The Astrophysical Journal, 665:1489–1506, 2007 August 20

SOME ASPECTS OF MEASUREMENT ERROR IN LINEAR REGRESSION OF ASTRONOMICAL DATA

Brandon C. Kelly
Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721; bkelly@as.arizona.edu
Received 2006 December 7; accepted 2007 May 8

ABSTRACT

I describe a Bayesian method to account for measurement errors in linear regression of astronomical data. The method allows for heteroscedastic and possibly correlated measurement errors and intrinsic scatter in the regression relationship. The method is based on deriving a likelihood function for the measured data, and I focus on the case when the intrinsic distribution of the independent variables can be approximated using a mixture of Gaussian functions. I generalize the method to incorporate multiple independent variables, nondetections, and selection effects (e.g., Malmquist bias). A Gibbs sampler is described for simulating random draws from the probability distribution of the parameters, given the observed data. I use simulation to compare the method with other common estimators. The simulations illustrate that the Gaussian mixture model outperforms other common estimators and can effectively give constraints on the regression parameters, even when the measurement errors dominate the observed scatter, source detection fraction is low, or the intrinsic distribution of the independent variables is not a mixture of Gaussian functions. I conclude by using this method to fit the X-ray spectral slope as a function of Eddington ratio using a sample of 39 $z \lesssim 0.8$ radio-quiet quasars. I confirm the correlation seen by other authors between the radio-quiet quasar X-ray spectral slope and the Eddington ratio, where the X-ray spectral slope softens as the Eddington ratio increases. IDL routines are made available for performing the regression.

Subject headings: methods: data analysis — methods: numerical — methods: statistical

1. INTRODUCTION

Linear regression is one of the most common statistical techniques used in astronomical data analysis. In general, linear regression in astronomy is characterized by intrinsic scatter about the regression line and measurement errors in both the independent and dependent variables. The source of intrinsic scatter is variations in the physical properties of astronomical sources that are not completely captured by the variables included in the regression. It is important to correctly account for both measurement error and intrinsic scatter, as both aspects can have a nonnegligible effect on the regression results. In particular, ignoring the intrinsic scatter and weighting the data points solely by the measurement errors can result in the higher precision measurements being given disproportionate influence on the regression results. Furthermore, when the independent variable is measured with error, the ordinary least-squares (OLS) estimate of the regression slope is biased toward zero (e.g., Fuller 1987; Akritas & Bershady 1996; Fox 1997). When there are multiple independent variables, measurement error can have an even stronger and more unpredictable effect (Fox 1997). In addition, the existence of nondetections, referred to as "censored data," in the data set will result in additional complications (e.g., Isobe et al. 1986). Therefore, when performing regression, it is essential to correctly account for the measurement errors and intrinsic scatter in order to ensure that the data analysis and, thus, the scientific conclusions based on it are trustworthy.

Many methods have been proposed for performing linear regression when intrinsic scatter is present and both variables are measured with error. These include methods that correct the observed moments of the data (e.g., Fuller 1987; Akritas & Bershady 1996; Freedman et al. 2004), minimize an "effective" χ^2 statistic (e.g., Clutton-Brock 1967; Barker & Diana 1974; Press et al. 1992; Tremaine et al. 2002), and assume a probability distribution for the true independent variable values (so-called structural equation models; e.g., Schafer 1987, 2001; Roy & Banerjee 2006); Bayesian approaches to these models have also been developed (e.g., Zellner 1971; Gull 1989; Dellaportas & Stephens 1995; Carroll et al. 1999; Scheines et al. 1999). In addition, methods have been proposed to account for measurement error in censored regression (e.g., Stapleton & Young 1984; Weiss 1993). The most commonly used methods in astronomy are the BCES estimator (Akritas & Bershady 1996) and the "FITEXY" estimator (Press et al. 1992). Both methods have their advantages and disadvantages, some of which have been pointed out by Tremaine et al. (2002). However, neither method is applicable when the data contain nondetections.

In this work I describe a Bayesian method for handling measurement errors in astronomical data analysis. My approach starts by computing the likelihood function of the complete data, i.e., the likelihood function of both the unobserved true values of the data and the measured values of the data. The measured data likelihood is then found by integrating the likelihood function for the complete data over the unobserved true values (e.g., Little & Rubin 2002; Gelman et al. 2004). This approach is known as "structural equation modeling" of measurement error problems and has been studied from both a frequentist approach (e.g., Fuller 1987; Carroll et al. 1995; Schafer 2001; Aitken & Rocci 2002) and a Bayesian approach (e.g., Müller & Roeder 1997; Richardson & Leblond 1997; Richardson et al. 2002). In this work I extend the statistical model of Carroll et al. (1999) to allow for measurement errors of different magnitudes (i.e., "heteroscedastic" errors), nondetections, and selection effects, so long as the selection function can be modeled mathematically. Our method models the distribution of independent variables as a weighted sum of Gaussian functions. The mixture of Gaussians model allows flexibility when estimating the distribution of the true values of the independent variable, thus increasing its robustness against model misspecification (e.g., Huang et al. 2006). The basic idea is that one can use a suitably large enough number of Gaussian functions to accurately approximate the true distribution of independent variables, even though in general the individual Gaussian functions have no physical meaning.

The paper is organized as follows. In § 2 we summarize some notation, and in § 3 I review the effects of measurement error on

the estimates for the regression slope and correlation coefficient. In § 4 I describe the statistical model and derive the likelihood functions, and in § 5 I describe how to incorporate knowledge of the selection effects and account for nondetections. In § 6.1 I describe the prior distribution for this model, and in § 6.2 I describe a Gibbs sampler for sampling from the posterior distributions. In § 7 I use simulation to illustrate the effectiveness of this structural model and compare with the OLS, BCES($Y|X$), and FITEXY estimators. Finally, in § 8 I illustrate the method using astronomical data by performing a regression of the X-ray photon index Γ_X on the Eddington ratio using a sample of 39 $z < 0.83$ radio-quiet quasars. Sections 4, 5, and 6 are somewhat technical, and the reader who is uninterested in the mathematical and computational details may skip them.

2. NOTATION

I will use the common statistical notation that an estimate of a quantity is denoted by placing a "hat" above it; e.g., $\hat{\theta}$ is an estimate of the true value of the parameter θ. In general, greek letters will denote the true value of a quantity, while roman letters will denote the contaminated measured value. I will frequently refer to the "bias" of an estimator. The bias of an estimator is $E(\hat{\theta}) - \theta_0$, where $E(\hat{\theta})$ is the expectation value of the estimator $\hat{\theta}$ and θ_0 is the true value of θ. An unbiased estimator is one such that $E(\hat{\theta}) = \theta_0$.

I will denote a normal density with mean μ and variance σ^2 as $N(\mu, \sigma^2)$, and I will denote as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a multivariate normal density with p-element mean vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. If I want to explicitly identify the argument of the Gaussian function, I will use the notation $N(x|\mu, \sigma^2)$, which should be understood to be a Gaussian function with mean μ and variance σ^2 as a function of x. Following Gelman et al. (2004), I denote the scaled inverse χ^2 density as Inv $\chi^2(\nu, s^2)$, where ν is the degrees of freedom and s^2 is the scale parameter, and we denote the inverse Wishart as Inv Wishart$_\nu$(**S**), where ν is the degrees of freedom and **S** is the scale matrix. The inverse Wishart distribution can be thought of as a multivariate generalization of the scaled inverse χ^2 distribution. I will often use the common statistical notation where a tilde means "is drawn from" or "is distributed as." For example, $x \sim N(\mu, \sigma^2)$ states that x is drawn from a normal density with mean μ and variance σ^2.

3. EFFECT OF MEASUREMENT ERROR ON CORRELATION AND REGRESSION

It is well known that measurement error can attenuate the estimate of the regression slope and correlation coefficient (e.g., Fuller 1987; Fox 1997). For completeness, I give a brief review of the effect of measurement error on correlation and regression analysis for the case of one independent variable.

Denote the independent variable as $\boldsymbol{\xi}$ and the dependent variable as $\boldsymbol{\eta}$; $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are also referred to as the "covariate" and the "response," respectively. I assume that $\boldsymbol{\xi}$ is a random vector of n data points drawn from some probability distribution. The dependent variable $\boldsymbol{\eta}$ depends on $\boldsymbol{\xi}$ according to the usual additive model

$$\eta_i = \alpha + \beta\xi_i + \epsilon_i, \tag{1}$$

where ϵ_i is a random variable representing the intrinsic scatter in η_i about the regression relationship and (α, β) are the regression coefficients. The mean of $\boldsymbol{\epsilon}$ is assumed to be zero, and the variance of $\boldsymbol{\epsilon}$ is assumed to be constant and is denoted as σ^2. We do not observe the actual values of $(\boldsymbol{\xi}, \boldsymbol{\eta})$, but instead observe values $(\boldsymbol{x}, \boldsymbol{y})$ which are measured with error. The measured values are assumed to be related to the actual values as

$$x_i = \xi_i + \epsilon_{x,i}, \tag{2}$$

$$y_i = \eta_i + \epsilon_{y,i}, \tag{3}$$

where $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are the random measurement errors on x_i and y_i, respectively. In general, the errors are normally distributed with known variances $\sigma^2_{x,i}$ and $\sigma^2_{y,i}$ and covariance $\sigma_{xy,i}$. For simplicity, throughout the rest of this section I assume that σ_x^2, σ_y^2, and σ_{xy} are the same for each data point.

When the data are measured without error, the least-squares estimate of the regression slope, $\hat{\beta}_{\rm OLS}$, and the estimated correlation coefficient, $\hat{\rho}$, are

$$\hat{\beta}_{\rm OLS} = \frac{{\rm Cov}(\boldsymbol{\xi}, \boldsymbol{\eta})}{{\rm Var}(\boldsymbol{\xi})}, \tag{4}$$

$$\hat{\rho} = \frac{{\rm Cov}(\boldsymbol{\xi}, \boldsymbol{\eta})}{\sqrt{{\rm Var}(\boldsymbol{\xi}){\rm Var}(\boldsymbol{\eta})}} = \hat{\beta}_{\rm OLS}\sqrt{\frac{{\rm Var}(\boldsymbol{\xi})}{{\rm Var}(\boldsymbol{\eta})}}, \tag{5}$$

where ${\rm Cov}(\boldsymbol{\xi}, \boldsymbol{\eta})$ is the sample covariance between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and ${\rm Var}(\boldsymbol{\xi})$ is the sample variance of $\boldsymbol{\xi}$. When the data are measured with error, the least-squares estimate of the regression slope, $\hat{b}_{\rm OLS}$, and the estimated correlation coefficient, $\hat{r}$, become

$$\hat{b}_{\rm OLS} = \frac{{\rm Cov}(\boldsymbol{x}, \boldsymbol{y})}{{\rm Var}(\boldsymbol{x})} = \frac{{\rm Cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) + \sigma_{xy}}{{\rm Var}(\boldsymbol{\xi}) + \sigma_x^2}, \tag{6}$$

$$\hat{r} = \frac{{\rm Cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{{\rm Var}(\boldsymbol{x}){\rm Var}(\boldsymbol{y})}} = \frac{{\rm Cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) + \sigma_{xy}}{\sqrt{\left[{\rm Var}(\boldsymbol{\xi}) + \sigma_x^2\right]\left[{\rm Var}(\boldsymbol{\eta}) + \sigma_y^2\right]}}. \tag{7}$$

From these equations it is apparent that the estimated slope and correlation are biased when the data are measured with error.

It is informative to assess the effect of measurement error in terms of the ratios $R_x = \sigma_x^2/{\rm Var}(\boldsymbol{x})$, $R_y = \sigma_y^2/{\rm Var}(\boldsymbol{y})$, and $R_{xy} = \sigma_{xy}/{\rm Cov}(\boldsymbol{x}, \boldsymbol{y})$, as these quantities can be calculated from the data. The fractional bias in the estimated slope and correlation may then be expressed as

$$\frac{\hat{b}}{\hat{\beta}} = \frac{1 - R_x}{1 - R_{xy}}, \tag{8}$$

$$\frac{\hat{r}}{\hat{\rho}} = \frac{\sqrt{(1 - R_x)(1 - R_y)}}{1 - R_{xy}}. \tag{9}$$

From equations (8) and (9) it is apparent that measurement errors have the following effects. First, covariate measurement error reduces the magnitude of the observed correlation between the independent variable and the response, as well as biasing the estimate of the slope toward zero. Second, measurement error in the response also reduces the magnitude of the observed correlation between the variables. Third, if the measurement errors are correlated, the effects depend on the sign of this correlation. If the measurement error correlation has the same sign as the intrinsic correlation between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, then the measurement errors cause a spurious increase in the observed correlation; otherwise, the measurement errors cause a spurious decrease in the observed correlation. The magnitude of these effects depend on how large the measurement errors are compared to the observed variance in $\boldsymbol{x}$ and $\boldsymbol{y}$.

In Figure 1 I plot the fractional bias in the correlation coefficient, $(\hat{\rho} - \hat{r})/\hat{\rho}$, as a function of R_x and R_y when the errors are uncorrelated. As can be seen, measurement error can have a significant

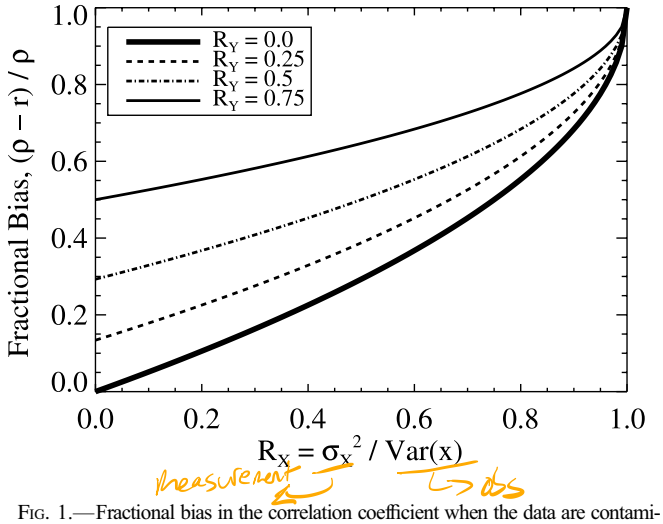

FIG. 1.—Fractional bias in the correlation coefficient when the data are contaminated with measurement error. The fractional bias is shown as a function of the contribution of measurement error to the observed variance in both x and y, for uncorrelated measurement errors. When the measurement errors make up ~50% of the observed variance in both x and y, the observed correlation coefficient is reduced by about ~50%.

effect on the estimation of the linear correlation coefficient. For example, when $R_x \approx 0.5$ and $R_y \approx 0.5$, the estimated correlation is ≈50% lower than the true correlation. Therefore, interpretation of correlation coefficients and regression slopes must be approached with caution when the data have been contaminated by measurement error. To ensure accurate results, it is necessary to employ statistical methods that correct for the measurement errors.

4. THE STATISTICAL MODEL

4.1. *Regression with One Independent Variable*

I assume that the independent variable ξ is drawn from a probability distribution $p(\xi|\psi)$, where ψ denotes the parameters for this distribution. The dependent variable is then drawn from the conditional distribution of η given ξ, denoted as $p(\eta|\xi, \boldsymbol{\theta})$; $\boldsymbol{\theta}$ denotes the parameters for this distribution. The joint distribution of ξ and η is then $p(\xi, \eta|\psi, \boldsymbol{\theta}) = p(\eta|\xi, \boldsymbol{\theta})p(\xi|\psi)$. In this work I assume the normal linear regression model given by equation (1), and thus, $p(\eta|\xi, \boldsymbol{\theta})$ is a normal density with mean $\alpha + \beta\xi$ and variance σ^2, and $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$.

Since the data are a randomly observed sample, we can derive the likelihood function for the measured data. The likelihood function of the measured data, $p(x, y|\boldsymbol{\theta}, \psi)$, is obtained by integrating the complete data likelihood over the missing data, ξ and η (e.g., Little & Rubin 2002; Gelman et al. 2004),

$$p(x, y|\boldsymbol{\theta}, \psi) = \int\int p(x, y, \xi, \eta|\boldsymbol{\theta}, \psi)d\xi\, d\eta, \tag{10}$$

where $p(x, y, \xi, \eta|\boldsymbol{\theta}, \psi)$ is the complete data likelihood function. Because of the hierarchical structure inherent in the measurement error model, it is helpful to decompose the complete data likelihood into conditional probability densities,

$$p(x, y|\boldsymbol{\theta}, \psi) = \int\int p(x, y|\xi, \eta)p(\eta|\xi, \boldsymbol{\theta})p(\xi|\psi)d\xi\, d\eta. \tag{11}$$

The density $p(x, y|\xi, \eta)$ describes the joint distribution of the measured values x and y at a given ξ and η and depends on the assumed distribution of the measurement errors, ϵ_x and ϵ_y. In this work I assume Gaussian measurement error, and thus, $p(x_i, y_i|\xi_i, \eta_i)$ is a multivariate normal density with mean (ξ_i, η_i) and covariance matrix $\boldsymbol{\Sigma}_i$, where $\Sigma_{11,i} = \sigma^2_{y,i}$, $\Sigma_{22,i} = \sigma^2_{x,i}$, and $\Sigma_{12,i} = \sigma_{xy,i}$. The statistical model may then be conveniently expressed hierarchically as

$$\xi_i \sim p(\boldsymbol{\xi}|\psi), \tag{12}$$

$$\eta_i|\xi_i \sim N(\alpha + \beta\xi_i, \sigma^2), \tag{13}$$

$$y_i, x_i|\eta_i, \xi_i \sim N_2([\eta_i, \xi_i], \boldsymbol{\Sigma}_i). \tag{14}$$

Note that if x_i is measured without error, then $p(x_i|\xi_i)$ is a Dirac δ-function, and $p(x_i, y_i|\xi_i, \eta_i) = p(y_i|\eta_i)\delta(x_i - \xi_i)$. An equivalent result holds if y_i is measured without error.

