Astrostatistics: Tue 05 Feb 2017

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

- Examples Classes (sheets provided ~1 week prior)
  - Fri Feb 16, Fri Mar 2, Wed Mar 14 (1pm, RoomTBD)
  - One more + Revision Class in Easter Term
- Fitting Statistical Models to Astronomical Data
  - Generative / Latent Variable Modeling / Bayes
  - Hogg, Bovy & Lang. “Data analysis recipes: Fitting a model to data”. https://arxiv.org/abs/1008.4686

Statistical Modelling Wisdom

- Have an objective function [e.g. Likelihood or posterior] that you optimise or sample to fit the data - not just a procedure/recipe
- Objective function helps you evaluate relative fits of data with under different parameter values / models
- Derive your objective function from your modelling assumptions (physical or statistical)
- Write down your assumptions!
- First question: what is the likelihood $L(\theta)$? Derive it from the assumptions underlying your sampling distribution $P(D \mid \theta)$!
- Second question: what is your prior $P(\theta)$? (if Bayesian)
- Third question: How do I optimise/sample objective function to fit the data?

Fitting Models to Astro Data

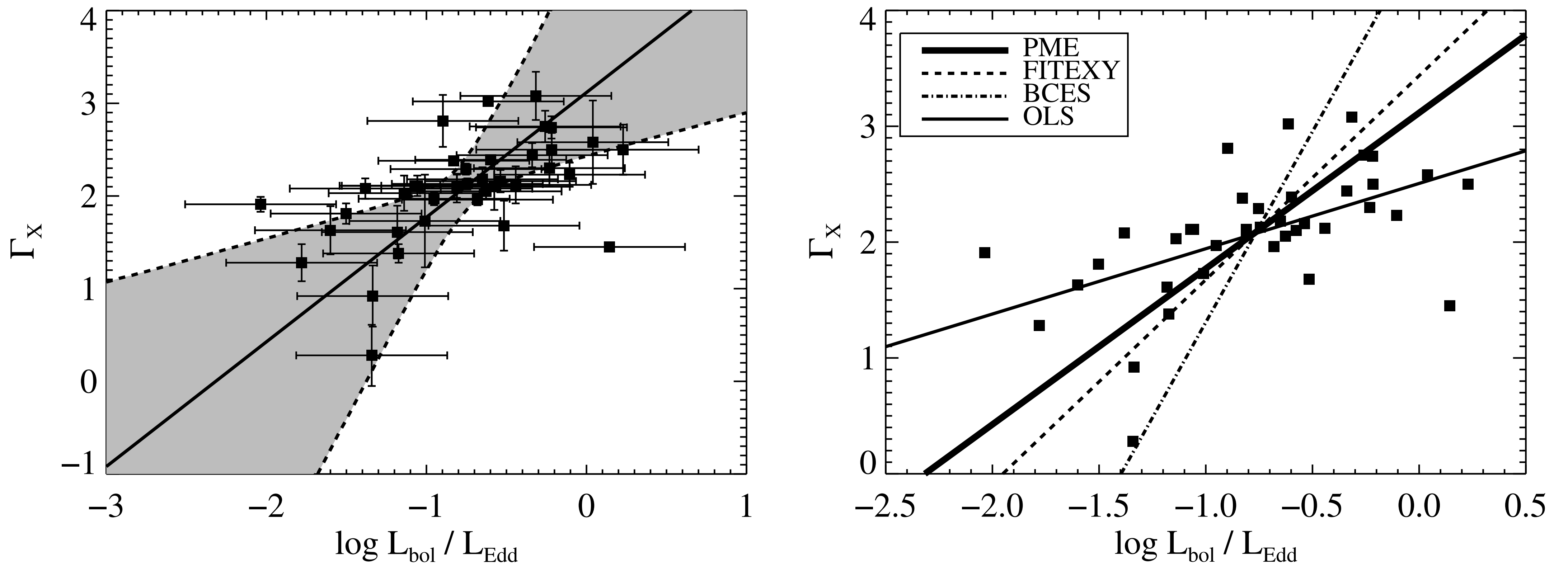


FIG. 10.—X-ray photon index Γ_X as a function of $\log L_{bol}/L_{Edd}$ for 39 $z \lesssim 0.8$ radio-quiet quasars. In both plots, the thick solid line shows the posterior median estimate (PME) of the regression line. In the left panel, the shaded region denotes the 95% (2 σ) pointwise confidence intervals 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

Ad-hoc “$\boldsymbol{\chi}^2$” approaches vs. Likelihood formulation

FITEXY Estimator

- Press et al.(1992, *Numerical Recipes*) define an ‘effective χ^2’ statistic:

$$\chi^2_{EXY} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2_{y,i} + \beta^2 \sigma^2_{x,i}}$$

