}, \boldsymbol{y} | \, \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} P(x_i, y_i | \, \boldsymbol{\theta}, \boldsymbol{\psi})$$

Knowns and Unknowns

Population Parameters: $\boldsymbol{\psi} = (\mu, \tau)$

Regression Parameters: $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$

Latent (true) Variables (ξ_i, η_i)

Observed Data: (x_i, y_i)

In Frequentist Statistics, distinction btw data and parameters: parameters are fixed and unknown, but not "random". Only "data" are random realisations of random variables

Knowns and Unknowns

What is the nature of the latent variables (ξ_i, η_i) ?

They have a probability distribution:

$$(\xi_i, \eta_i) \sim P(\xi_i, \eta_i | \boldsymbol{\theta}, \boldsymbol{\psi}) = P(\eta_i | \, \xi_i, \boldsymbol{\theta}) P(\xi_i | \, \boldsymbol{\psi})$$

Often called “nuisance parameters”
parameters you need to introduce to complete the model
but are not the parameters of interest: $(\boldsymbol{\theta}, \boldsymbol{\psi})$

Also referred to as “missing data”
quantities that you didn’t observe, but wish you had!
but relate to actual measurements (x,y)

Are the latent variables “data” or “parameters”?

Bayesian viewpoint

- There is a symmetry between data D and parameters θ - both are random variables described by probability distributions
- Actually they are described by a joint probability $P(D, \theta)$
- Data are random variables whose realisations are observed, parameters are RVs not observed
- Goal is to infer the unobserved parameters from the observed data using the rules of probability:
- Conditional Probability: $P(\theta|\, D) = P(D, \theta)/P(D)$
- Bayes' Theorem: $P(\theta|\, D) = P(D|\, \theta)P(\theta)/P(D)$
- Probability interpreted as degree of belief / uncertainty in hypotheses

Bayes' Theorem

Joint Probability of Data and Parameters:

$$P(D,\theta) = P(D|\theta)P(\theta) = P(\theta|D)P(D)$$

Probability of Parameters Given Data:

$$P(\theta|D) = P(D|\theta)P(\theta)/P(D)$$

Normalisation Constant

Posterior probability: Degree of Belief

Likelihood: Sampling Distribution

Prior probability: Degree of Belief

Simple Gaussian Example