Equation (11) may be used to obtain the observed data likelihood function for any assumed distribution of $\boldsymbol{\xi}$. In this work, I model $p(\boldsymbol{\xi}|\psi)$ as a mixture of K Gaussian functions,

$$p(\xi_i|\psi) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi\tau_k^2}} \exp\left[-\frac{1}{2}\frac{(\xi_i - \mu_k)^2}{\tau_k^2}\right], \tag{15}$$

where $\sum_{k=1}^{K} \pi_k = 1$. Note that π_k may be interpreted as the probability of drawing a data point from the kth Gaussian function. I will use the convenient notation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_K^2)$; note that $\psi = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}^2)$. It is useful to model $p(\boldsymbol{\xi}|\psi)$ using this form, because it is flexible enough to adapt to a wide variety of distributions, but is also conjugate for the regression relationship (eq. [1]) and the measurement error distribution, thus simplifying the mathematics.

Assuming the Gaussian mixture model for $p(\boldsymbol{\xi}|\psi)$, the measured data likelihood for the ith data point can be directly calculated using equation (11). Denoting the measured data as $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x})$, the measured data likelihood function for the ith data point is then a mixture of bivariate normal distributions with weights $\boldsymbol{\pi}$, means $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K)$, and covariance matrices $\boldsymbol{V}_i = (\mathbf{V}_{1,i}, \ldots, \mathbf{V}_{K,i})$. Because the data points are statistically independent, the full measured data likelihood is then the product of the likelihood functions for the individual data points,

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

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

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

where $\boldsymbol{z}^T$ denotes the transpose of $\boldsymbol{z}$. Equation (16) may be maximized to compute the maximum likelihood estimate (MLE). When $K > 1$, the expectation-maximization (EM) algorithm (Dempster et al. 1977) is probably the most efficient tool for calculating the MLE. Roy & Banerjee (2006) describe an EM algorithm when $p(\boldsymbol{\xi})$ is assumed to be a mixture of normals and the measurement error distribution is multivariate t, and their results can be extended to the statistical model described in this work.

It is informative to decompose the measured data likelihood, $p(x_i, y_i|\boldsymbol{\theta}, \psi) = p(y_i|x_i, \boldsymbol{\theta}, \psi)p(x_i|\psi)$, as this representation is useful when the data contain nondetections (see § 5.2). The marginal distribution of x_i is

$$p(x_i|\psi) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi(\tau_k^2 + \sigma^2_{x,i})}} \exp\left[-\frac{1}{2}\frac{(x_i - \mu_k)^2}{\tau_k^2 + \sigma^2_{x,i}}\right], \tag{19}$$

and the conditional distribution of y_i given x_i is

$$p(y_i|x_i,\boldsymbol{\theta},\boldsymbol{\psi}) = \sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi \mathrm{Var}(y_i|x_i,k)}} \times \exp\left\{-\frac{1}{2}\frac{[y_i - E(y_i|x_i,k)]^2}{\mathrm{Var}(y_i|x_i,k)}\right\}, \quad (20)$$

$$\gamma_k = \frac{\pi_k N(x_i|\mu_k, \tau_k^2 + \sigma_{x,i}^2)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j, \tau_j^2 + \sigma_{x,i}^2)}, \quad (21)$$

$$E(y_i|x_i,k) = \alpha + \frac{\beta\tau_k^2 + \sigma_{xy,i}}{\tau_k^2 + \sigma_{x,i}^2} x_i + \frac{\beta\sigma_{x,i}^2 - \sigma_{xy,i}}{\tau_k^2 + \sigma_{x,i}^2}\mu_k, \quad (22)$$

$$\mathrm{Var}(y_i|x_i,k) = \beta^2\tau_k^2 + \sigma^2 + \sigma_{y,i}^2 - \frac{(\beta\tau_k^2 - \sigma_{xy,i})^2}{\tau_k^2 + \sigma_{x,i}^2}, \quad (23)$$

where γ_k can be interpreted as the probability that the ith data point was drawn from the kth Gaussian function given x_i, $E(y_i|x_i,k)$ gives the expectation value of y_i at x_i, given that the data point was drawn from the kth Gaussian function, and $\mathrm{Var}(y_i|x_i,k)$ gives the variance in y_i at x_i, given that the data point was drawn from the kth Gaussian function.

4.2. *Relationship between Uniformly Distributed Covariates and Effective χ^2 Estimators*

It is informative to investigate the case where the distribution of $\boldsymbol{\xi}$ is assumed to be uniform, $p(\boldsymbol{\xi}) \propto 1$. Interpreting $p(\boldsymbol{\xi})$ as a "prior" on $\boldsymbol{\xi}$, one may be tempted to consider assuming $p(\boldsymbol{\xi}) \propto 1$ as a more objective alternative to the normal distribution. A uniform distribution for $\boldsymbol{\xi}$ may be obtained as the limit $\boldsymbol{\tau}^2 \to \infty$, and thus, the likelihood function for $p(\boldsymbol{\xi}) \propto 1$ can be calculated from equation (20) by taking $\boldsymbol{\tau}^2 \to \infty$ and $K = 1$. When the measurement errors are uncorrelated, the likelihood for uniform $p(\boldsymbol{\xi})$ is

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2)}} \times \exp\left[-\frac{1}{2}\frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2 + \sigma_{y,i}^2 + \beta^2\sigma_{x,i}^2}\right]. \quad (24)$$

The argument of the exponential is the FITEXY goodness of fit statistic, χ^2_{EXY}, as modified by Tremaine et al. (2002) to account for intrinsic scatter; this fact has also been recognized by Weiner et al. (2006). Despite this connection, minimizing χ^2_{EXY} is not the same as maximizing the conditional likelihood of $\boldsymbol{y}$ given $\boldsymbol{x}$, as both β and σ^2 appear in the normalization of the likelihood function as well.

For a given value of σ^2, minimizing χ^2_{EXY} can be interpreted as minimizing a weighted sum of squared errors, where the weights are given by the variances in y_i at a given x_i and one assumes a uniform distribution for $\boldsymbol{\xi}$. Unfortunately, this is only valid for a fixed value of σ^2. Moreover, little is known about the statistical properties of the FITEXY estimator, such as its bias and variance, although bootstrapping (e.g., Efron 1979; Davison & Hinkley 1997) may be used to estimate them. Furthermore, it is ambiguous how to calculate the FITEXY estimates when there is an intrinsic scatter term. The FITEXY goodness of fit statistic, χ^2_{EXY}, cannot be simultaneously minimized with respect to α, β, and σ^2, as χ^2_{EXY} is a strictly decreasing function of σ^2. As such, it is unclear how to proceed in the optimization beyond an ad hoc approach. Many authors have followed the approach adopted by Tremaine et al. (2002) and increase σ^2 until $\chi^2_{\mathrm{EXY}}/(n-2) = 1$ or assume $\sigma^2 = 0$ if $\chi^2_{\mathrm{EXY}}/(n-2) < 1$.

Despite the fact that minimizing χ^2_{EXY} is not the same as maximizing equation (24), one may still be tempted to calculate a MLE based on equation (24). However, it can be shown that if one assumes $p(\boldsymbol{\xi}) \propto 1$ and if all of the $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same respective measurement error variances, σ_x^2 and σ_y^2, the MLE estimates for α and β are just the OLS estimates (Zellner 1971). While this is not necessarily true when the magnitudes of the measurement errors vary between data points, one might expect that the MLE will behave similarly to the OLS estimate. I confirm this fact using simulation in § 7.1. Unfortunately, this implies that the MLE for $p(\boldsymbol{\xi}) \propto 1$ inherits the bias in the OLS estimate, and thus, nothing is gained. Furthermore, as argued by Gull (1989) one can easily be convinced that assuming $p(\boldsymbol{\xi}) \propto 1$ is incorrect by examining a histogram of $\boldsymbol{x}$.

4.3. *Regression with Multiple Independent Variables*

The formalism developed in § 4.1 can easily be generalized to multiple independent variables. In this case, equation (1) becomes

$$\eta_i = \alpha + \boldsymbol{\beta}^T\boldsymbol{\xi}_i + \epsilon_i, \quad (25)$$

where $\boldsymbol{\beta}$ is now a p-element vector and $\boldsymbol{\xi}_i$ is a p-element vector containing the values of the independent variables for the ith data point. Similar to before, we assume that the distribution of $\boldsymbol{\xi}_i$ can be approximated using a mixture of K multivariate normal densities with p-element mean vectors $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, $p \times p$ covariance matrices $\boldsymbol{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_K)$, and weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. The measured value of $\boldsymbol{\xi}_i$ is the p-element vector $\boldsymbol{x}_i$, and the Gaussian measurement errors on $(y_i, \boldsymbol{x}_i)$ have $(p+1)\times(p+1)$ covariance matrix $\boldsymbol{\Sigma}_i$. The statistical model is then

$$\boldsymbol{\xi}_i \sim \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{\mu}_k, \mathbf{T}_k), \quad (26)$$

$$\eta_i|\boldsymbol{\xi}_i \sim N(\alpha + \boldsymbol{\beta}^T\boldsymbol{\xi}_i, \sigma^2), \quad (27)$$

$$y_i, \boldsymbol{x}_i|\eta_i, \boldsymbol{\xi}_i \sim N_{p+1}([\eta_i, \boldsymbol{\xi}_i], \boldsymbol{\Sigma}_i). \quad (28)$$

Denoting $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)$, the measured data likelihood is

$$p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\psi}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{(p+1)/2}|\mathbf{V}_{k,i}|^{1/2}} \times \exp\left[-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\zeta}_k)^T\mathbf{V}_{k,i}^{-1}(\boldsymbol{z}_i - \boldsymbol{\zeta}_k)\right], \quad (29)$$

$$\boldsymbol{\zeta}_k = (\alpha + \boldsymbol{\beta}^T\boldsymbol{\mu}_k, \boldsymbol{\mu}_k), \quad (30)$$

$$\mathbf{V}_{k,i} = \begin{pmatrix} \boldsymbol{\beta}^T\mathbf{T}_k\boldsymbol{\beta} + \sigma^2 + \sigma_{y,i}^2 & \boldsymbol{\beta}^T\mathbf{T}_k + \boldsymbol{\sigma}_{xy,i}^T \\ \mathbf{T}_k\boldsymbol{\beta} + \boldsymbol{\sigma}_{xy,i} & \mathbf{T}_k + \boldsymbol{\Sigma}_{x,i} \end{pmatrix}, \quad (31)$$

where $\boldsymbol{\zeta}_k$ is the $(p+1)$-element mean vector of $\boldsymbol{z}_i$ for Gaussian function k, $\mathbf{V}_{k,i}$ is the $(p+1)\times(p+1)$ covariance matrix of $\boldsymbol{z}_i$ for Gaussian function k, $\sigma_{y,i}^2$ is the variance in the measurement error on y_i, $\boldsymbol{\sigma}_{xy,i}$ is the p-element vector of covariances between the measurement errors on y_i and $\boldsymbol{x}_i$, and $\boldsymbol{\Sigma}_{x,i}$ is the $p \times p$ covariance matrix of the measurement errors on $\boldsymbol{x}_i$.

Similar to the case for one independent variable, the measured data likelihood can be decomposed as $p(x, y|\boldsymbol{\theta}, \psi) = p(y|x, \boldsymbol{\theta}, \psi) \times p(x|\psi)$, where $p(\boldsymbol{x}_i|\psi) = \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \mathbf{T}_k + \boldsymbol{\Sigma}_{x,i})$

$$p(\boldsymbol{x}_i|\psi) = \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \mathbf{T}_k + \boldsymbol{\Sigma}_{x,i}), \tag{32}$$

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}, \psi) = \sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi \mathrm{Var}(y_i|\boldsymbol{x}_i, k)}} \times \exp\left\{ -\frac{1}{2} \frac{[y_i - E(y_i|\boldsymbol{x}_i, k)]^2}{\mathrm{Var}(y_i|\boldsymbol{x}_i, k)} \right\}, \tag{33}$$

$$\gamma_k = \frac{\pi_k N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \mathbf{T}_k + \boldsymbol{\Sigma}_{x,i})}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \mathbf{T}_j + \boldsymbol{\Sigma}_{x,i})}, \tag{34}$$

$$E(y_i|\boldsymbol{x}_i, k) = \alpha + \boldsymbol{\beta}^T \boldsymbol{\mu}_k + (\boldsymbol{\beta}^T \mathbf{T}_k + \boldsymbol{\sigma}_{xy,i}^T) \times (\mathbf{T}_k + \boldsymbol{\Sigma}_{x,i})^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k), \tag{35}$$

$$\mathrm{Var}(y_i|\boldsymbol{x}_i, k) = \boldsymbol{\beta}^T \mathbf{T}_k \boldsymbol{\beta} + \sigma^2 + \sigma_{y,i}^2 - (\boldsymbol{\beta}^T \mathbf{T}_k + \boldsymbol{\sigma}_{xy,i}^T) \times (\mathbf{T}_k + \boldsymbol{\Sigma}_{x,i})^{-1} (\mathbf{T}_k \boldsymbol{\beta} + \boldsymbol{\sigma}_{xy,i}). \tag{36}$$

5. DATA COLLECTION ISSUES: SELECTION EFFECTS AND NONDETECTIONS

There are several issues common in the collection of astronomical data that violate the simple assumptions made in § 4. Astronomical data collection consists almost entirely of passive observations, and thus, selection effects are a common concern. Instrumental detection limits often result in the placement of upper or lower limits on quantities, and astronomical surveys are frequently flux limited. In this section I modify the likelihood functions described in § 4 to include the effects of data collection.

General methods for dealing with missing data are described in Little & Rubin (2002) and Gelman et al. (2004), and I apply the methodology described in these references to the measurement error model developed here. Although in this work I focus on linear regression, many of these results can be applied to more general statistical models, such as estimating luminosity functions.

5.1. *Selection Effects*

Suppose that one collects a sample of n sources out of a possible N sources. One is interested in understanding how the observable properties of these sources are related, but is concerned about the effects of the selection procedure on the data analysis. For example, one may perform a survey that probes some area of the sky. There are N sources located within this solid angle, where N is unknown. Because of the survey's selection method, the sample only includes n sources. In this case, the astronomer is interested in how measurement error and the survey's selection method affect statistical inference.

I investigate selection effects within the framework of our statistical model by introducing an indicator variable, $\boldsymbol{I}$, which denotes whether a source is included in the sample. If the ith source is included in the sample, then $I_i = 1$, otherwise $I_i = 0$. In addition, I assume that the selection function only depends on the measured values, $\boldsymbol{x}$ and $\boldsymbol{y}$. Under this assumption, the selection function of the sample is the probability of including a source with a given $\boldsymbol{x}$ and $\boldsymbol{y}$, $p(\boldsymbol{I}|\boldsymbol{x}, \boldsymbol{y})$. This is commonly the case in astronomy, where sources are collected based on their measured properties. For example, one may select sources for a sample based on their measured properties as reported in the literature. In addition, if one performs a flux-limited survey, then a source will only be considered detected if its measured flux falls above some set flux limit. If a sample is from a survey with a simple flux limit, then $p(I_i = 1|y_i) = 1$ if the measured source flux y_i is above the flux limit or $p(I_i = 1|y_i) = 0$ if the measured source flux is below the flux limit. Since the selection function depends on the measured flux value and not the true flux value, sources with true flux values above the flux limit can be missed by the survey, and sources with true flux below the limit can be detected by the survey. This effect is well known in astronomy and is commonly referred to as Malmquist bias (e.g., Landy & Szalay 1992).

Including the variable $\boldsymbol{I}$, the complete data likelihood can be written as

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{I}|\boldsymbol{\theta}, \psi) = p(\boldsymbol{I}|\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\eta}) p(\boldsymbol{\eta}|\boldsymbol{\xi}, \boldsymbol{\theta}) p(\boldsymbol{\xi}|\psi). \tag{37}$$

Equation (37) is valid for any number of independent variables, and thus, x_i and ξ_i may be either scalars or vectors. Integrating equation (37) over the missing data, the observed data likelihood is

$$p(\boldsymbol{x}_{\rm obs}, \boldsymbol{y}_{\rm obs}|\boldsymbol{\theta}, \psi, N) \propto C_n^N \prod_{i \in \mathcal{A}_{\rm obs}} p(x_i, y_i|\boldsymbol{\theta}, \psi) \times \prod_{j \in \mathcal{A}_{\rm mis}} \int p(I_j = 0|x_j, y_j) p(x_j, y_j|\xi_j, \eta_j) \times p(\eta_j|\xi_j, \boldsymbol{\theta}) p(\xi_j|\psi) dx_j\, dy_j\, d\xi_j\, d\eta_j, \tag{38}$$

where C_n^N is the binomial coefficient, $\mathcal{A}_{\rm obs}$ denotes the set of n included sources, $\boldsymbol{x}_{\rm obs}$ and $\boldsymbol{y}_{\rm obs}$ denote the values of $\boldsymbol{x}$ and $\boldsymbol{y}$ for the included sources, and $\mathcal{A}_{\rm mis}$ denotes the set of $N - n$ missing sources. In addition, I have omitted terms that do not depend on $\boldsymbol{\theta}$, ψ, or N. Note that N is unknown and is thus also a parameter of the statistical model. The binomial coefficient is necessary because it gives the number of possible ways to select a sample of n sources from a set of N sources.