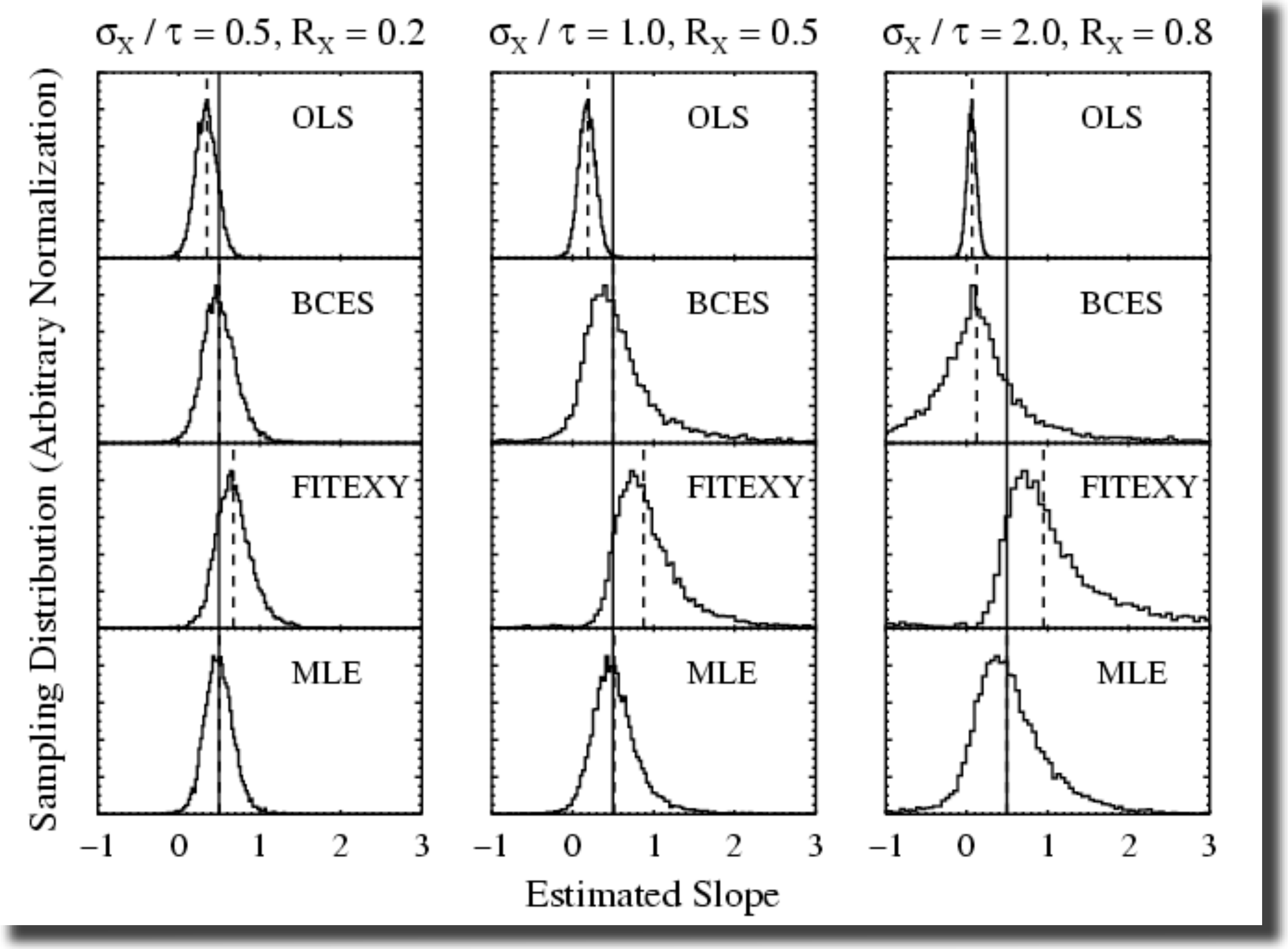
- Choose values of α and β that minimize χ^2_{EXY}
- Modified by Tremaine et al.(2002, ApJ, 574, 740), to account for intrinsic scatter:

$$\chi^2_{EXY} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2 + \sigma^2_{y,i} + \beta^2 \sigma^2_{x,i}}$$

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

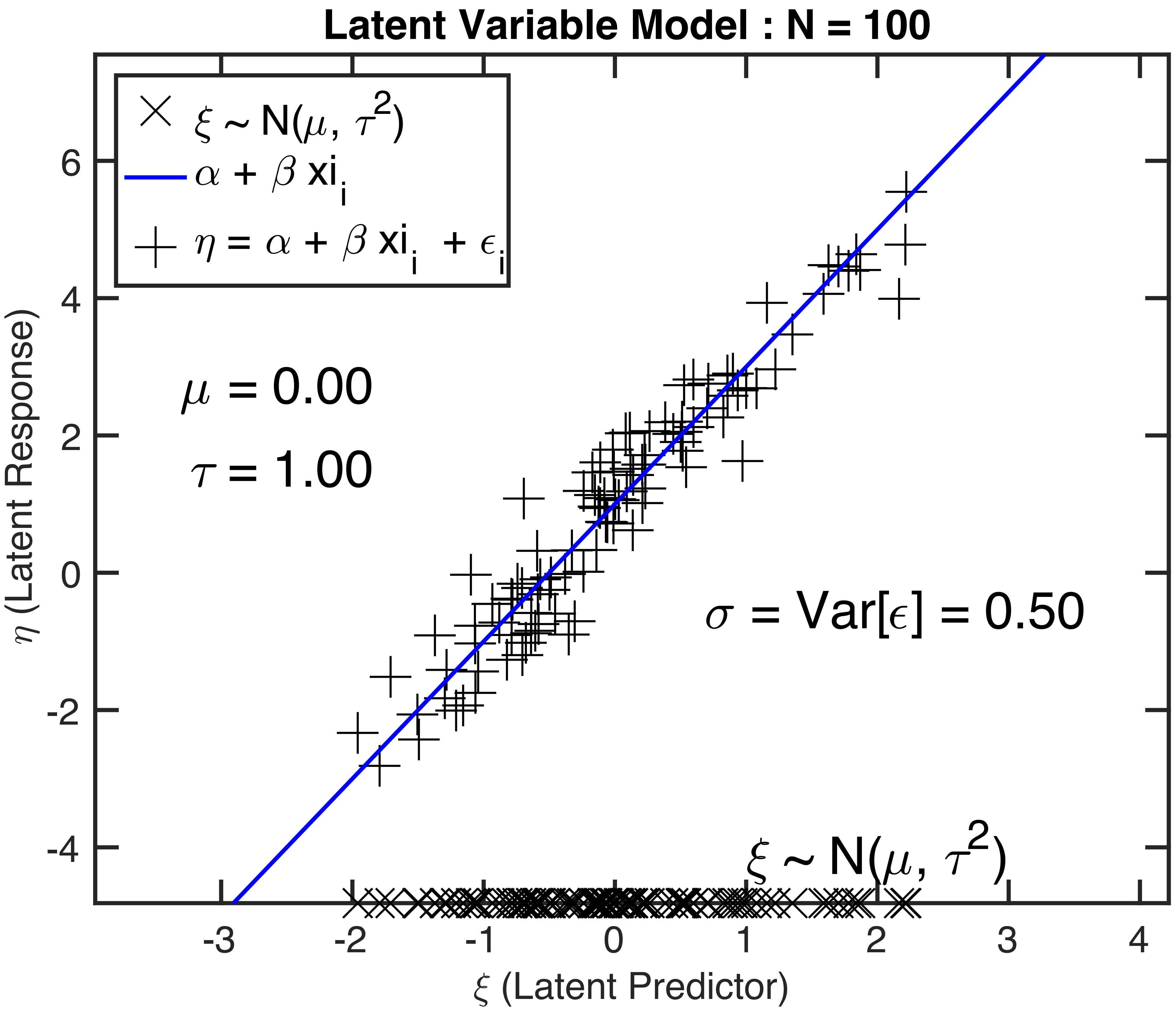
Kelly et al. 2017, Latent Variable Likelihood approach vs. Bad