It is apparent from equation (38) that statistical inference on the regression parameters is unaffected if the selection function is independent of $\boldsymbol{y}$ and $\boldsymbol{x}$ (e.g., Little & Rubin 2002; Gelman et al. 2004). In this case the selection function may be ignored.

5.1.1. *Selection Based on Measured Independent Variables*

It is commonly the case that a sample is selected based only on the measured independent variables. For example, suppose one performs a survey in which all sources with measured optical flux greater than some threshold are included. Then, these optically selected sources are used to fit a regression in order to understand how the X-ray luminosity of these objects depends on their optical luminosity and redshift. In this case, the probability of including a source only depends on the measured values of the optical luminosity and redshift and is thus independent of the X-ray luminosity.

When the sample selection function is independent of $\boldsymbol{y}$, given $\boldsymbol{x}$, then $p(\boldsymbol{I}|\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{I}|\boldsymbol{x})$. Because we are primarily interested in the regression parameters $\boldsymbol{\theta}$, I model the distributions of $\boldsymbol{\xi}$ for the included and missing sources separately, with the parameters for the distribution of included sources denoted as $\psi_{\rm obs}$. In addition, I assume that the measurement errors between $\boldsymbol{y}$ and $\boldsymbol{x}$ are statistically independent. Then the $N - n$ integrals over $\boldsymbol{y}$ and $\boldsymbol{\eta}$ for

the missing sources in equation (38) are equal to unity, and we can write the observed data likelihood as

$$p(\boldsymbol{x}_{\rm obs},\boldsymbol{y}_{\rm obs}|\boldsymbol{\theta},\boldsymbol{\psi}_{\rm obs}) \propto \prod_{i=1}^{n}\int\int p(x_i|\xi_i)p(y_i|\eta_i) \times p(\eta_i|\xi_i,\boldsymbol{\theta})p(\xi_i|I_i=1,\boldsymbol{\psi}_{\rm obs})d\xi_i\,d\eta_i, \quad (39)$$

where $p(\xi_i|I_i=1,\boldsymbol{\psi}_{\rm obs})$ is the distribution of those $\boldsymbol{\xi}$ included in one's sample. Here I have omitted terms depending on N, because one is primarily interested in inference on the regression parameters $\boldsymbol{\theta}$. Equation (39) is identical to equation (11), with the exception that $p(\boldsymbol{\xi}|\boldsymbol{\psi})$ now only models the distribution of those $\boldsymbol{\xi}$ that have been included in one's sample, and I have now assumed that the measurement errors on $\boldsymbol{y}$ and $\boldsymbol{x}$ are independent. In particular, for the Gaussian mixture models described in § 4.1 and § 4.3, the observed data likelihood is given by equations (16) and (29), where $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\tau}^2$ (or $\mathbf{T}$) should be understood as referring to the parameters for the distribution of the observed $\boldsymbol{\xi}$. As is evident from the similarity between equations (39) and (11), if the sample is selected based on the measured independent variables, then inference on the regression parameters $\boldsymbol{\theta}$ is unaffected by selection effects.

5.1.2. *Selection Based on Measured Dependent and Independent Variables*

If the method in which a sample is selected depends on the measured dependent variable, $\boldsymbol{y}$, the observed data likelihood becomes more complicated. As an example, one might encounter this situation if one uses an X-ray–selected sample to investigate the dependence of X-ray luminosity on optical luminosity and redshift. In this case, the selection function of the sample depends on both the X-ray luminosity and redshift and is thus no longer independent of the dependent variable. Such data sets are said to be "truncated."

If the selection function depends on $\boldsymbol{y}$, one cannot simply ignore the terms depending on N, since the $N-n$ integrals in equation (38) depend on $\boldsymbol{\theta}$. However, we can eliminate the dependence of equation (38) on the unknown N by applying a Bayesian approach. The posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and N is related to the observed data likelihood function as $p(\boldsymbol{\theta},\boldsymbol{\psi},N|\boldsymbol{x}_{\rm obs},\boldsymbol{y}_{\rm obs}) \propto p(\boldsymbol{\theta},\boldsymbol{\psi},N)p(\boldsymbol{x}_{\rm obs},\boldsymbol{y}_{\rm obs}|\boldsymbol{\theta},\boldsymbol{\psi},N)$, where $p(\boldsymbol{\theta},\boldsymbol{\psi},N)$ is the prior distribution of $(\boldsymbol{\theta},\boldsymbol{\psi},N)$. If we assume a uniform prior on $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\log N$, then one can show (e.g., Gelman et al. 2004) that the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is

$$p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{x}_{\rm obs},\boldsymbol{y}_{\rm obs}) \propto [\,p(I=1|\boldsymbol{\theta},\boldsymbol{\psi})]^{-n}\prod_{i=1}^{n}p(x_i,y_i|\boldsymbol{\theta},\boldsymbol{\psi}), \quad (40)$$

where $p(x_i,y_i|\boldsymbol{\theta},\boldsymbol{\psi})$ is given by equation (11) and $p(I=1|\boldsymbol{\theta},\boldsymbol{\psi})$ is the probability of including a source in one's sample, given the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$,

$$p(I=1|\boldsymbol{\theta},\boldsymbol{\psi}) = \int\int p(I=1|x,y)p(x,y|\boldsymbol{\theta},\boldsymbol{\psi})dx\,dy. \quad (41)$$

I have left off the subscripts for the data points in equation (41), because the integrals are the same for each (x_j,y_j,ξ_j,η_j). If one assumes the Gaussian mixture model of §§ 4.1 and 4.3, then $p(x_i,y_i|\boldsymbol{\theta},\boldsymbol{\psi})$ is given by equations (16) or (29). The posterior mode can then be used as an estimate of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, which is found by maximizing equation (40).

5.2. *Nondetections*

In addition to issues related to the sample selection method, it is common in astronomical data to have nondetections. Such nondetections are referred to as "censored" data, and the standard procedure is to place an upper and/or lower limit on the censored data point. Methods of data analysis for censored data have been reviewed and proposed in the astronomical literature (e.g., Feigelson & Nelson 1985; Schmitt 1985; Marshall 1992; Akritas & Siebert 1996), and Isobe et al. (1986) describe censored regression when the variables are measured without error. See Feigelson (1992) for a review of censored data in astronomy.

To facilitate the inclusion of censored data, I introduce an additional indicator variable, $\boldsymbol{D}$, indicating whether a data point is censored or not on the dependent variable. If y_i is detected, then $D_i=1$, else if y_i is censored, then $D_i=0$. It is commonly the case that a source is considered "detected" if its measured flux falls above some multiple of the background noise level, say $3\,\sigma$. Then, in this case, the probability of detecting the source given the measured source flux y_i is $p(D_i=1|\,y_i)=1$ if $y_i>3\,\sigma$, and $p(D_i=0|\,y_i)=1$ if $y_i<3\,\sigma$. Since source detection depends on the measured flux, some sources with intrinsic flux $\boldsymbol{\eta}$ above the flux limit will have a measured flux $\boldsymbol{y}$ that falls below the flux limit. Similarly, some sources with intrinsic flux below the flux limit will have a measured flux above the flux limit.

I assume that a sample is selected based on the independent variables, i.e., $p(\boldsymbol{I}|\boldsymbol{x},\boldsymbol{y})=p(\boldsymbol{I}|\boldsymbol{x})$. It is difficult to imagine obtaining a censored sample if the sample is selected based on its dependent variable, as some of the values of $\boldsymbol{y}$ are censored and thus unknown. Therefore, I only investigate the effects of censoring on $\boldsymbol{y}$ when the probability that a source is included in the sample is independent of $\boldsymbol{y}$, given $\boldsymbol{x}$. In addition, I do not address the issue of censoring on the independent variable. Although such methods can be developed, it is probably simpler to just omit such data, as inference on the regression parameters is unaffected when a sample is selected based only on the independent variables (see § 5.1.1).

The observed data likelihood for an $\boldsymbol{x}$-selected sample is given by equation (39). We can modify this likelihood to account for censored $\boldsymbol{y}$ by including the indicator variable $\boldsymbol{D}$ and again integrating over the missing data,

$$p(\boldsymbol{x}_{\rm obs},\boldsymbol{y}_{\rm obs},\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\psi}_{\rm obs}) \propto \prod_{i\in\mathcal{A}_{\rm det}}p(x_i,y_i|\boldsymbol{\theta},\boldsymbol{\psi}_{\rm obs})\prod_{j\in\mathcal{A}_{\rm cens}}p(x_j|\boldsymbol{\psi}_{\rm obs}) \times\int p(D_j=0|\,y_j,x_j)p(\,y_j|x_j,\boldsymbol{\theta},\boldsymbol{\psi}_{\rm obs})dy_j, \quad (42)$$

where the first product is over the set of data points with detections, $\mathcal{A}_{\rm det}$, and the second product is over the set of data points with nondetections, $\mathcal{A}_{\rm cens}$. The conditional distribution $p(\,y_j|x_j,\boldsymbol{\theta},\boldsymbol{\psi}_{\rm obs})$ and the marginal distribution $p(x_j|\boldsymbol{\psi}_{\rm obs})$ for the Gaussian mixture model are both given in §§ 4.1 and 4.3. If the data points are measured without error and one assumes the normal regression model $p(\eta|\xi,\boldsymbol{\theta})=N(\eta|\alpha+\beta\xi,\sigma^2)$, then equation (42) becomes the censored data likelihood function described in Isobe et al. (1986). A MLE for censored regression with measurement errors is then obtained by maximizing equation (42).

6. COMPUTATIONAL METHODS

In this section I describe a Bayesian method for computing estimates of the regression parameters $\boldsymbol{\theta}$ and their uncertainties. The Bayesian approach calculates the posterior probability distribution of the model parameters, given the observed data, and therefore is accurate for both small and large sample sizes. The

posterior distribution follows from Bayes' formula as $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{x}, \boldsymbol{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi})$, where $p(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the prior distribution of the parameters. I describe some Markov chain methods for drawing random variables from the posterior, which can then be used to estimate quantities such as standard errors and confidence intervals on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Gelman et al. (2004) is a good reference on Bayesian methods, and Loredo (1992) gives a review of Bayesian methods intended for astronomers. Further details of Markov chain simulation, including methods for making the simulations more efficient, can be found in Gelman et al. (2004).

6.1. *The Prior Density*

In order to ensure a proper posterior for the Gaussian mixture model, it is necessary to invoke a proper prior density on the mixture parameters (Roeder & Wasserman 1997). I adopt a uniform prior on the regression parameters $(\alpha, \beta, \sigma^2)$ and take $\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(1, \ldots, 1)$. The Dirichlet density is a multivariate extension of the Beta density, and the Dirichlet$(1, \ldots, 1)$ prior adopted in this work is equivalent to a uniform prior on $\boldsymbol{\pi}$, under the constraint $\sum_{k=1}^{K} \pi_k = 1$.

The prior on $\boldsymbol{\mu}$ and $\boldsymbol{\tau}^2$ (or $\mathbf{T}$) adopted in this work is very similar to that advocated by Roeder & Wasserman (1997) and Carroll et al. (1999). I adopt a normal prior on the individual μ_k with mean μ_0 and variance u^2 (or covariance matrix $\mathbf{U}$). This reflects our prior belief that the distribution of ξ is more likely to be fairly unimodal and, thus, that we expect it to be more likely that the individual Gaussian functions will be close together than far apart. If there is only one covariate, then I adopt a scaled inverse χ^2 prior on the individual τ_k^2 with scale parameter w^2 and one degree of freedom; otherwise, if there are $p > 1$ covariates, I adopt an inverse Wishart prior on the individual $\mathbf{T}_k$ with scale matrix $\mathbf{W}$ and p degrees of freedom. This reflects our prior expectation that the variances for the individual Gaussian components should be similar, but the low number of degrees of freedom accommodates a large range of scales. Both the Gaussian means and variances are assumed to be independent in their prior distribution, and the "hyperparameters" μ_0, u^2 (or $\mathbf{U}$), and w^2 (or $\mathbf{W}$) are left unspecified. By leaving the parameters for the prior distribution unspecified, they become additional parameters in the statistical model and therefore are able to adapt to the data.

Since the hyperparameters are left as free parameters, they also require a prior density. I assume a uniform prior on μ_0 and w^2 (or $\mathbf{W}$). If there is one covariate, then I assume a scaled inverse χ^2 prior for u^2 with scale parameter w^2 and one degree of freedom; otherwise, if there are multiple covariates, I assume an inverse Wishart prior for $\mathbf{U}$ with scale matrix $\mathbf{W}$ and p degrees of freedom. The prior on u^2 (or $\mathbf{U}$) reflects the prior expectation that the dispersion of the Gaussian components about their mean μ_0 should be on the order of the typical dispersion of each individual Gaussian function. The prior density for one covariate is then $p(\boldsymbol{\theta}, \boldsymbol{\psi}, \mu_0, u^2, w^2) \propto p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\mu_0, u^2)p(\tau^2|w^2)p(u^2|w^2)$ and is summarized hierarchically as

$$\alpha, \beta \sim \text{Uniform}(-\infty, \infty), \tag{43}$$

$$\sigma^2 \sim \text{Uniform}(0, \infty), \tag{44}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1, \ldots, 1), \tag{45}$$

$$\mu_1, \ldots, \mu_K|\mu_0, u^2 \sim N(\mu_0, u^2), \tag{46}$$

$$\tau_1^2, \ldots, \tau_K^2, u^2|w^2 \sim \text{Inv } \chi^2(1, w^2), \tag{47}$$

$$\mu_0 \sim \text{Uniform}(-\infty, \infty), \tag{48}$$

$$w^2 \sim \text{Uniform}(0, \infty). \tag{49}$$

The prior density for multiple covariates is just the multivariate extension of equations (43)–(49).

6.2. *Markov Chains for Sampling from the Posterior Distribution*

The posterior distribution summarizes our knowledge about the parameters in the statistical model, given the observed data and the priors. Direct computation of the posterior distribution is too computationally intensive for the model described in this work. However, we can obtain any number of random draws from the posterior using Markov chain Monte Carlo (MCMC) methods. In MCMC methods, we simulate a Markov chain that performs a random walk through the parameter space, saving the locations of the walk at each iteration. Eventually, the Markov chain converges to the posterior distribution, and the saved parameter values can be treated as a random draw from the posterior. These random draws can then be used to estimate the posterior median for each parameter, the standard error for each parameter, or plot a histogram as an estimate of the marginal probability distribution for each parameter.

6.2.1. *Gibbs Sampler for the Gaussian Model*

The easiest method for sampling from the posterior is to construct a Gibbs sampler. The basic idea behind the Gibbs sampler is to construct a Markov chain, where new values of the model parameters and missing data are simulated at each iteration, conditional on the values of the observed data and the current values of the model parameters and the missing data. Within the context of the measurement error model considered in this work, the Gibbs sampler undergoes four different stages.

The first stage of the Gibbs sampler simulates values of the missing data, given the measured data and current parameter values, a process known as data augmentation. In this work, the missing data are $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and any nondetections. In addition, I introduce an additional latent variable, $\boldsymbol{G}_i$, which gives the class membership for the ith data point. The vector $\boldsymbol{G}_i$ has K elements, where $G_{ik} = 1$ if the ith data point comes from the kth Gaussian function, and $G_{ij} = 0$ if $j \neq k$. I will use $\boldsymbol{G}$ to refer to the set of n vectors $\boldsymbol{G}_i$. Noting that π_k gives the probability of drawing a data point from the kth Gaussian function, the mixture model for $\boldsymbol{\xi}$ may then be expressed hierarchically as

$$\boldsymbol{G}_i|\boldsymbol{\pi} \sim \text{Multinom}(1, \pi_1, \ldots, \pi_K), \tag{50}$$

$$\xi_i|G_{ik} = 1, \mu_k, \tau_k^2 \sim N\left(\mu_k, \tau_k^2\right), \tag{51}$$

where Multinom$(m, p_1, \ldots, p_K)$ is a multinomial distribution with m trials, where p_k is the probability of success for the kth class on any particular trial. The vector $\boldsymbol{G}_i$ is also considered to be missing data and is introduced to simplify construction of the Gibbs sampler.

The new values of the missing data simulated in the data augmentation step are then used to simulate new values of the regression and Gaussian mixture parameters. The second stage of the Gibbs sampler simulates values of the regression parameters $\boldsymbol{\theta}$, given the current values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. The third stage simulates values of the mixture parameters $\boldsymbol{\psi}$, given the current values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. The fourth stage uses the new values of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ to update the parameters of the prior density. The values of the parameters are saved, and the process is repeated, creating a Markov chain. After a large number of iterations, the Markov chain converges, and the saved values of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ from the latter part of the algorithm

may then be treated as a random draw from the posterior distribution, $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{x}, \boldsymbol{y})$. Methods for simulating random variables from the distributions used for this Gibbs sampler are described in various works (e.g., Ripley 1987; Press et al. 1992; Gelman et al. 2004).

A Gibbs sampler for the Gaussian mixture model is presented below.

1. Start with initial guesses for $\boldsymbol{\eta}$, $\boldsymbol{G}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and the prior parameters.

2. If there are any nondetections, then draw y_i for the censored data points from $p(y_i|\eta_i, D_i = 0) \propto p(D_i = 0|y_i)p(y_i|\eta_i)$. This may be done by first drawing y_i from $p(y_i|\eta_i)$,

$$y_i|\eta_i \sim N(\eta_i, \sigma^2_{y,i}). \tag{52}$$

One then draws a random variable u_i, uniformly distributed on $[0, 1]$. If $u_i < p(D_i = 0|y_i)$, then the value of y_i is kept; otherwise, one draws a new value of y_i and u_i until $u_i < p(D_i = 0|y_i)$.

3. Draw values of $\boldsymbol{\xi}$ from $p(\boldsymbol{\xi}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{G}, \boldsymbol{\theta}, \boldsymbol{\psi})$. The distribution $p(\boldsymbol{\xi}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{G}, \boldsymbol{\theta}, \boldsymbol{\psi})$ can be derived from equations (12)–(14) or (26)–(28) and the properties of the multivariate normal distribution.

a) If there is only one independent variable, then ξ_i is updated as

$$\xi_i|x_i, y_i, \eta_i, \boldsymbol{G}_i, \boldsymbol{\theta}, \boldsymbol{\psi} \sim N(\hat{\xi}_i, \sigma^2_{\hat{\xi},i}), \tag{53}$$

$$\hat{\xi}_i = \sum_{k=1}^{K} G_{ik}\hat{\xi}_{ik}, \tag{54}$$

$$\hat{\xi}_{ik} = \sigma^2_{\hat{\xi},i}\left[\frac{\hat{\xi}_{xy,i}}{\sigma^2_{x,i}(1-\rho^2_{xy,i})} + \frac{\beta(\eta_i - \alpha)}{\sigma^2} + \frac{\mu_k}{\tau_k^2}\right], \tag{55}$$

$$\hat{\xi}_{xy,i} = x_i + \frac{\sigma_{xy,i}}{\sigma^2_{y,i}}(\eta_i - y_i), \tag{56}$$

$$\sigma^2_{\hat{\xi},i} = \sum_{k=1}^{K} G_{ik}\sigma^2_{\hat{\xi},ik}, \tag{57}$$

$$\sigma^2_{\hat{\xi},ik} = \left[\frac{1}{\sigma^2_{x,i}(1-\rho^2_{xy,i})} + \frac{\beta^2}{\sigma^2} + \frac{1}{\tau_k^2}\right]^{-1}, \tag{58}$$

where $\rho_{xy,i} = \sigma_{xy,i}/(\sigma_{x,i}\sigma_{y,i})$ is the correlation between the measurement errors on x_i and y_i. Note that ξ_i is updated using only information from the kth Gaussian function, since $G_{ij} = 1$ only for $j = k$ and $G_{ij} = 0$ otherwise.

b) If there are multiple independent variables, I have found it easier and computationally faster to update the values of $\boldsymbol{\xi}_i$ using a scalar Gibbs sampler. In this case, the p elements of $\boldsymbol{\xi}_i$ are updated individually. I denote ξ_{ij} to be the value of the jth independent variable for the ith data point and x_{ij} to be the measured value of ξ_{ij}. In addition, I denote $\boldsymbol{\xi}_{i,-j}$ to be the $(p-1)$-element vector obtained by removing ξ_{ij} from $\boldsymbol{\xi}_i$, i.e., $\boldsymbol{\xi}_{i,-j} = (\xi_{i1}, \ldots, \xi_{i(j-1)}, \xi_{i(j+1)}, \ldots, \xi_{ip})$. Similarly, $\boldsymbol{\beta}_{-j}$ denotes the $(p-1)$-element vector of regression coefficients obtained after removing β_j from $\boldsymbol{\beta}$. Then, ξ_{ij} is updated as

$$\xi_{ij}|\boldsymbol{x}_i, y_i, \boldsymbol{G}_i, \boldsymbol{\xi}_{i,-j}, \eta_i, \boldsymbol{\theta}, \boldsymbol{\psi} \sim N(\hat{\xi}_{ij}, \sigma^2_{\hat{\xi},ij}), \tag{59}$$

$$\hat{\xi}_{ij} = \sum_{k=1}^{K} G_{ik}\hat{\xi}_{ijk}, \tag{60}$$

$$\hat{\xi}_{ijk} = \frac{(\Sigma_i^{-1}\boldsymbol{z}_i^*)_{j+1} + (T_k^{-1}\mu_{ik}^*)_j + \beta_j(\eta_i - \alpha - \boldsymbol{\beta}_{-j}^T\boldsymbol{\xi}_{i,-j})/\sigma^2}{(\Sigma_i^{-1})_{(j+1)(j+1)} + (T_k^{-1})_{jj} + \beta_j^2/\sigma^2}, \tag{61}$$

$$(\boldsymbol{z}_i^*)_l = \begin{cases} y_i - \eta_i, & l = 1, \\ x_{il}, & l = j+1, \\ x_{il} - \xi_{il}, & l \neq j+1, \end{cases} \tag{62}$$

$$(\mu_{ik}^*)_l = \begin{cases} (\mu_k)_l, & l = j, \\ (\mu_k)_l - \xi_{il}, & l \neq j, \end{cases} \tag{63}$$

$$\sigma^2_{\hat{\xi},ij} = \sum_{k=1}^{K} G_{ik}\sigma^2_{\hat{\xi},ijk}, \tag{64}$$

$$\sigma^2_{\hat{\xi},ijk} = \left[(\Sigma_i^{-1})_{(j+1)(j+1)} + (T_k^{-1})_{jj} + \frac{\beta_j^2}{\sigma^2}\right]^{-1}, \tag{65}$$

where $\boldsymbol{z}_i^*$ is a $(p+1)$-element vector obtained by subtracting (η_i, ξ_i) from $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)$, with the exception of the jth element of $\boldsymbol{\xi}_i$; instead, the $(j+1)$th element of $\boldsymbol{z}_i^*$ is just x_{ij}. The p-element vector $\boldsymbol{\mu}_{ik}^*$ is obtained in an equivalent manner. The $(p+1) \times (p+1)$ matrix $\boldsymbol{\Sigma}_i$ is the covariance matrix of the measurement errors on $\boldsymbol{z}_i$. The term $(\Sigma_i^{-1}\boldsymbol{z}_i^*)_{(j+1)}$ denotes the $(j+1)$th element of the vector $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{z}_i^*$ and likewise for $(T_k^{-1}\mu_{ik}^*)_j$. The terms $(\Sigma_i^{-1})_{(j+1)(j+1)}$ and $(T_k^{-1})_{jj}$ denote the $(j+1)$th and jth elements of the diagonals of $\boldsymbol{\Sigma}_i^{-1}$ and $\mathbf{T}_k^{-1}$, respectively. This step is repeated until all p independent variables have been updated for each data point.

If any of the $\boldsymbol{\xi}_i$ are measured without error, then one simply sets $\boldsymbol{\xi}_i = \boldsymbol{x}_i$ for those data points.

4. Draw values of $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\theta})$. Similar to $\boldsymbol{\xi}$, the distribution $p(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\theta})$ can be derived from equations (12)–(14) or (26)–(28) and the properties of the multivariate normal distribution.

a) If there is only one covariate, then $\boldsymbol{\eta}$ is updated as

$$\eta_i|x_i, y_i, \xi_i, \boldsymbol{\theta} \sim N(\hat{\eta}_i, \sigma^2_{\hat{\eta},i}), \tag{66}$$

$$\hat{\eta}_i = \sigma^2_{\hat{\eta},i}\left[\frac{y_i + \sigma_{xy,i}(\xi_i - x_i)/\sigma^2_{x,i}}{\sigma^2_{y,i}(1-\rho^2_{xy,i})} + \frac{\alpha + \beta\xi_i}{\sigma^2}\right], \tag{67}$$

$$\sigma^2_{\hat{\eta},i} = \left[\frac{1}{\sigma^2_{y,i}(1-\rho^2_{xy,i})} + \frac{1}{\sigma^2}\right]^{-1}. \tag{68}$$

b) If there are multiple covariates, then $\boldsymbol{\eta}$ is updated as

$$\eta_i|\boldsymbol{x}_i, y_i, \boldsymbol{\xi}_i, \boldsymbol{\theta} \sim N(\hat{\eta}_i, \sigma^2_{\hat{\eta},i}), \tag{69}$$

$$\hat{\eta}_i = \frac{(\Sigma_i^{-1}\boldsymbol{z}_i^*)_1 + (\alpha + \boldsymbol{\beta}^T\boldsymbol{\xi}_i)/\sigma^2}{(\Sigma_i^{-1})_{11} + 1/\sigma^2}, \tag{70}$$

$$\sigma^2_{\hat{\eta},i} = \left[(\Sigma_i^{-1})_{11} + \frac{1}{\sigma^2}\right]^{-1}, \tag{71}$$

$$\boldsymbol{z}_i^* = (y_i, \boldsymbol{x}_i - \boldsymbol{\xi}_i), \tag{72}$$

where $(\Sigma_i^{-1}\boldsymbol{z}_i^*)_1$ is the first element of the vector $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{z}_i^*$, $\boldsymbol{z}_i^*$ is a $(p+1)$-element vector whose first element is y_i and remaining elements are $\boldsymbol{x}_i - \boldsymbol{\xi}_i$, and $(\Sigma_i^{-1})_{11}$ is the first diagonal element of $\boldsymbol{\Sigma}_i^{-1}$.

If any of the $\boldsymbol{\eta}$ are measured without error, then one sets $\boldsymbol{\eta} = \boldsymbol{y}$ for those data points.

5. Draw new values of the Gaussian labels, $\boldsymbol{G}$. The conditional distribution of $\boldsymbol{G}_i$ is multinomial with number of trials $m = 1$ and group probabilities $q_k = p(G_{ik} = 1|\xi_i, \psi)$,

$$\boldsymbol{G}_i|\xi_i, \boldsymbol{\psi} \sim \text{Multinom}(1, q_1, \ldots, q_K), \tag{73}$$

$$q_k = \frac{\pi_k N_p(\xi_i|\mu_k, T_k)}{\sum_{j=1}^{K} \pi_j N_p(\xi_i|\mu_j, T_j)}. \tag{74}$$

Note that if there is only one covariate then $p = 1$ and $T_k = \tau_k^2$.

6. Draw (α, β) from $p(\alpha, \beta|\boldsymbol{\xi}, \boldsymbol{\eta}, \sigma^2)$. Given $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and σ^2, the distribution of α and β is obtained by ordinary regression,

$$\alpha, \beta|\boldsymbol{\xi}, \boldsymbol{\eta}, \sigma^2 \sim N_{p+1}(\hat{\mathbf{c}}, \boldsymbol{\Sigma}_{\hat{\mathbf{c}}}), \tag{75}$$

$$\hat{\mathbf{c}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta}, \tag{76}$$

$$\boldsymbol{\Sigma}_{\hat{\mathbf{c}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2, \tag{77}$$

where $\mathbf{X}$ is a $n \times (p + 1)$ matrix, where the first column is a column of ones, the second column contains the n values of $\boldsymbol{\xi}_i$ for the first independent variable, the third column contains the n values of $\boldsymbol{\xi}_i$ for the second independent variable, etc.

7. Draw a new value of σ^2 from $p(\sigma^2|\boldsymbol{\xi}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta})$. The distribution $p(\sigma^2|\boldsymbol{\xi}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta})$ is derived by noting that given α, β, and $\boldsymbol{\xi}_i$, η_i is normally distributed with mean $\alpha + \boldsymbol{\beta}^T\boldsymbol{\xi}_i$ and variance σ^2. Reexpressing this distribution in terms of σ^2 instead of $\boldsymbol{\eta}$ and taking the product of the distributions for each data point, it follows that σ^2 has a scaled inverse χ^2 distribution,

$$\sigma^2|\boldsymbol{\xi}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta} \sim \text{Inv } \chi^2\left(\nu, s^2\right), \tag{78}$$

$$\nu = n - 2, \tag{79}$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\eta_i - \alpha - \boldsymbol{\beta}^T\boldsymbol{\xi}_i\right)^2. \tag{80}$$

8. Draw new values of the group proportions, $\boldsymbol{\pi}$. Given $\boldsymbol{G}$, $\boldsymbol{\pi}$ follows a Dirichlet distribution,

$$\boldsymbol{\pi}|\boldsymbol{G} \sim \text{Dirichlet}(n_1 + 1, \ldots, n_K + 1), \tag{81}$$

$$n_k = \sum_{i=1}^{n} G_{ik}. \tag{82}$$

Note that n_k is the number of data points that belong to the kth Gaussian function.

9. Draw a new value of μ_k from $p(\mu_k|\boldsymbol{\xi}, \boldsymbol{G}, \mathbf{T}_k, \mu_0, \mathbf{U})$. If there is only one independent variable, then $\mathbf{T}_k = \tau_k^2$ and $\mathbf{U} = u^2$. The new value of μ_k is simulated as

$$\mu_k|\boldsymbol{\xi}, \boldsymbol{G}, \mathbf{T}_k, \mu_0, \mathbf{U} \sim N_p\left(\hat{\mu}_k, \boldsymbol{\Sigma}_{\hat{\mu}_k}\right), \tag{83}$$

$$\hat{\mu}_k = \left(\mathbf{U}^{-1} + n_k\mathbf{T}_k^{-1}\right)^{-1}\left(\mathbf{U}^{-1}\mu_0 + n_k\mathbf{T}_k^{-1}\bar{\xi}_k\right), \tag{84}$$

$$\bar{\xi}_k = \frac{1}{n_k}\sum_{i=1}^{n} G_{ik}\xi_i, \tag{85}$$

$$\boldsymbol{\Sigma}_{\hat{\mu}_k} = \left(\mathbf{U}^{-1} + n_k\mathbf{T}_k^{-1}\right)^{-1}. \tag{86}$$

10. Draw a new value of τ_k^2 or $\mathbf{T}_k$. The distribution of $\boldsymbol{\tau}^2|\boldsymbol{\xi}, \boldsymbol{\mu}$ or $\mathbf{T}_k|\boldsymbol{\xi}, \boldsymbol{\mu}$ is derived in a manner similar to $\sigma^2|\boldsymbol{\xi}, \boldsymbol{\eta}, \alpha, \boldsymbol{\beta}$, after noting that the prior is conjugate for this likelihood. The distribution of $\tau_k^2|\boldsymbol{\xi}, \boldsymbol{\mu}$ is a scaled inverse χ^2 distribution, and the distribution of $\mathbf{T}_k|\boldsymbol{\xi}, \boldsymbol{\mu}$ is an inverse Wishart distribution.

a) If there is only one independent variable, then draw

$$\tau_k^2|\boldsymbol{\xi}, \boldsymbol{G}, \mu_k, w^2 \sim \text{Inv } \chi^2\left(\nu_k, t_k^2\right), \tag{87}$$

$$\nu_k = n_k + 1, \tag{88}$$

$$t_k^2 = \frac{1}{n_k + 1}\left[w^2 + \sum_{i=1}^{n} G_{ik}(\xi_i - \mu_k)^2\right]. \tag{89}$$

b) If there are multiple independent variables, then draw

$$\mathbf{T}_k|\boldsymbol{\xi}, \boldsymbol{G}, \boldsymbol{\mu}_k, \mathbf{W} \sim \text{Inv Wishart}_{\nu_k}(\mathbf{S}_k), \tag{90}$$

$$\nu_k = n_k + p, \tag{91}$$

$$\mathbf{S}_k = \mathbf{W} + \sum_{i=1}^{n} G_{ik}(\boldsymbol{\xi}_i - \boldsymbol{\mu}_k)(\boldsymbol{\xi}_i - \boldsymbol{\mu}_k)^T. \tag{92}$$

11. Draw a new value for $\boldsymbol{\mu}_0|\boldsymbol{\mu}, \mathbf{U}$. Noting that conditional on $\boldsymbol{\mu}_0$ and $\mathbf{U}$, $\mu_1, \ldots, \mu_K$ are independently distributed as $N_p(\boldsymbol{\mu}_0, \mathbf{U})$, it is straightforward to show that

$$\boldsymbol{\mu}_0|\boldsymbol{\mu}, \mathbf{U} \sim N_p(\bar{\boldsymbol{\mu}}, \mathbf{U}/K), \tag{93}$$

$$\bar{\mu} = \frac{1}{K}\sum_{k=1}^{K}\mu_k. \tag{94}$$

If there is only one covariate, then $p = 1$ and $\mathbf{U} = u^2$.