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


Latent Variable Model : N = 100
$\xi \sim$ N(μ, τ^2)
α + β xi$_i$
η = α + β xi$_i$ + ϵ_i
μ = 0.00
τ = 1.00
σ = Var[ϵ] = 0.50
$\xi \sim$ N(μ, τ^2)
6
4
2
0
-2
-4
η (Latent Response)
-3
-2
-1
0
1
2
3
4
ξ (Latent Predictor)

Probabilistic Generative Modelling

- Forward Model comprises series of probabilistic steps describing conceptually how the observed data was generated from the parameters of interest

- Can introduce intermediate parameters / unobserved latent variables $\boldsymbol{\alpha}$ (e.g. true values corresponding to the observed data).

- From Forward model, derive the sampling distribution, e.g. $P(D \mid \theta) = \int P(D \mid \boldsymbol{\alpha})\, P(\boldsymbol{\alpha} \mid \theta)\, d\boldsymbol{\alpha}$

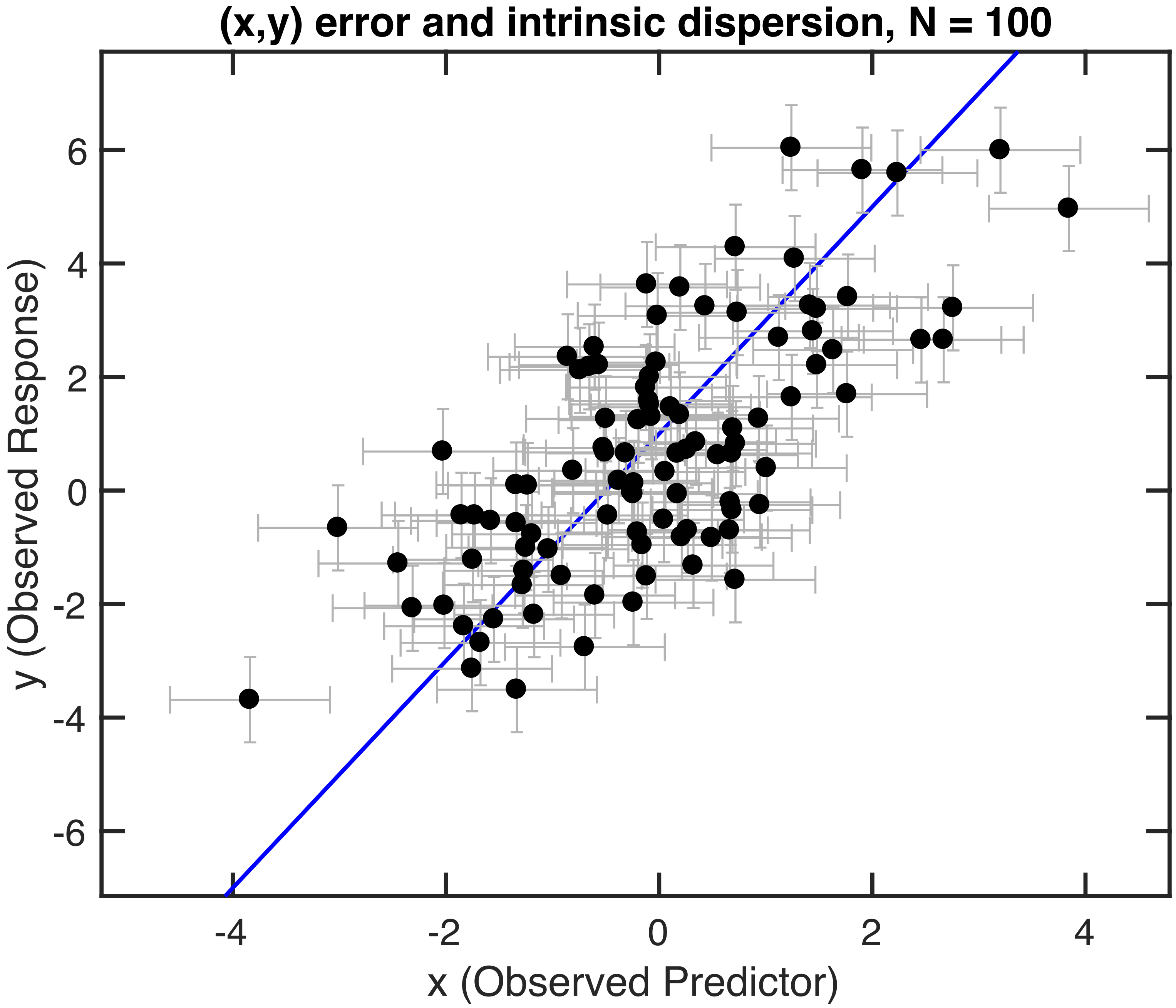
- Using observed data D, draw inference from Likelihood function: $L(\theta) = P(D \mid \theta)$

- Or if Bayesian with prior $P(\theta)$: sample posterior: $P(\theta \mid D) = P(D \mid \theta)\, P(\theta)$

Example: Structural Model for Linear Regression (B. Kelly et al. 2007, “Some Aspects of Measurement Error in Linear Regression of Astronomical Data.” ApJ, 665, 1489)