12. Draw a new value for u^2 or $\mathbf{U}$, given μ_0, $\boldsymbol{\mu}$, and w^2 (or $\mathbf{W}$). Similar to the case for τ_k^2 or $\mathbf{T}_k$, the conditional distribution of u^2 or $\mathbf{U}$ is scaled inverse χ^2 or inverse Wishart.

a) If there is only one covariate, then

$$u^2|\mu_0, \boldsymbol{\mu}, w^2 \sim \text{Inv } \chi^2\left(\nu_u, \hat{u}^2\right), \tag{95}$$

$$\nu_u = K + 1, \tag{96}$$

$$\hat{u}^2 = \frac{1}{\nu_u}\left[w^2 + \sum_{k=1}^{K}(\mu_k - \mu_0)^2\right]. \tag{97}$$

b) If there are multiple covariates, then

$$\mathbf{U}|\boldsymbol{\mu}_0, \boldsymbol{\mu}, \mathbf{W} \sim \text{Inv Wishart}_{\nu_U}\left(\hat{\mathbf{U}}\right), \tag{98}$$

$$\nu_U = K + p, \tag{99}$$

$$\hat{\mathbf{U}} = \mathbf{W} + \sum_{k=1}^{K}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T. \tag{100}$$

13. Finally, draw a new value of $w^2|u^2, \tau^2$ or $\mathbf{W}|\mathbf{U}, \mathbf{T}$.

a) If there is only one covariate, then $w^2|u^2, \tau^2$ is drawn from a Gamma distribution. This can be derived by noting that $p(w^2|u^2, \tau^2) \propto p(u^2|w^2)p(\tau^2|w^2)$ has the form of a Gamma distribution as a function of w^2. The new value of w^2 is then simulated as

$$w^2|u^2, \tau^2 \sim \text{Gamma}(a, b), \tag{101}$$

$$a = \frac{1}{2}(K + 3), \tag{102}$$

$$b = \frac{1}{2}\left(\frac{1}{u^2} + \sum_{k=1}^{K}\frac{1}{\tau_k^2}\right). \tag{103}$$

b) If there are multiple covariates, then $\mathbf{W}|\mathbf{U}, \mathbf{T}$ is drawn from a Wishart distribution. This can be derived by noting that

$p(\mathbf{W}|\mathbf{U},\mathbf{T}) \propto p(\mathbf{U}|\mathbf{W})p(\mathbf{T}|\mathbf{W})$ has the form of a Wishart distribution as a function of $\mathbf{W}$. The new value of $\mathbf{W}$ is then simulated as

$$\mathbf{W}|\mathbf{U},\mathbf{T} \sim \text{Wishart}_{\nu_W}\left(\hat{\mathbf{W}}\right), \tag{104}$$

$$\nu_W = (K+2)p + 1, \tag{105}$$

$$\hat{\mathbf{W}} = \left(\mathbf{U}^{-1} + \sum_{k=1}^{K} \mathbf{T}_k^{-1}\right)^{-1}. \tag{106}$$

After completing steps 2–13 above, an iteration of the Gibbs sampler is complete. One then uses the new simulated values of $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, $\boldsymbol{\theta}$, ψ, and the prior parameters and repeats steps 2–13. The algorithm is repeated until convergence, and the values of $\boldsymbol{\theta}$ and ψ at each iteration are saved. Upon reaching convergence, one discards the values of $\boldsymbol{\theta}$ and ψ from the beginning of the simulation, and the remaining values of α, $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\mu}$, and $\boldsymbol{\tau}^2$ (or $\mathbf{T}$) may be treated as a random draw from the posterior distribution, $p(\boldsymbol{\theta}, \psi|\boldsymbol{x}, \boldsymbol{y})$. One can then use these values to calculate estimates of the parameters and their corresponding variances and confidence intervals. The posterior distribution of the parameters can also be estimated from these values of $\boldsymbol{\theta}$ and ψ using histogram techniques. Techniques for monitoring convergence of the Markov chains can be found in Gelman et al. (2004).

The output from the Gibbs sampler may be used to perform Bayesian inference on other quantities of interest. In particular, the Pearson linear correlation coefficient, ρ, is often used in assessing the strength of a relationship between the $\boldsymbol{x}$ and $\boldsymbol{y}$. A random draw from the posterior distribution for the correlation between $\boldsymbol{\eta}$ and ξ_j, denoted as ρ_j, can be calculated from equation (5) for each draw from the Gibbs sampler. For the Gaussian mixture model, the variance $\text{Var}(\boldsymbol{\eta})$ and covariance matrix $\boldsymbol{\Sigma}_\xi \equiv \text{Var}(\boldsymbol{\xi})$ are

$$\text{Var}(\eta) = \boldsymbol{\beta}^T \boldsymbol{\Sigma}_\xi \boldsymbol{\beta} + \sigma^2, \tag{107}$$

$$\boldsymbol{\Sigma}_\xi = \sum_{k=1}^{K} \pi_k \left(\mathbf{T}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right) - \bar{\boldsymbol{\xi}}\bar{\boldsymbol{\xi}}^T, \tag{108}$$

$$\bar{\boldsymbol{\xi}} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k, \tag{109}$$

and $\text{Var}(\xi_j)$ is the jth diagonal element of $\boldsymbol{\Sigma}_\xi$. The simplification for one covariate is self-evident.

If there is considerable posterior probability near $\sigma^2 \approx 0$ or $\tau_k^2 \approx 0$, then the Gibbs sampler can get "stuck." For example, if $\tau_k^2 \approx 0$, then step 3*a* of the Gibbs sampler will draw values of $\boldsymbol{\xi}|\boldsymbol{G} \approx \boldsymbol{\mu}_k$. Then, step 9 will produce a new value of $\boldsymbol{\mu}_k$ that is almost identical to the previous iteration, step 10*a* will produce a new value of $\tau_k^2 \approx 0$, and so on. The Gibbs sampler will eventually get "unstuck," but this can take a long time and result in very slow convergence. In particular, it is very easy for the Gibbs sampler to get stuck if the measurement errors are large relative to σ^2 or τ_k^2 or if the number of data points is small. In this situation I have found it useful to use the Metropolis-Hastings algorithm instead.

6.2.2. *Metropolis-Hastings Algorithm*

If the selection function is not independent of $\boldsymbol{y}$, given the independent variables (see eq. [40]), then posterior simulation based on the Gibbs sampler is more complicated. In addition, if the measurement errors are large compared to the intrinsic dispersion in the data or if the sample size is small, then the Gibbs sampler can become stuck and extremely inefficient. In both of these cases, one can use the Metropolis-Hastings algorithm (Metropolis & Ulam 1949; Metropolis et al. 1953; Hastings 1970) to sample from the posterior distribution, as the Metropolis-Hastings algorithm can avoid constructing Markov chains for $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. For a description of the Metropolis-Hastings algorithm, we refer the reader to Chib & Greenberg (1995) or Gelman et al. (2004).

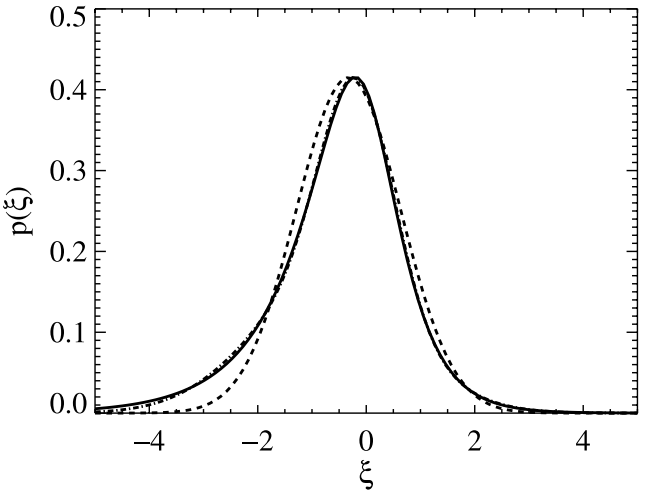

FIG. 2.—Actual distribution of ξ (*solid line*) for the simulations, compared with the best-fitting one (*dashed line*) and two (*dash-dotted line*) Gaussian fit. The two Gaussian fit is nearly indistinguishable from the true $p(\xi)$. Although the one Gaussian fit provides a reasonable approximation to the distribution of ξ, it is not able to pick up the asymmetry in $p(\xi)$.

7. SIMULATIONS

In this section I perform simulations to illustrate the effectiveness of the Gaussian structural model for estimating the regression parameters, even in the presence of severe measurement error and censoring. In addition, I compare the OLS, BCES($Y|X$), and FITEXY estimators with a MLE based on the Gaussian mixture model with $K = 1$ Gaussian function.

7.1. *Simulation Without Nondetections*

The first simulation I performed is for a simple regression with one independent variable. I generated 2.7×10^5 data sets by first drawing n values of the independent variable, ξ, from a distribution of the form

$$p(\xi) \propto e^{\xi}\left(1 + e^{2.75\xi}\right)^{-1}. \tag{110}$$

The distribution of ξ is shown in Figure 2, along with the best-fitting one and two Gaussian approximations. In this case, the two Gaussian mixture is nearly indistinguishable from the actual distribution of ξ and, thus, should provide an excellent approximation to $p(\xi)$. The values for ξ had a mean of $\mu = -0.493$ and a dispersion of $\tau = 1.200$. I varied the number of data points in the simulated data sets as $n = 25$, 50, and 100. I then simulated values of η according to equation (1), with $\alpha = 1.0$ and $\beta = 0.5$. The intrinsic scatter, ϵ, had a normal distribution with mean zero and standard deviation $\sigma = 0.75$, and the correlation between η and ξ was $\rho \approx 0.62$. The joint distribution of ξ and η for one simulated data set with $n = 50$ is shown in Figure 3.

Measured values for ξ and η were simulated according to equations (2) and (3). The measurement errors had a zero mean normal distribution of varying dispersion and were independent

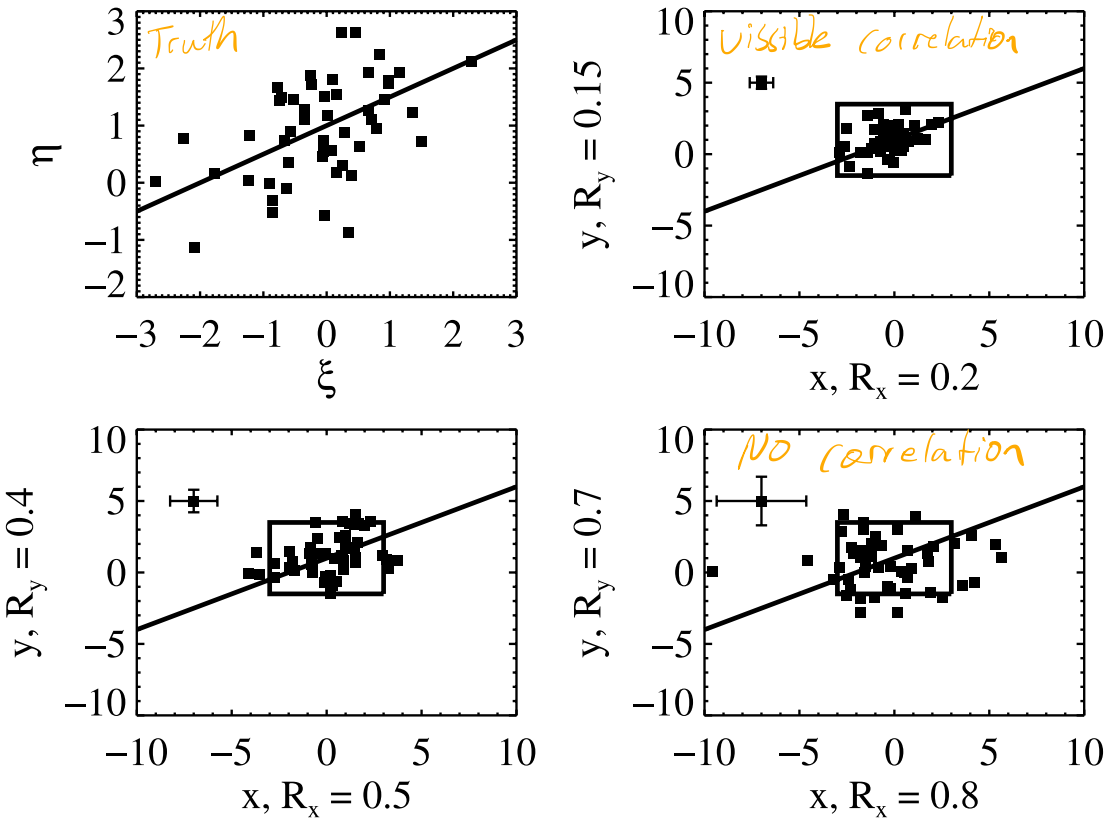

FIG. 3.—Distributions of the simulated data for various levels of measurement error (see § 7.1). The top left panel shows the distribution of η as a function of ξ for one simulated data set; the solid line is the true value of the regression line. The remaining panels show the distributions of the observed values, y and x, for various levels of measurement error. The data point with error bars in each panel is a fictitious data point and is used to illustrate the median values of the error bars. The box outlines the bounds of the plot of η against ξ. As can be seen, large measurement errors wash out any visual evidence for a correlation between the variables.

for x and y. The variances in the measurement errors, $\sigma_{x,i}^2$ and $\sigma_{y,i}^2$, were different for each data point and drawn from a scaled inverse χ^2 distribution. For the inverse χ^2 distribution, there were $\nu = 5$ degrees of freedom, and the scale parameters are denoted as t and s for the x and y measurement error variances, respectively. The scale parameters dictate the typical size of the measurements errors and were varied as $t = 0.5\tau$, τ, and 2τ and $s = 0.5\ \sigma$, σ, and $2\ \sigma$. These values corresponded to values of $R_x \sim 0.2$, 0.5, and 0.8 and $R_y \sim 0.15$, 0.4, and 0.6, respectively. I simulated 10^4 data sets for each grid point of t, s, and n, giving a total of 2.7×10^5 simulated data sets. The joint distributions of x and y for varying values of t/τ and s/σ are also shown in Figure 3. These values of x and y are the "measured" values of the simulated data set shown in the plot of η as a function of ξ.

For each simulated data set, I calculated the MLE, found by maximizing equation (16). For simplicity, I only use $K = 1$ Gaussian function. I also calculated the OLS, BCES($Y|X$), and FITEXY estimates for comparison. I calculated an OLS estimate of σ^2 by subtracting the average σ_y^2 from the variance in the regression residuals. If the OLS estimate of σ^2 was negative, I set $\hat{\sigma}_{\rm OLS} = 0$. Following Fuller (1987), I estimate σ^2 for a BCES($Y|X$)-type estimator as $\hat{\sigma}^2_{\rm BCES} = {\rm Var}(y) - \bar{\sigma}_y^2 - \hat{\beta}_{\rm BCES}{\rm Cov}(x, y)$, where $\bar{\sigma}_y^2$ is the average measurement error variance in y and $\hat{\beta}_{\rm BCES}$ is the BCES($Y|X$) estimate of the slope. If $\hat{\sigma}^2_{\rm BCES}$ is negative, I set $\hat{\sigma}_{\rm BCES} = 0$. Following Tremaine et al. (2002), I compute a FITEXY estimate of σ by increasing σ^2 until $\chi^2_{\rm EXY}/(n-2) = 1$ or assume $\sigma^2 = 0$ if $\chi^2_{\rm EXY}/(n-2) < 1$. The sampling distributions of the slope and intrinsic scatter estimators for $n = 50$ are shown in Figures 4 and 5 as functions of t/τ, and the results of the simulations are summarized in Table 1.

The bias of the OLS estimate is apparent, becoming more severe as the measurement errors in the independent variable increase. In addition, the variance in the OLS slope estimate decreases as the measurement errors in ξ increase, giving one the false impression that one's estimate of the slope is more precise when the measurement errors are large. This has the effect of concentrating the OLS estimate of β around $\hat{\beta}_{\rm OLS} \sim 0$, thus effectively erasing any evidence of a relationship between the two variables. When the measurement errors are large, the OLS estimate of the intrinsic scatter, $\hat{\sigma}^2_{\rm OLS}$, is occasionally zero.

The BCES($Y|X$) estimator performs better than the OLS and FITEXY estimators, being approximately unbiased when the measurement errors are $\sigma_x/\tau \lesssim 1$. However, the BCES estimate of the slope, $\hat{\beta}_{\rm BCES} = {\rm Cov}(x, y)/[{\rm Var}(x) - \bar{\sigma}_x^2]$, suffers some bias when the measurement errors are large and/or the sample size is small. In addition, the variance in $\hat{\beta}_{\rm BCES}$ is larger than the MLE, and $\hat{\beta}_{\rm BCES}$ becomes considerably unstable when the measurement errors on ξ are large. This instability results because the denominator in the equation for $\hat{\beta}_{\rm BCES}$ is ${\rm Var}(x) - \bar{\sigma}_x^2$. If $\bar{\sigma}_x^2 \approx {\rm Var}(x)$, then the denominator is $\approx$0, and $\hat{\beta}_{\rm BCES}$ can become very large. Similar to the OLS and FITEXY estimates, the estimate of the intrinsic variance for the BCES-type estimator is often zero when the measurement errors are large, suggesting the false conclusion that there is no intrinsic scatter about the regression line.

The FITEXY estimator performed poorly in the simulations, being both biased and highly variable. The bias of the FITEXY estimator is such that $\hat{\beta}_{\rm EXY}$ tends to overestimate β, the severity of which tends to increase as R_y decreases. This upward bias in $\hat{\beta}_{\rm EXY}$ has been noted by Weiner et al. (2006), who also performed simulations comparing $\hat{\beta}_{\rm EXY}$ with $\hat{\beta}_{\rm BCES}$. They note that when

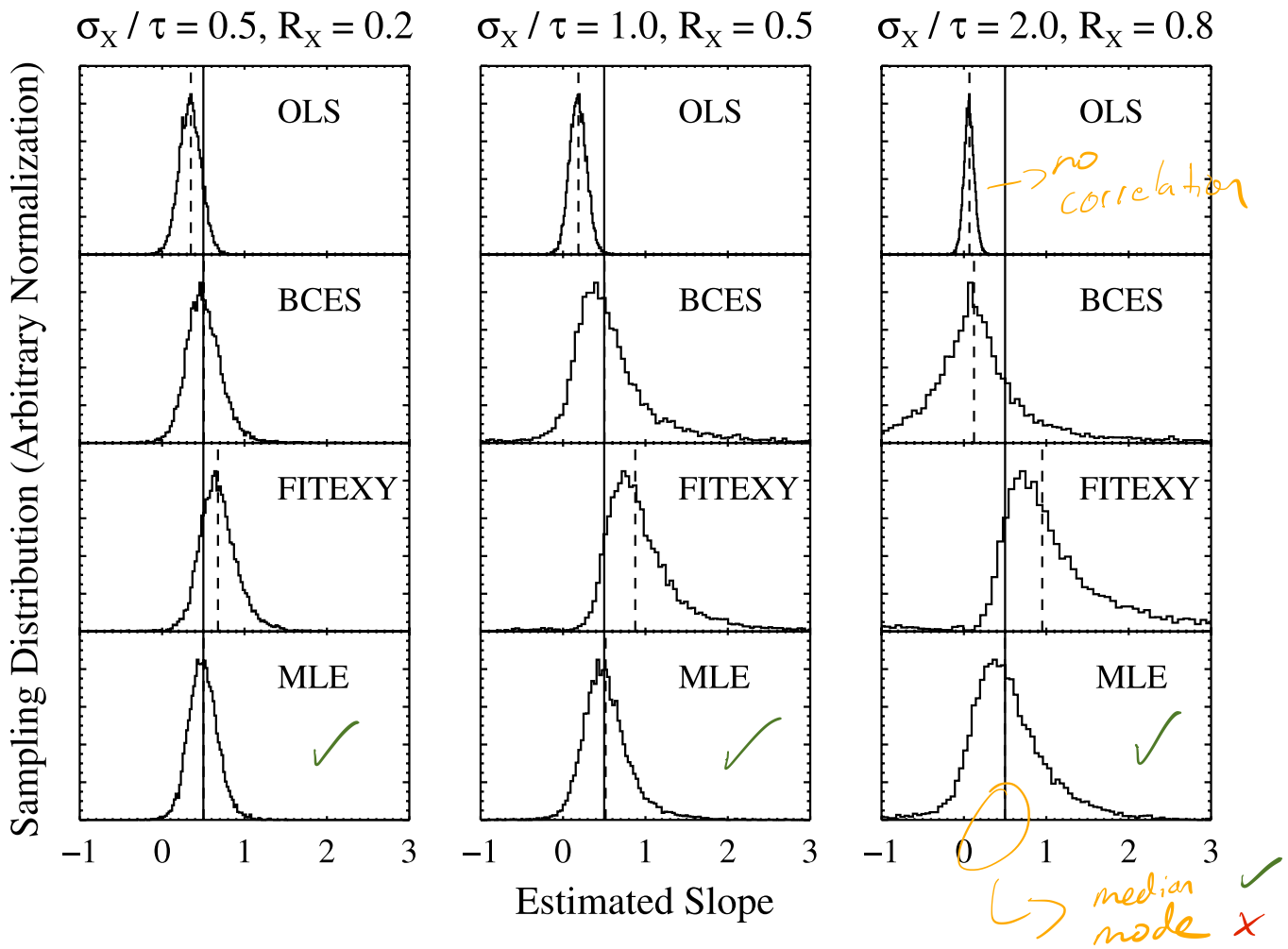

FIG. 4.—Sampling distributions of the slope estimators as functions of covariate measurement error magnitude for $n = 50$ data points and $\sigma_y \sim \sigma$, inferred from simulations (see § 7.1). The estimators are the ordinary least-squares estimator (OLS), the BCES($Y|X$) estimator, the FITEXY estimator, and the maximum-likelihood estimator (MLE) of the $K = 1$ Gaussian structural model. The solid vertical lines mark the true value of $\beta = 0.5$, and the dashed vertical lines mark the median values of each respective estimator. The OLS estimator is biased toward zero, while the FITEXY estimator is biased away from zero; in both cases, the bias gets worse for larger measurement errors. The BCES($Y|X$) estimator is, in general, unbiased, but can become biased and highly variable if the measurement errors becomes large. The MLE of the Gaussian model performs better than the other estimators, as it is approximately unbiased and less variable.

one minimizes $\chi^2_{\rm EXY}$ alternatively with respect to β and σ^2 and iterates until convergence, then the bias in $\hat{\beta}_{\rm EXY}$ can be improved. I have tested this and also find that the bias in $\hat{\beta}_{\rm EXY}$ is reduced, but at the cost of a considerable increase in variance in $\hat{\beta}_{\rm EXY}$. In general, our simulations imply that the variance of the FITEXY estimator is comparable to that of the BCES($Y|X$) estimator if one does not iterate the minimization of $\chi^2_{\rm EXY}$, and the variance of $\hat{\beta}_{\rm EXY}$ is larger if one does iterate. However, since $\hat{\beta}_{\rm BCES}$ is approximately unbiased when R_x is not too large, $\hat{\beta}_{\rm BCES}$ should be preferred over $\hat{\beta}_{\rm EXY}$. In addition, when the measurement errors are large the FITEXY estimate of σ is commonly $\hat{\sigma}_{\rm EXY} = 0$, similar to the BCES-type estimate of the intrinsic dispersion.

The MLE based on the Gaussian structural model performs better than the OLS, BCES, and FITEXY estimators and gives fairly consistent estimates even in the presence of severe measurement error and low sample size. The MLE is approximately unbiased, in spite of the fact that the MLE incorrectly assumes that the independent variables are normally distributed. The variance in the MLE of the slope, $\hat{\beta}_{\rm MLE}$, is smaller than that of $\hat{\beta}_{\rm BCES}$ and $\hat{\beta}_{\rm EXY}$, particularly when R_x is large. In contrast to the OLS estimate of the slope, the dispersion in $\hat{\beta}_{\rm MLE}$ increases as the measurement errors increase, reflecting the additional uncertainty in $\hat{\beta}_{\rm MLE}$ caused by the measurement errors. Finally, in contrast to the other estimators, the MLE of the intrinsic variance is always positive, and the probability of obtaining $\hat{\sigma}_{\rm MLE} = 0$ is negligible for these simulations.

I argued in § 4.1 that assuming a uniform distribution on ξ does not lead to better estimates than the usual OLS case. I also used these simulations to estimate the sampling density of the MLE assuming $p(\xi) \propto 1$. The results were nearly indistinguishable from the OLS estimator, supporting our conjecture that assuming $p(\xi) \propto 1$ does not offer an improvement over OLS.

While it is informative to compare the sampling distribution of our proposed MLE with those of the OLS, BCES($Y|X$), and FITEXY estimators, I do not derive the uncertainties in the regression parameters from the sampling distribution of the MLE. As described in § 6.2, we derive the uncertainties in the regression parameters by simulating draws from the posterior distribution, $p(\boldsymbol{\theta}, \boldsymbol{\psi}|x, y)$. This allows a straightforward method of interpreting the parameter uncertainties that does not rely on large-sample approximations, as the posterior distribution is the probability distribution of the parameters, given the observed data. The posterior distributions of ρ, β, and σ for a simulated data set with $n = 50$, $\sigma_x \sim \tau$, and $\sigma_y \sim \sigma$ is shown in Figure 6. When estimating these posteriors, I used $K = 2$ Gaussian functions in the mixture model. As can be seen from Figure 6, the true values of

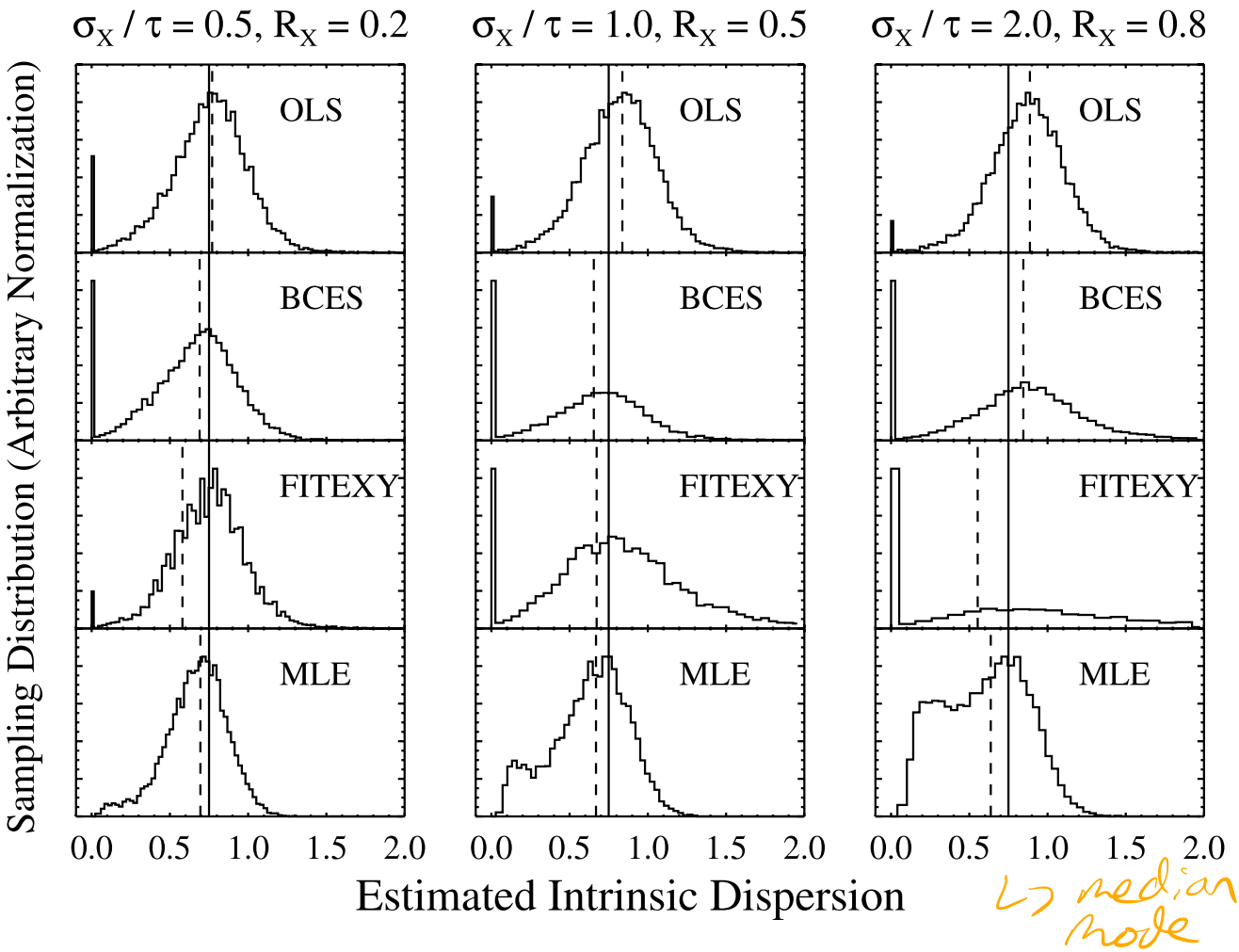

FIG. 5.—Same as Fig. 4, but for the standard deviation of the intrinsic scatter, σ. The solid vertical lines mark the true value of $\sigma = 0.75$, and the dashed vertical lines mark the median values of each respective estimator. All of the estimators exhibit some bias, and the BCES and FITEXY estimators can exhibit significant variance. Moreover, the BCES and FITEXY estimators both commonly have values of $\hat{\sigma} = 0$, misleading one into concluding that there is no intrinsic scatter; this occasionally occurs for the OLS estimate as well. In contrast, the MLE based on the Gaussian model does not suffer from this problem, at least for these simulations.

TABLE 1

DEPENDENCE OF THE ESTIMATOR SAMPLING DISTRIBUTIONS ON MEASUREMENT ERROR AND SAMPLE SIZE

		OLS		BCES($Y|X$)		FITEXY		MLE	
$t/\tau = s/\sigma$[a]	n[b]	$\hat{\beta}$[c]	$\hat{\sigma}$[d]	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\sigma}$
0.5.....	25	$0.357^{+0.242}_{-0.246}$	$0.784^{+0.717}_{-0.608}$	$0.518^{+0.513}_{-0.349}$	$0.687^{+0.714}_{-0.650}$	$0.896^{+1.127}_{-0.425}$	$0.855^{+1.616}_{-0.757}$	$0.513^{+0.393}_{-0.315}$	$0.677^{+0.663}_{-0.580}$
	50	$0.355^{+0.166}_{-0.164}$	$0.801^{+0.591}_{-0.528}$	$0.510^{+0.306}_{-0.233}$	$0.716^{+0.601}_{-0.540}$	$0.898^{+0.572}_{-0.298}$	$0.873^{+1.042}_{-0.668}$	$0.506^{+0.242}_{-0.212}$	$0.717^{+0.555}_{-0.507}$
	100	$0.354^{+0.117}_{-0.114}$	$0.810^{+0.488}_{-0.447}$	$0.506^{+0.197}_{-0.164}$	$0.743^{+0.494}_{-0.466}$	$0.895^{+0.352}_{-0.218}$	$0.885^{+0.786}_{-0.587}$	$0.504^{+0.162}_{-0.149}$	$0.732^{+0.456}_{-0.429}$
1.0.....	25	$0.190^{+0.255}_{-0.239}$	$0.798^{+1.047}_{-0.798}$	$0.442^{+2.763}_{-2.167}$	$0.610^{+1.418}_{-0.610}$	$0.827^{+2.293}_{-1.687}$	$0.727^{+2.899}_{-0.727}$	$0.524^{+0.907}_{-0.576}$	$0.572^{+0.903}_{-0.564}$
	50	$0.191^{+0.172}_{-0.164}$	$0.839^{+0.869}_{-0.752}$	$0.519^{+1.816}_{-0.707}$	$0.643^{+1.023}_{-0.643}$	$0.870^{+1.195}_{-0.459}$	$0.814^{+1.754}_{-0.814}$	$0.519^{+0.552}_{-0.370}$	$0.669^{+0.745}_{-0.643}$
	100	$0.189^{+0.121}_{-0.116}$	$0.862^{+0.726}_{-0.640}$	$0.520^{+0.913}_{-0.348}$	$0.687^{+0.784}_{-0.687}$	$0.895^{+0.665}_{-0.329}$	$0.855^{+1.246}_{-0.788}$	$0.502^{+0.337}_{-0.242}$	$0.714^{+0.623}_{-0.604}$
2.0.....	25	$0.066^{+0.243}_{-0.228}$	$0.565^{+1.797}_{-0.565}$	$0.036^{+2.761}_{-2.944}$	$0.663^{+2.544}_{-0.663}$	$0.443^{+3.793}_{-2.836}$	$0.000^{+2.994}_{-0.000}$	$0.366^{+1.468}_{-1.395}$	$0.381^{+1.223}_{-0.362}$
	50	$0.067^{+0.164}_{-0.158}$	$0.768^{+1.525}_{-0.768}$	$0.116^{+2.878}_{-2.951}$	$0.743^{+2.271}_{-0.743}$	$0.634^{+3.276}_{-3.027}$	$0.258^{+2.983}_{-0.258}$	$0.426^{+1.055}_{-0.918}$	$0.559^{+1.082}_{-0.529}$
	100	$0.065^{+0.113}_{-0.106}$	$0.843^{+1.293}_{-0.843}$	$0.209^{+2.936}_{-2.962}$	$0.743^{+1.932}_{-0.743}$	$0.765^{+2.492}_{-2.024}$	$0.627^{+2.928}_{-0.627}$	$0.444^{+0.698}_{-0.548}$	$0.673^{+0.921}_{-0.621}$

NOTES.—The values given for $\hat{\beta}$ and $\hat{\sigma}$ are the median and interval containing 90% of the estimates over the simulations. For example, when $t/\tau = s/\sigma = 0.5$ and $n = 25$, the median value of the OLS slope estimator is 0.357, and 90% of the values of $\hat{\beta}_{\rm OLS}$ are contained within $0.357^{+0.242}_{-0.246}$.

[a] Typical value of the measurement error magnitude for the simulations.

[b] The number of data points in the simulated data sets.

[c] The estimate of the slope, β. The true value is $\beta = 0.5$.

[d] The estimate of the dispersion in the intrinsic scatter, σ. The true value is $\sigma = 0.75$.