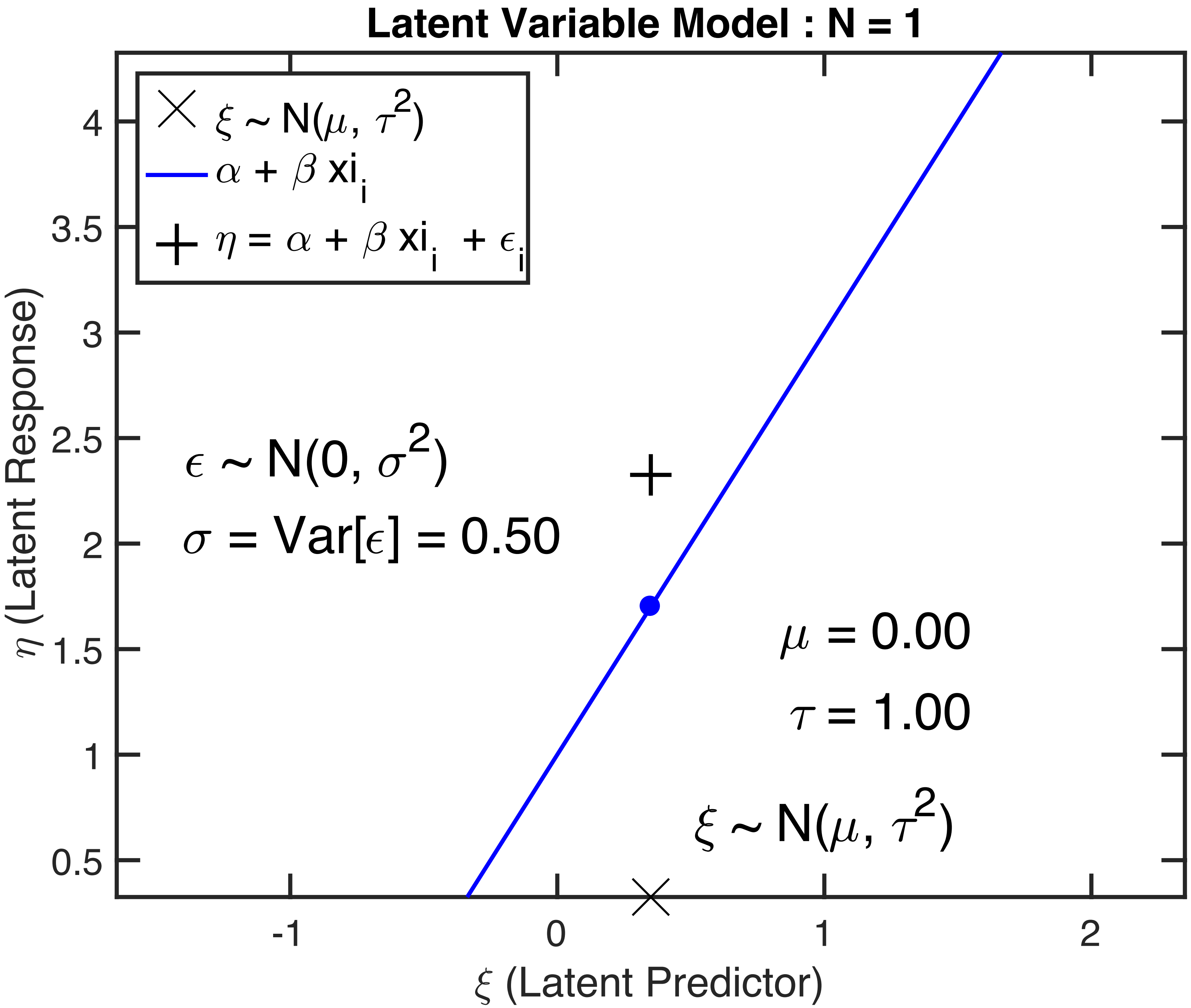
- Observed data has x and y meas. errors and intrinsic dispersion

- Estimate the true slope (and other parameters)



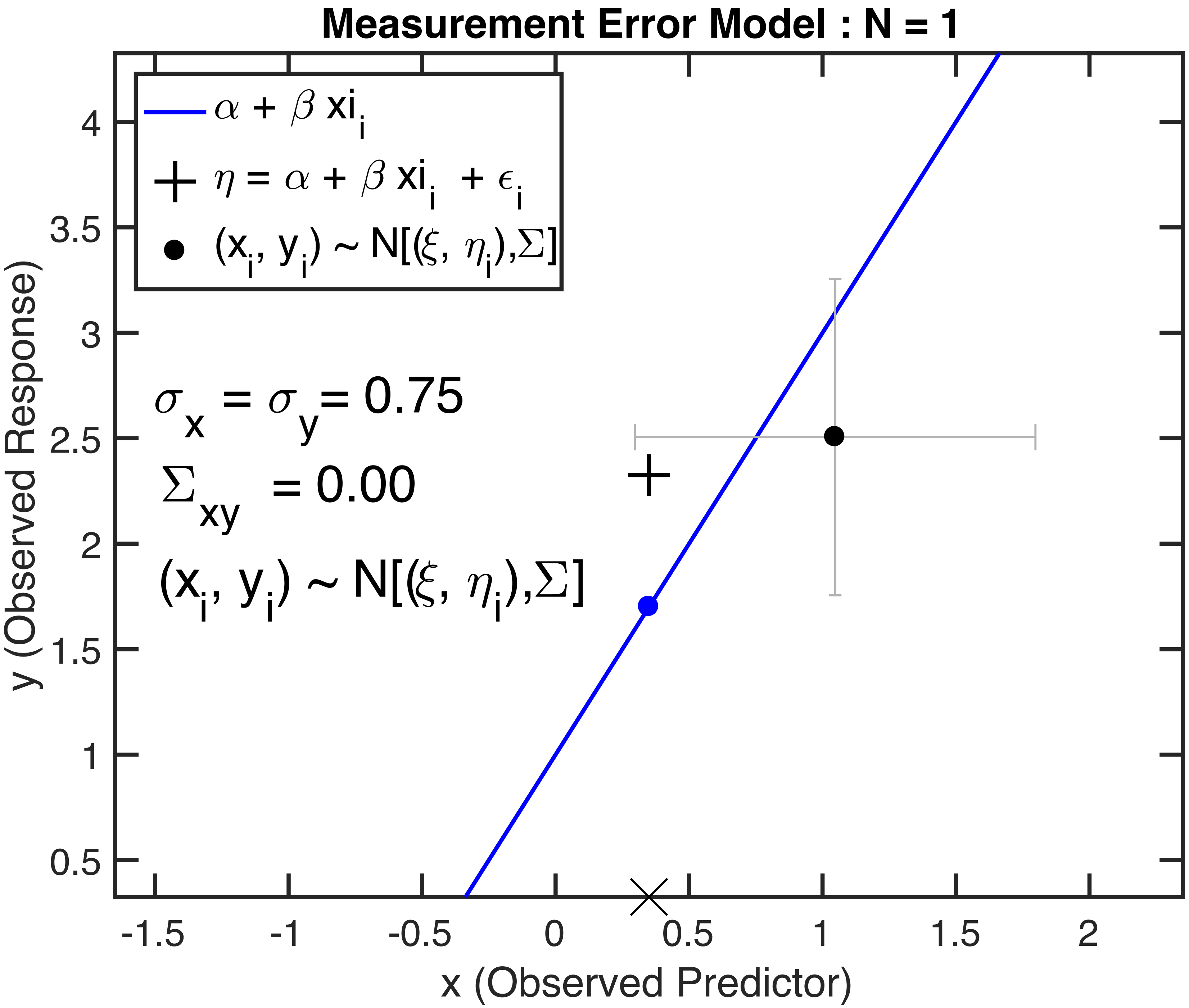
Step 1:
Generating Latent Variables from Parameters:

$$P(\eta_i, \xi_i | \alpha, \beta, \sigma, \mu, \tau) = P(\eta_i | \xi_i, \alpha, \beta, \sigma) \times P(\xi_i | \mu, \tau)$$

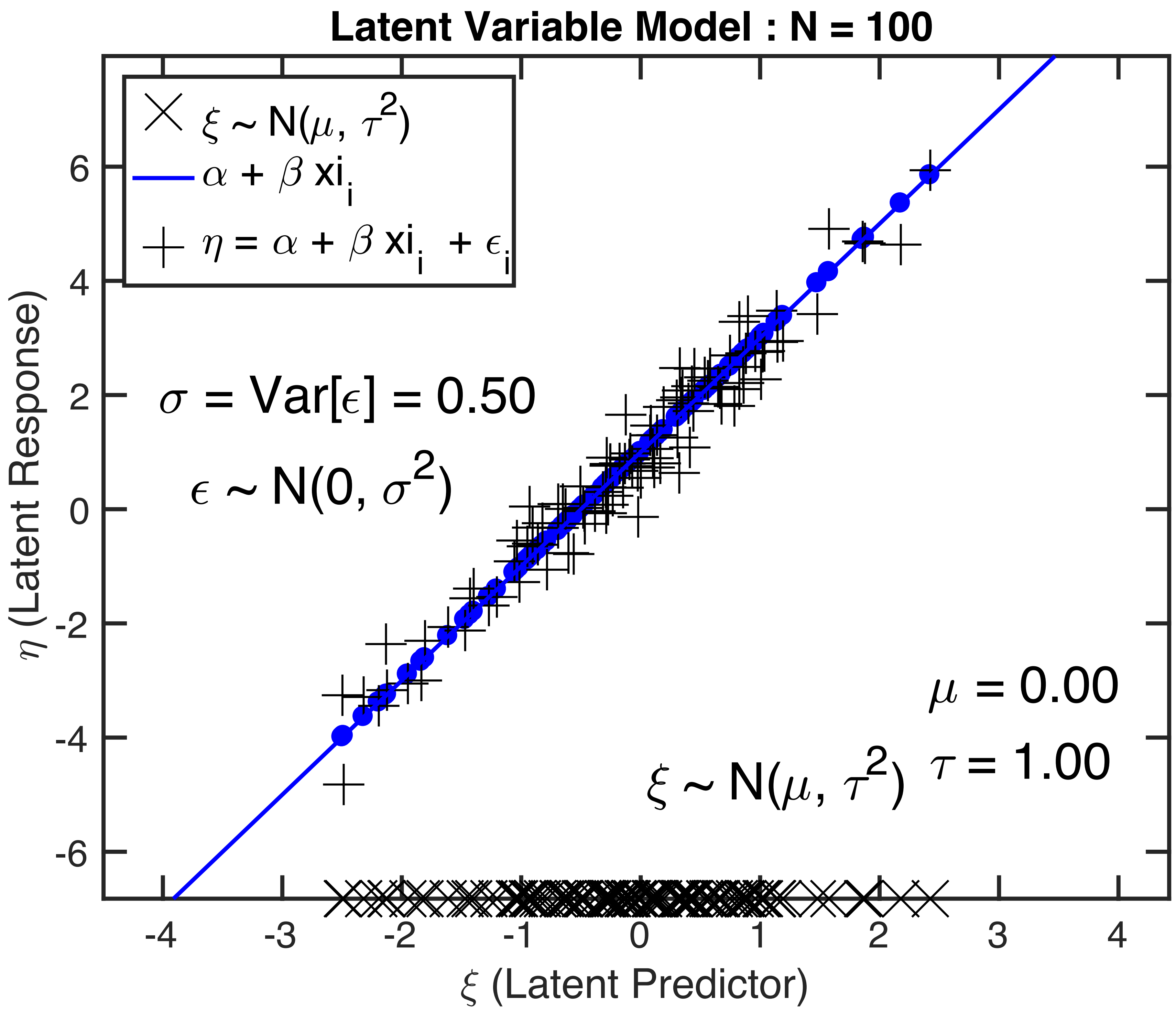


Step 2: Generating Observed Data from Latent Variables

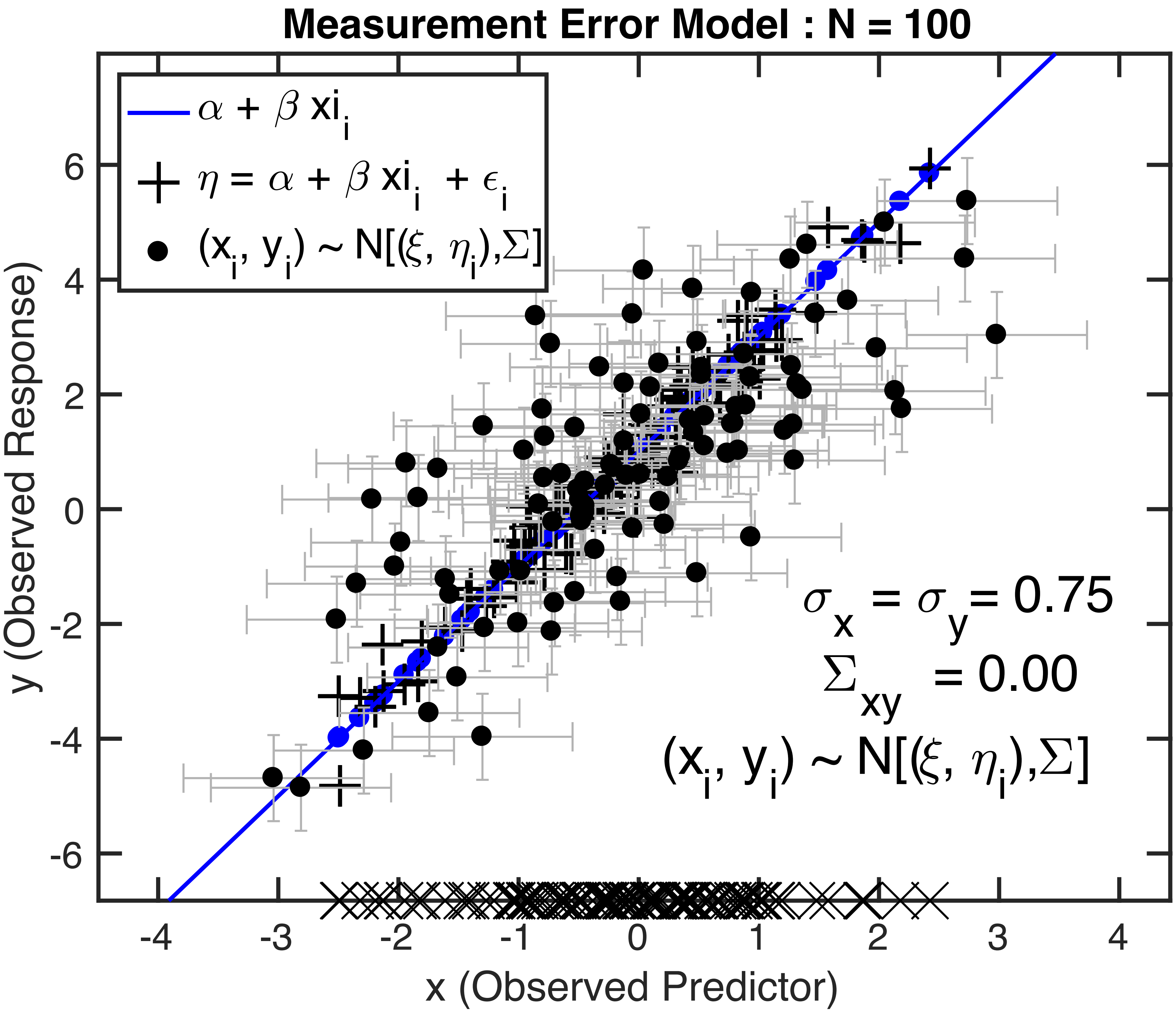
$$P([x_i, y_i]|\eta_i, \xi_i) = N([x_i, y_i]|[\eta_i, \xi_i], \mathbf{\Sigma})$$



Now repeat for N=100 objects



Now repeat for N=100 objects



Knowns and Unknowns

Regression Parameters	$\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$
Independent Variable Population Distribution “Hyperparameters”	$\boldsymbol{\psi} = (\mu, \tau)$
Latent (true) Variables	(ξ_i, η_i)
Observed Data	(x_i, y_i)