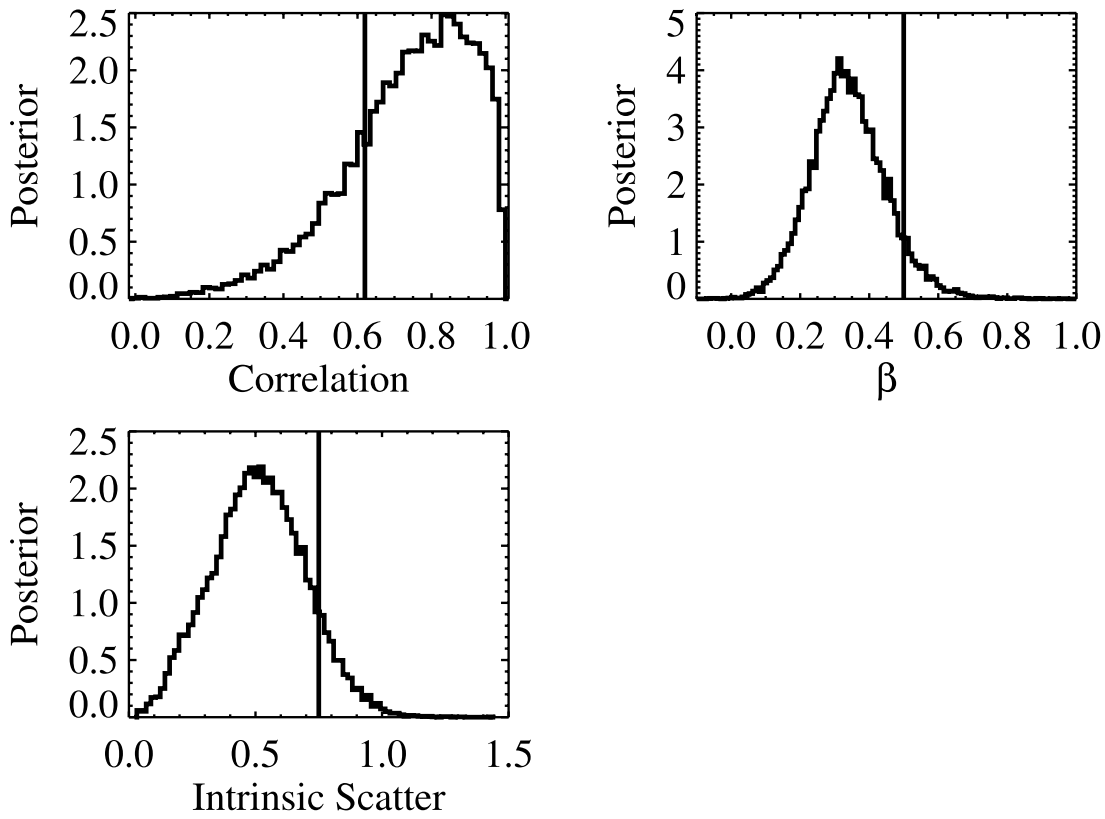

FIG. 6.—Marginal posterior distributions of the linear correlation coefficient, the regression slope, and the intrinsic dispersion for a simulated data set of $n = 50$ data points with $\sigma_x \sim \tau$ and $\sigma_y \sim \sigma$. The vertical lines mark the true values of the parameters. The true values of the regression parameters are contained within the spread of the marginal posteriors, implying that bounds on the regression parameters inferred from the posterior are trustworthy.

ρ, β, and σ are contained within the regions of nonnegligible posterior probability. I have estimated posteriors for other simulated data sets, varying the number of data points and the degree of measurement error. As one would expect, the uncertainties in the regression parameters, represented by the widths of the posterior distributions, increase as the size of the measurement errors increase and the sample size decreases.

A common frequentist approach is to compute the covariance matrix of the MLE by inverting the estimated Fisher information matrix evaluated at the MLE. Then, under certain regularity conditions, the MLE of the parameters is asymptotically normally distributed with mean equal to the true value of the parameters and covariance matrix equal to the inverse of the Fisher information matrix. Furthermore, under these regularity conditions the posterior distribution and sampling distribution of the MLE are asymptotically the same. Figure 7 compares the posterior distribution of the slope for a simulated data set with that inferred from the MLE. The posterior and MLE were calculated assuming $K = 1$ Gaussian function. As can be seen, the posterior distribution for β is considerably different from the approximation based on the MLE of β, and thus, the two have not converged for this sample. In particular, the posterior is more skewed and heavy tailed, placing more probability on values of $\beta > 0$ than does the distribution approximated by the MLE. Therefore, uncertainties in the MLE should be interpreted with caution if using the asymptotic approximation to the sampling distribution of the MLE.

7.2. *Simulation With Nondetections*

To assess the effectiveness of the Gaussian structural model in dealing with censored data sets with measurement error, I introduced nondetections into the simulations. The simulations were performed in an identical manner as that described in § 7.1, but now I only consider sources to be "detected" if $y > 1.5$. For those sources that were "censored" ($y < 1.5$), I placed an upper limit on them of $y = 1.5$.

I focus on the results for a simulated data set with $n = 100$ data points and measurement errors similar to the intrinsic dispersion in the data, $\sigma_y \sim \sigma$ and $\sigma_x \sim \tau$. The detection threshold of $y > 1.5$ resulted in a detection fraction of ~30%. This

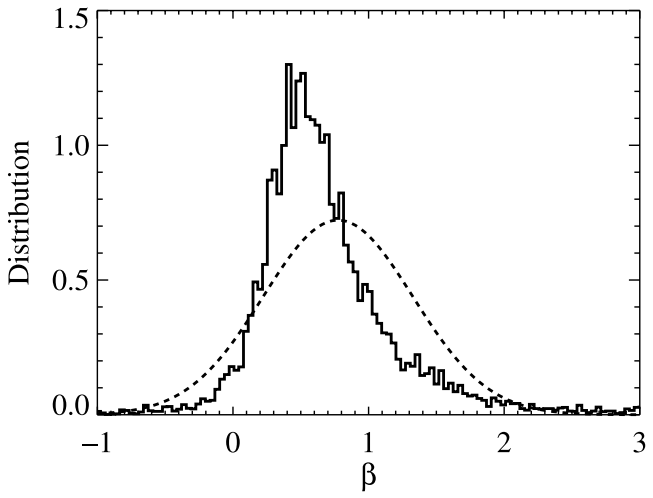

FIG. 7.—Posterior distributions of the slope (*solid histogram*), compared with the posterior approximated from the MLE and Fisher information matrix (*dashed line*), for a simulated data set of $n = 50$ data points with $\beta = 0.5$, $\sigma_x \sim \tau$, and $\sigma_y \sim \sigma$. The two distributions have not converged and the Bayesian and frequentist inference differ in this case, with the Bayesian approach placing more probability near $\beta \approx 0.5$ and on positive values of β.

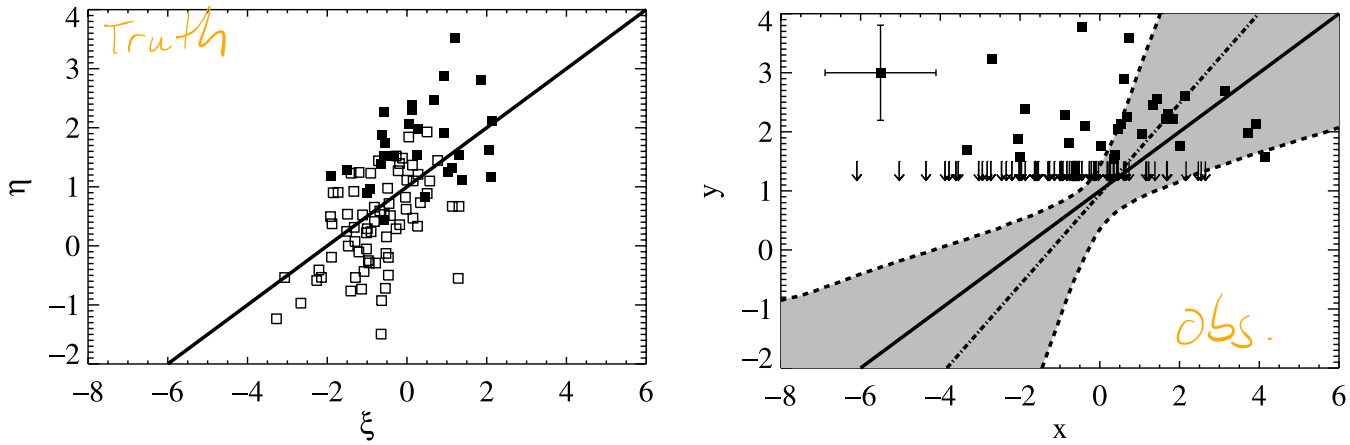

FIG. 8.—Distribution of η and ξ (*left*) and the measured values of y and x (*right*), from a simulated censored data set of $n = 50$ data points, $\sigma_x \sim \tau$, and $\sigma_y \sim \sigma$ (see § 7.2). In the plot of η and ξ, the filled squares denote the values of ξ and η for the detected data points, and the open squares denote the values of ξ and η for the undetected data points. The solid line in both plots is the true regression line. In the plot of y and x, the squares denote the measured values of x and y for the detected data points, and the arrows denote the "upper limits" on y for the undetected data points. The fictitious data point with error bars illustrates the median values of the error bars. The dash-dotted line shows the best-fit regression line, as calculated from the posterior median of α and β, and the shaded region defines the approximate 95% (2 σ) pointwise confidence intervals on the regression line. The true values of the regression line are contained within the 95% confidence intervals.

simulation represents a rather extreme case of large measurement errors and low detection fraction and provides an interesting test of the method. In Figure 8 I show the distribution of ξ and η, as well as the distribution of their measured values, for one of the simulated data sets. For this particular data set, there were 29 detections and 71 nondetections. As can be seen, the significant censoring and large measurement errors have effectively erased any visual evidence for a relationship between the two variables.

I estimated the posterior distribution of the regression parameters for this data set using the Gibbs sampler (cf. § 6.2.1) with $K = 2$ Gaussian functions. The posterior median of the regression line, as well as the 95% (2 σ) pointwise confidence intervals[1] on the regression line are shown in Figure 8. The posterior distributions for ρ, β, and σ are shown in Figure 9. As can be

[1] Technically, these are called "credibility intervals," as I am employing a Bayesian approach. These intervals contain 95% of the posterior probability. While the difference between confidence intervals and credibility intervals is not purely semantical, I do not find the difference to be significant within the context of my work, so I use the more familiar term "confidence interval."

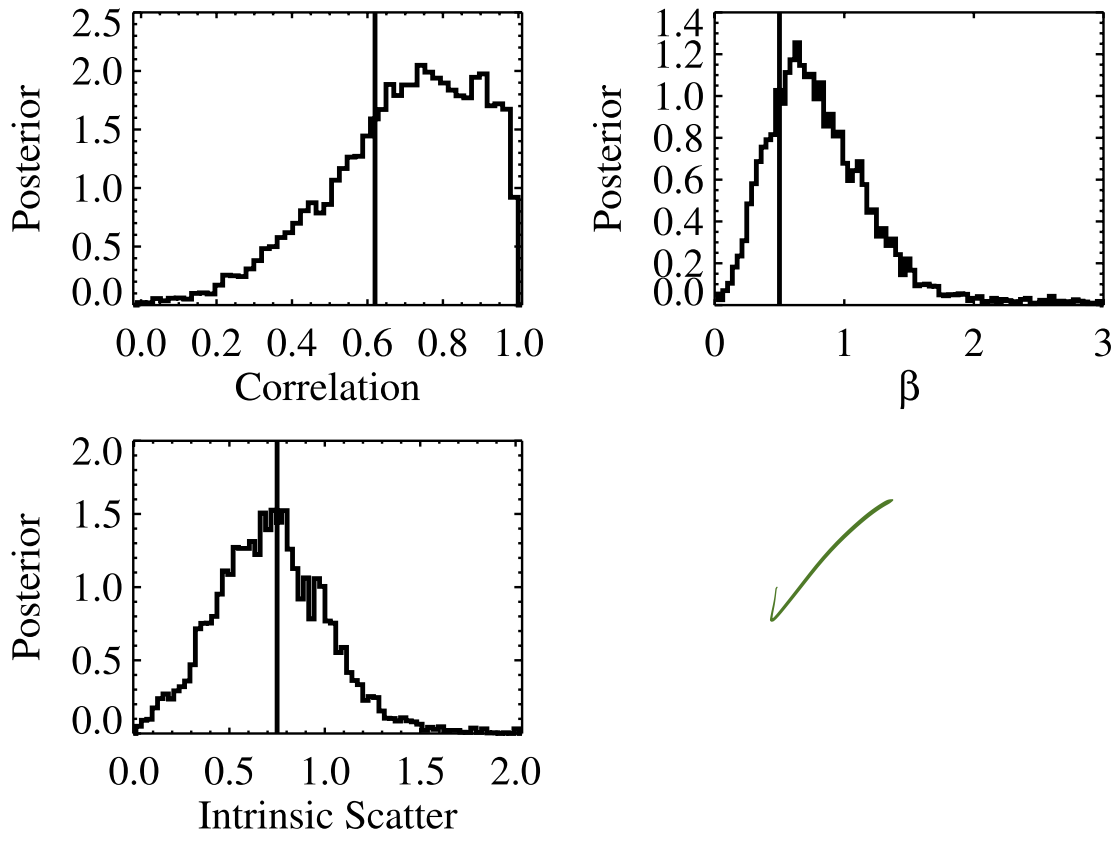

FIG. 9.—Same as Fig. 6, but for the censored data set shown in Fig. 8. The true values of the regression parameters are contained within the spread of the posteriors, implying that bounds on the regression parameters inferred from the posterior are trustworthy.

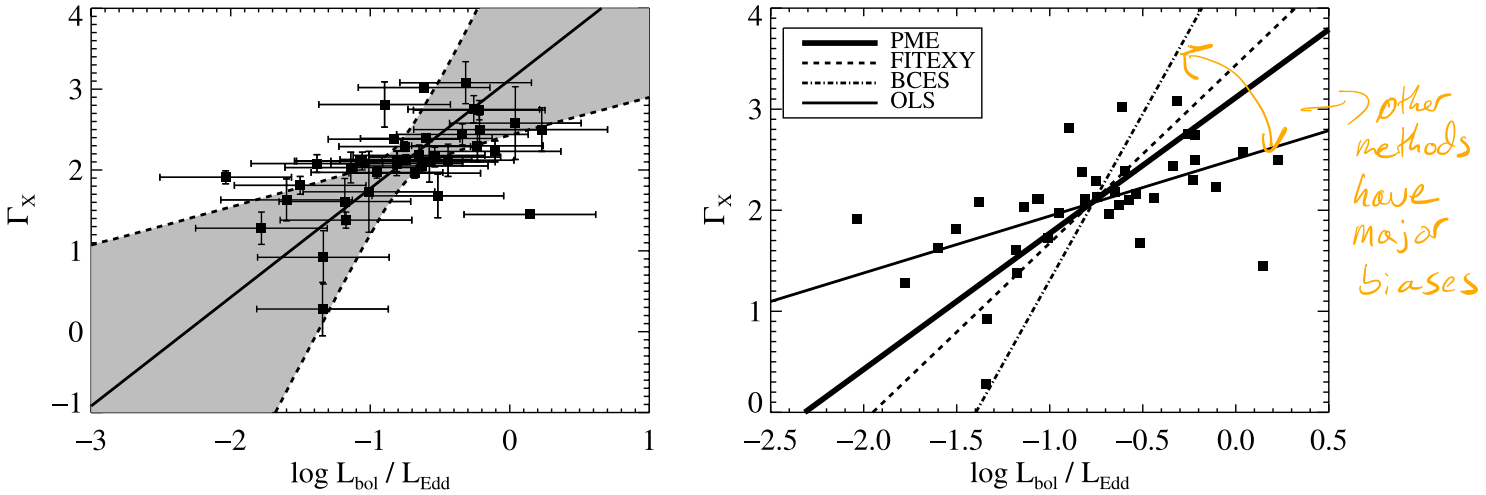

FIG. 10.—X-ray photon index Γ_X as a function of $\log L_{bol}/L_{Edd}$ for 39 $z \lesssim 0.8$ radio-quiet quasars. In both plots, the thick solid line shows the posterior median estimate (PME) of the regression line. In the left panel, the shaded region denotes the 95% (2 σ) pointwise confidence intervals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

seen, the true value of the parameters is contained within the 95% probability regions, although the uncertainty is large. For this particular data set, we can put limits on the value of the correlation coefficient as $0.2 \lesssim \rho \lesssim 1$ and the slope as $0 \lesssim \beta \lesssim 2.0$. For comparison, the usual MLE that ignores the measurement error (e.g., Isobe et al. 1986) concludes $\hat{\beta} = 0.229 \pm 0.077$. This estimate is biased and differs from the true value of β at a level of 3.5 σ.

The posterior constraints on the regression parameters are broad, reflecting our considerable uncertainty in the slope, but they are sufficient for finding a positive correlation between the two variables, ξ and η. Therefore, despite the high level of censoring and measurement error in this data set, we would still be able to conclude that η increases as ξ increases.

8. APPLICATION TO REAL ASTRONOMICAL DATA: DEPENDENCE OF Γ_X ON L_{bol}/L_{Edd} FOR RADIO-QUIET QUASARS

To further illustrate the effectiveness of the method, I apply it to a data set drawn from my work on investigating the X-ray properties of radio-quiet quasars (RQQs). Recent work has suggested a correlation between quasar X-ray spectral slope, $\alpha_X, f_\nu \propto \nu^{-\alpha_X}$, and quasar Eddington ratio, L_{bol}/L_{Edd} (e.g., Porquet et al. 2004; Piconcelli et al. 2005; Shemmer et al. 2006). In this section I apply the regression method to a sample of 39 $z < 0.83$ RQQs and confirm the Γ_X–L_{bol}/L_{Edd} correlation. Because the purpose of this section is to illustrate the use of this regression method on real astronomical data, I defer a more in-depth analysis to a future paper.