Generative Model

Population Distribution

$$\xi \sim N(\mu|\tau^2)$$

Regression

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

Measurement Error

$$[x_i, y_i]|\xi_i, \eta_i \sim N([\xi_i, \eta_i], \Sigma])$$

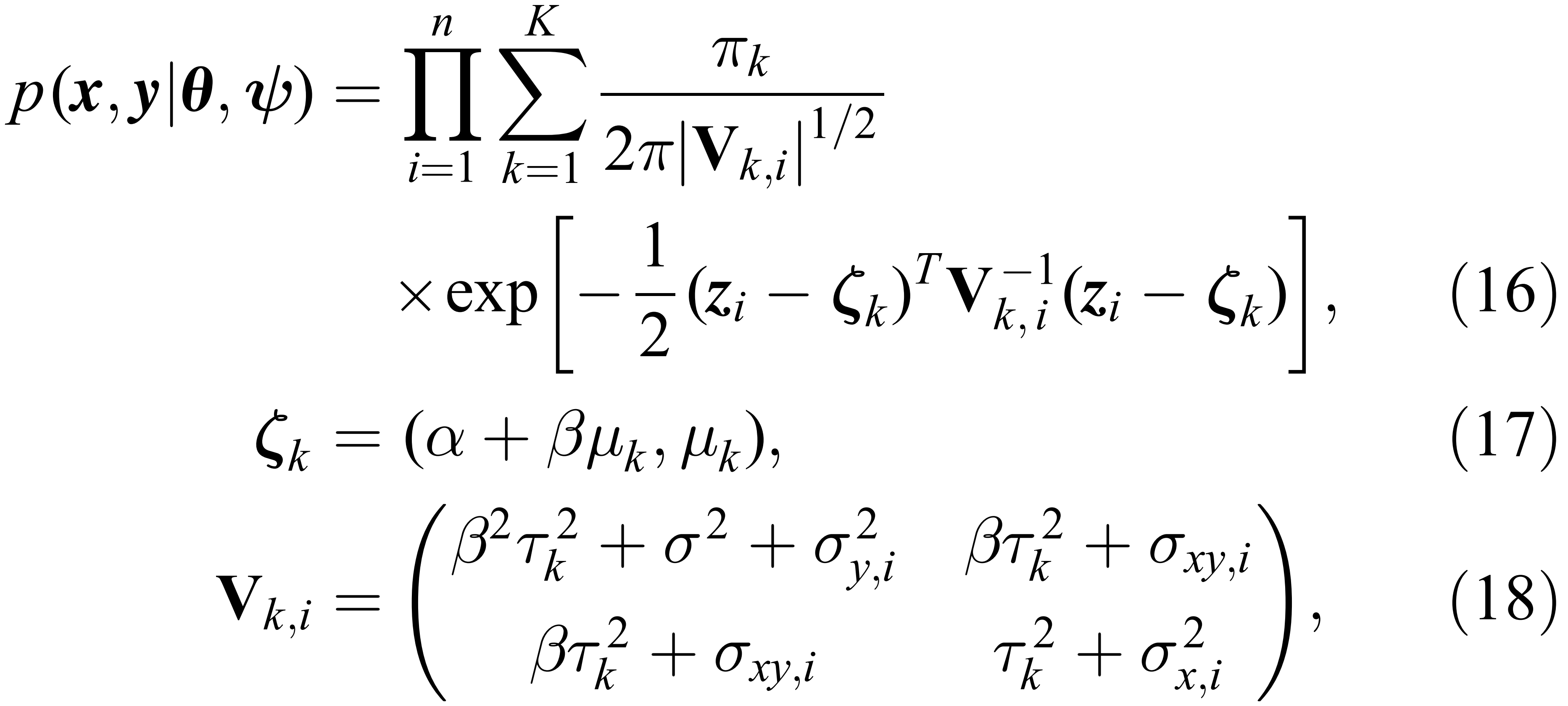
Formulating Likelihood Function: Marginalising (integrating out) latent variables

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

$$P(x_i, y_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int \int \underbrace{P(x_i, y_i|\xi_i, \eta_i)}_{\text{Measurement Error}} \underbrace{P(\eta_i|\xi_i, \boldsymbol{\theta})}_{\text{Regression}} \underbrace{P(\xi_i|\boldsymbol{\psi})}_{\text{Population Distribution}} \, d\xi_i \, d\eta$$

Solution: (Kelly 2007, Eqs. 16-23)

Gave More General Solution when P(ξ|Ψ)
is a Mixture of Gaussians
(set K=1, π_1 = 1 for us)

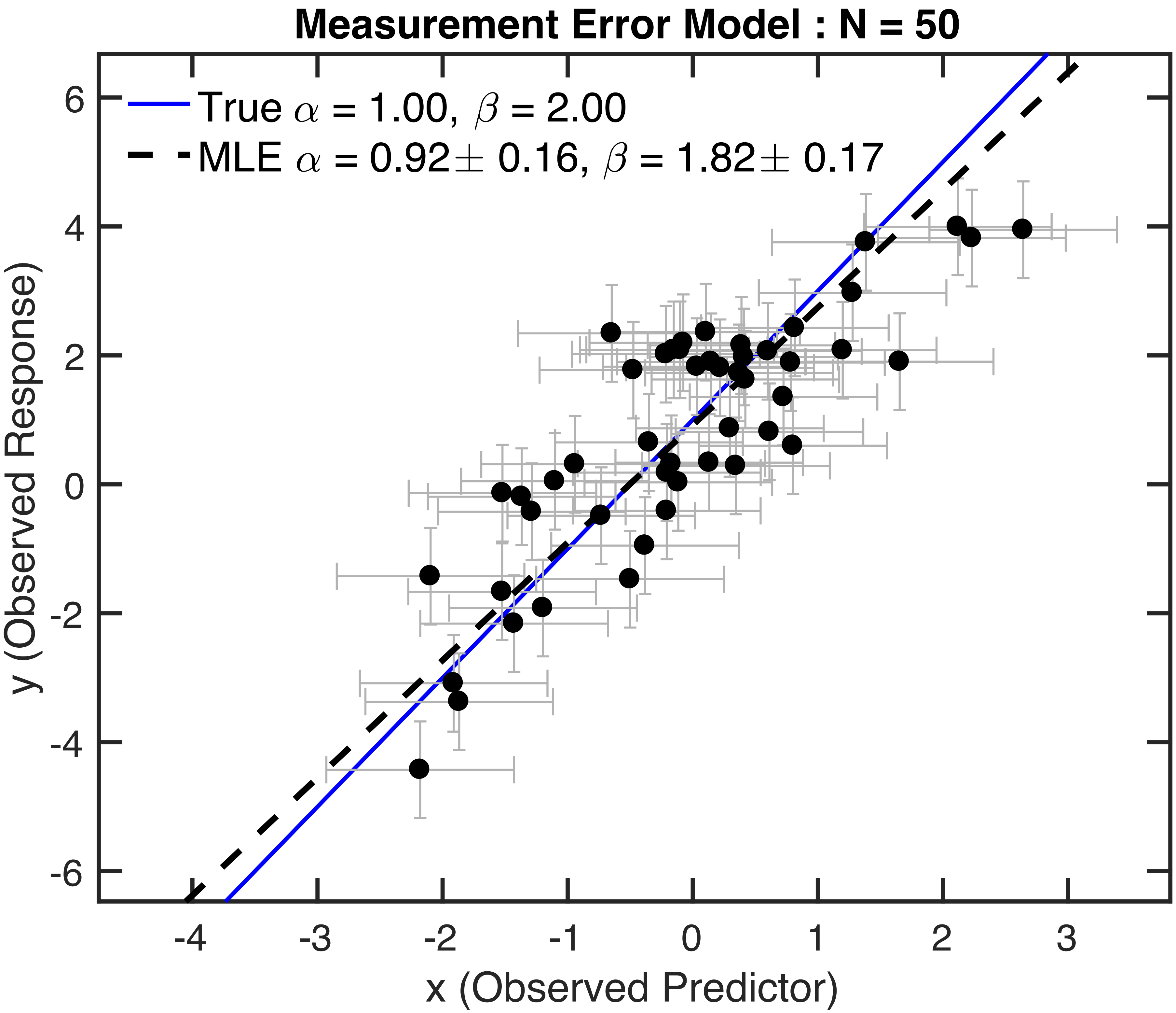


$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \psi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_k}{2\pi|\mathbf{V}_{k,i}|^{1/2}} \times \exp\left[-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\zeta}_k)^T \mathbf{V}_{k,i}^{-1}(\boldsymbol{z}_i - \boldsymbol{\zeta}_k)\right], \quad (16)$$

$$\boldsymbol{\zeta}_k = (\alpha + \beta\mu_k, \mu_k), \quad (17)$$

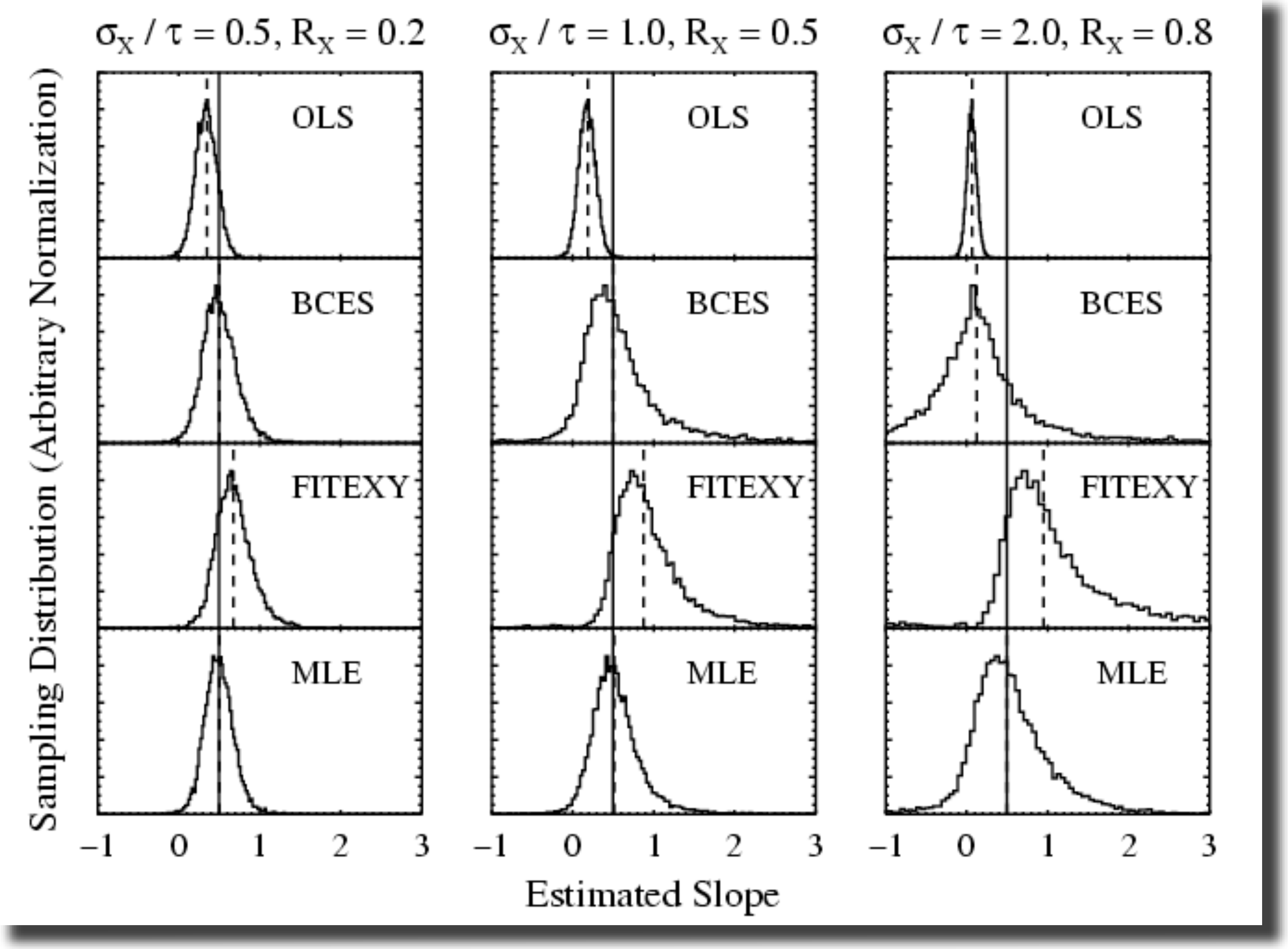
$$\mathbf{V}_{k,i} = \begin{pmatrix} \beta^2\tau_k^2 + \sigma^2 + \sigma_{y,i}^2 & \beta\tau_k^2 + \sigma_{xy,i} \\ \beta\tau_k^2 + \sigma_{xy,i} & \tau_k^2 + \sigma_{x,i}^2 \end{pmatrix}, \quad (18)$$

Example


Measurement Error Model : N = 50
True α = 1.00, β = 2.00
MLE α = 0.92± 0.16, β = 1.82± 0.17
6
4
2
0
-2
-4
-6
-4
-3
-2
-1
0
1
2
3
y (Observed Response)
x (Observed Predictor)

Kelly et al. 2017, Latent Variable Likelihood approach vs. Bad

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