Estimation of the Eddington luminosity, $L_{Edd} \propto M_{BH}$, requires an estimate of the black hole mass, M_{BH}. Black hole virial masses may be estimated as $M_{BH} \propto Rv^2$, where R is the broad line region size and v is the velocity dispersion of the gas emitting the broad emission lines. A correlation has been found between the luminosity of a source and the size of its broad line region (the R-L relationship; e.g., Kaspi et al. 2005). One can then exploit this relationship and use the broad line FWHM as an estimate for v, obtaining virial mass estimates $\hat{M}_{BH} \propto L^\theta v^2$ (e.g., Wandel et al. 1999), where the exponent is $\theta \approx 0.5$ (e.g., Vestergaard & Peterson 2006). Unfortunately, the uncertainty on the broad line estimates of M_{BH} can be considerable, having a standard deviation of $\sigma_m \sim 0.4$ dex (e.g., McLure & Jarvis 2002; Vestergaard & Peterson 2006; Kelly & Bechtold 2007). For ease of comparison with previous work, I estimate M_{BH} using only the Hβ emission line. The logarithm of the virial mass estimates were calculated using the Hβ luminosity and FWHM according to the relationship given by Vestergaard & Peterson (2006).

My sample consists of a subset of the sample of Kelly et al. (2007). These sources have measurements of the X-ray photon index, $\Gamma_X = \alpha_X + 1$, obtained from *Chandra* observations and measurements of the optical/UV luminosity at 2500 Å, denoted as L_{2500}, obtained from Sloan Digital Sky Survey (SDSS) spectra. The Hβ profile was modeled as a sum of Gaussian functions and extracted from the SDSS spectra according to the procedure described in Kelly & Bechtold 2007. I estimated the Hβ FWHM and luminosity from the line profile fits.

I estimate the bolometric luminosity L_{bol} from the luminosity at 2500 Å, assuming a constant bolometric correction $L_{bol} = 5.6L_{2500}$ (Elvis et al. 1994). The standard deviation in this bolometric correction reported by Elvis et al. (1994) is 3.1, implying an uncertainty in $\log L_{bol}$ of $\sigma_{bol} \sim 0.25$ dex. Combining this with the $\sim$0.4 dex uncertainty on $\log M_{BH}$, the total "measurement error" on $\log L_{bol}/L_{Edd}$ becomes $\sigma_x \sim 0.47$ dex. The distribution of Γ_X as a function of $\log L_{bol}/L_{Edd}$ is shown in Figure 10. As can be seen, the measurement errors on both Γ_X and $\log L_{bol}/L_{Edd}$ are large and make a considerable contribution to the observed scatter in both variables, where $R_y \sim 0.1$ and $R_x \sim 0.8$. Therefore, we expect the measurement errors to have a significant effect on the correlation and regression analysis.

I performed the regression assuming the linear form $\Gamma_X = \alpha + \beta \log L_{bol}/L_{Edd}$ and modelling the intrinsic distribution of $\log L_{bol}/L_{Edd}$ using $K = 2$ Gaussian functions. Draws from the posterior were obtained using the Gibbs sampler. The marginal posterior distributions for β, σ, and the correlation between Γ_X and $\log L_{bol}/L_{Edd}$, ρ, are shown in Figure 11, and the posterior median and 95% (2 σ) pointwise intervals on the regression line are shown in Figure 10. The posterior median estimate of the parameters are $\hat{\alpha} = 3.12 \pm 0.41$ for the constant, $\hat{\beta} = 1.35 \pm 0.54$ for the slope, $\hat{\sigma} = 0.26 \pm 0.11$ for the intrinsic scatter about the regression line, $\hat{\mu}_\xi = -0.77 \pm 0.10$ for the mean of $\log L_{bol}/L_{Edd}$, and $\hat{\sigma}_\xi = 0.32 \pm 0.12$ dex for the dispersion in $\log L_{bol}/L_{Edd}$. Here, I have used a robust estimate of the posterior standard deviation as an

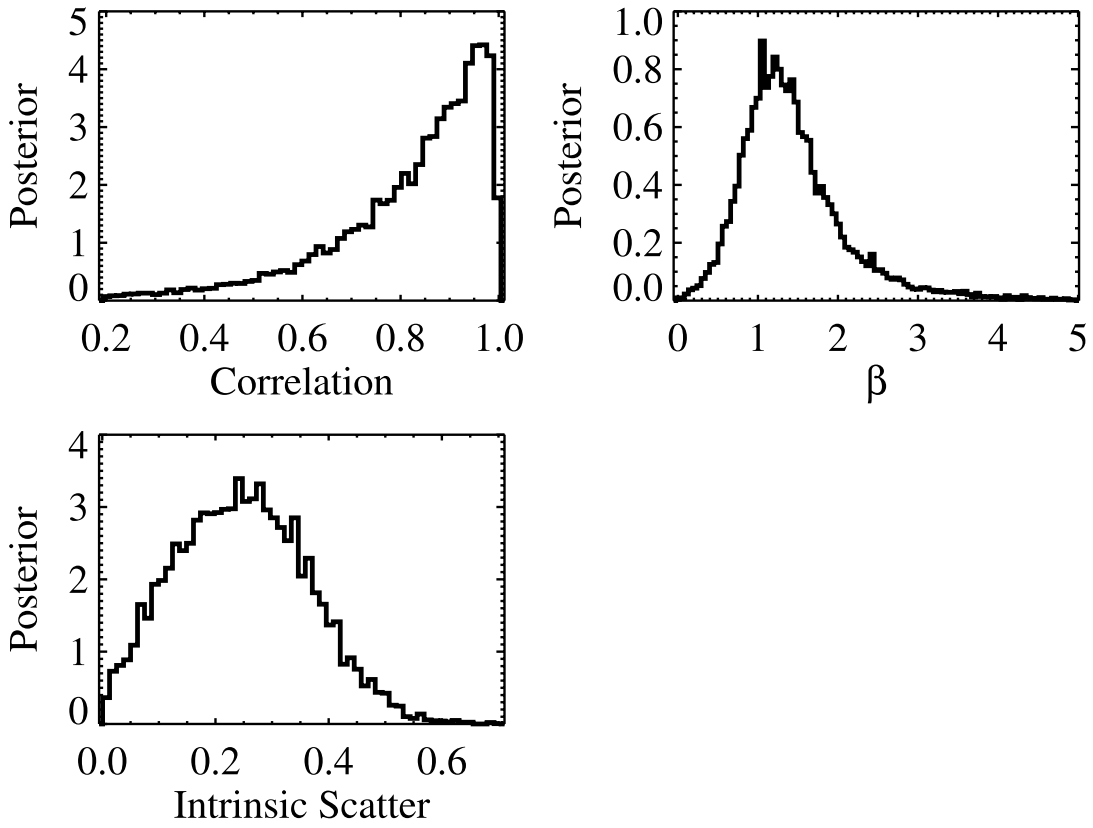

FIG. 11.—Same as Fig. 6, but for the $\Gamma_X - \log L_{bol}/L_{Edd}$ regression. Although the uncertainty on the slope and correlation are large, the bounds on them implied by the data are $0 \lesssim \beta \lesssim 3.5$ and $0.2 \lesssim \rho \lesssim 1.0$.

"error bar" on the parameters. These results imply that the observed scatter in $\log L_{bol}/L_{Edd}$ is dominated by measurement error, $\sigma_x/\tau \sim 1.5$, as expected from the large value of R_x.

For comparison, the BCES($Y|X$) estimate of the slope is $\hat{\beta}_{BCES} = 3.29 \pm 3.34$, the FITEXY estimate is $\hat{\beta}_{EXY} = 1.76 \pm 0.49$, and the OLS estimate is $\hat{\beta}_{OLS} = 0.56 \pm 0.14$; the standard error on $\hat{\beta}_{EXY}$ was estimated using bootstrapping. Figure 10 also compares the OLS, BCES, and FITEXY best-fit lines with the posterior median estimate. The 95% confidence region on the slope implied by the posterior draws is $0.46 < \beta < 3.44$, whereas the approximate 95% confidence region implied by the BCES, FITEXY, and OLS standard errors are $-3.26 < \beta < 9.84$, $0.80 < \beta < 2.72$, and $0.42 < \beta < 0.70$, respectively. The OLS and FITEXY estimates and the Bayesian approach give "statistically significant" evidence for a correlation between $\log L_{bol}/L_{Edd}$ and Γ_X; however, the BCES estimate is too variable to rule out the null hypothesis of no correlation. As noted before, the large measurement errors on $\log L_{bol}/L_{Edd}$ bias the OLS estimate of β toward shallower values and the FITEXY estimate of β toward steeper values. Because of this bias, confidence regions based on $\hat{\beta}_{OLS}$ and $\hat{\beta}_{EXY}$ are not valid, because they are not centered on the true value of β and, thus, do not contain the true value with the stated probability (e.g., 95%). On the other hand, confidence regions based on the BCES estimate are likely to be approximately valid; however, in this example the large measurement errors have caused $\hat{\beta}_{BCES}$ to be too variable to give meaningful constraints on the regression slope.

The BCES-type estimate of the intrinsic dispersion was $\hat{\sigma}_{BCES} = 0.32$, and the OLS estimate of the intrinsic dispersion was $\hat{\sigma}_{OLS} = 0.41$, where both were calculated in the same manner as in § 7.1. The FITEXY estimate of the intrinsic dispersion was $\hat{\sigma}_{EXY} = 0$, as $\chi^2_{EXY}/(n-2) < 1$. The BCES-type estimate of σ is similar to the Bayesian posterior median estimate, while $\hat{\sigma}_{OLS}$ overestimates the scatter compared to the Bayesian estimate by ≈58%. In contrast, the FITEXY estimator does not find any evidence for intrinsic scatter in the regression, which is inconsistent with the posterior distribution of σ.

From the posterior distribution, we can constrain the correlation between Γ_X and $\log L_{bol}/L_{Edd}$ to be $0.328 \lesssim \rho \lesssim 0.998$ with ≈95% probability, confirming the positive correlation between Γ_X and Eddington ratio seen previously. The posterior median estimate of the correlation is $\hat{\rho} = 0.87$, compared with an estimate of $\hat{r} = 0.54$ if one naively calculates the correlation directly from the measured data. The large measurement errors significantly attenuate the observed correlation, making the observed correlation between Γ_X and $\log L_{bol}/L_{Edd}$ appear weaker than if one does not correct for the measurement errors.

9. CONCLUSIONS

In this work I have derived a likelihood function for handling measurement errors in linear regression of astronomical data. Our probability model assumes that the measurement errors are Gaussian with zero mean and known variance, that the intrinsic scatter in the dependent variable about the regression line is Gaussian, and that the intrinsic distribution of the independent variables can be well approximated as a mixture of Gaussian functions. I extend this model to enable the inclusion of nondetections and describe how to incorporate the data selection process. A Gibbs sampler is described to enable simulating random draws from the posterior distribution.

I illustrated the effectiveness of the structural Gaussian mixture model using simulation. For the specific simulations performed, a MLE based on the Gaussian structural model performed better than the OLS, BCES($Y|X$), and FITEXY estimators, especially

when the measurement errors were large. In addition, our method also performed well when the measurement errors were large and the detection fraction was small, with the posterior distributions giving reasonable bounds on the regression parameters. These results were in spite of the fact that the intrinsic distribution of the independent variable was not a sum of Gaussian functions for the simulations, suggesting that approximating the distribution of the independent variable as a mixture of Gaussian functions does not lead to a significant bias in the results. Finally, I concluded by using the method to fit the radio-quiet quasar X-ray photon index as a function of $\log L_{\rm bol}/L_{\rm Edd}$, using a sample of 39 $z < 0.83$ sources. The posterior distribution for this data set constrained the slope to be $0 \lesssim \beta \lesssim 3.5$ and the linear correlation coefficient to be $0.2 \lesssim \rho \lesssim 1.0$, confirming the correlation between X-ray spectral slope and Eddington ratio seen by other authors.

Although I have focused on linear regression in this work, the approach that I have taken is quite general and can be applied to other applications. In particular, equations (11), (40), and (42) are derived under general conditions and are not limited to regression. In this work, I assume forms for the respective probability densities that are appropriate for linear regression; however, these equations provide a framework for constructing more general probability models of one's data, as in, for example, nonlinear fitting or estimation of luminosity functions.

IDL routines for constructing Markov chains for sampling from the posterior are publicly available from B. Kelly.

This work was supported in part by NSF grant AST 03-07384. The author would like to thank the referee for comments that contributed to the improvement of this paper and for providing some of the references to the statistics literature. The author would also like to thank Jill Bechtold, Eric Feigelson, and Aneta Siemiginowska for looking over and commenting on an early version of this paper.

REFERENCES

Aitken, M., & Rocci, R. 2002, Statistics and Computing, 12, 163
Akritas, M. G., & Bershady, M. A. 1996, ApJ, 470, 706
Akritas, M. G., & Siebert, J. 1996, MNRAS, 278, 919
Barker, D. R., & Diana, L. M. 1974, Am. J. Phys., 42, 224
Carroll, R. J., Roeder, K., & Wasserman, L. 1999, Biometrics, 55, 44
Carroll, R. J., Ruppert, D., & Stefanski, L. A. 1995, Measurement Error in Nonlinear Models (London: Chapman & Hall)
Chib, S., & Greenberg, E. 1995, Amer. Stat., 49, 327
Clutton-Brock, M. 1967, Technometrics, 9, 261
Davison, A. C., & Hinkley, D. V. 1997, Bootstrap Methods and Their Application (Cambridge: Cambridge Univ. Press)
Dellaportas, P., & Stephens, D. A. 1995, Biometrics, 51, 1085
Dempster, A., Laird, N., & Rubin, D. 1977, J. R. Statis. Soc. B., 39, 1
Efron, B. 1979, Ann. Statist., 7, 1
Elvis, M., et al. 1994, ApJS, 95, 1
Feigelson, E. D. 1992, in Statistical Challenges in Modern Astronomy, ed. E. Feigelson & G. Babu (New York: Springer), 221
Feigelson, E. D., & Nelson, P. I. 1985, ApJ, 293, 192
Fox, J. 1997, Applied Regression Analysis, Linear Models, and Related Methods (Thousand Oaks: Sage)
Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., & Carrol, R. J. 2004, Biometrics, 60, 172
Fuller, W. A. 1987, Measurement Error Models (New York: Wiley)
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2004, Bayesian Data Analysis (2nd ed.; Boca Raton: Chapman & Hall)
Gull, S. F. 1989, in Maximum Entropy and Bayesian Methods, ed. J. Skilling (Dordrecht: Kluwer), 511
Hastings, W. K. 1970, Biometrika, 57, 97
Huang, X., Stefanski, L. A., & Davidian, M. 2006, Biometrika, 93, 53
Isobe, T., Feigelson, E. D., & Nelson, P. I. 1986, ApJ, 306, 490
Kaspi, S., Maoz, D., Netzer, H., Peterson, B. M., Vestergaard, M., & Jannuzi, B. T. 2005, ApJ, 629, 61
Kelly, B. C., & Bechtold, J. 2007, ApJS, 168, 1
Kelly, B. C., Bechtold, J., Siemiginowska, A., Aldcroft, T., & Sobolewska, M. 2007, ApJ, 657, 116
Landy, S. D., & Szalay, A. S. 1992, ApJ, 391, 494
Little, R. J. A., & Rubin, D. B. 2002, Statistical Analysis with Missing Data (2nd ed.; Hoboken: Wiley)
Loredo, T. J. 1992, in Statistical Challenges in Modern Astronomy, ed. E. Feigelson & G. Babu (New York: Springer), 275
Marshall, H. L. 1992, in Statistical Challenges in Modern Astronomy, ed. E. Feigelson & G. Babu (New York: Springer), 247
McLure, R. J., & Jarvis, M. J. 2002, MNRAS, 337, 109
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Metropolis, N., & Ulam, S. 1949, J. Amer. Stat. Assoc., 44, 335
Müller, P., & Roeder, K. 1997, Biometrika, 84, 523
Piconcelli, E., Jimenez-Bailón, E., Guainazzi, M., Schartel, N., Rodríguez-Pascual, P. M., & Santos-Lleó, M. 2005, A&A, 432, 15
Porquet, D., Reeves, J. N., O'Brien, P., & Brinkmann, W. 2004, A&A, 422, 85
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes (2nd ed.; Cambridge: Cambridge Univ. Press)
Richardson, S., & Leblond, L. 1997, Statistics in Medicine, 16, 203
Richardson, S., Leblond, L., Jaussent, I., & Green, P. J. 2002, J. R. Statis. Soc. A, 165, 549
Ripley, B. D. 1987, Stochastic Simulation (New York: Wiley)
Roeder, K., & Wasserman, L. 1997, J. Amer. Stat. Assoc., 92, 894
Roy, S., & Banerjee, T. 2006, Ann. Inst. Statis. Math., 58, 153
Schafer, D. W. 1987, Biometrika, 74, 385
———. 2001, Biometrics, 57, 53
Scheines, R., Hoijtink, H., & Boomsma, A. 1999, Psychometrika, 64, 37
Schmitt, J. H. M. M. 1985, ApJ, 293, 178
Shemmer, O., Brandt, W. N., Netzer, H., Maiolino, R., & Kaspi, S. 2006, ApJ, 646, L29
Stapleton, D. C., & Young, D. J. 1984, Econometrica, 52, 737
Tremaine, S., et al. 2002, ApJ, 574, 740
Vestergaard, M., & Peterson, B. M. 2006, ApJ, 641, 689
Wandel, A., Peterson, B. M., & Malkan, M. A. 1999, ApJ, 526, 579
Weiner, B. J., et al. 2006, ApJ, 653, 1049
Weiss, A. A. 1993, J. Econometrics., 56, 169
Zellner, A. 1971, An Introduction to Bayesian Inference in Econometrics (New York: Wiley)