Electronic Journal of Statistics
Vol. 0 (2017)
ISSN: 1935-7524

A Note on Parameter Estimation for Misspecified Regression Models with Heteroskedastic Errors*

James P. Long†

Department of Statistics
3143 TAMU
College Station, TX 77843-3143
e-mail: jlong@stat.tamu.edu

Abstract: Misspecified models often provide useful information about the true data generating distribution. For example, if y is a non–linear function of x the least squares estimator $\widehat{\beta}$ is an estimate of β, the slope of the best linear approximation to the non–linear function. Motivated by problems in astronomy, we study how to incorporate observation measurement error variances into fitting parameters of misspecified models. Our asymptotic theory focuses on the particular case of linear regression where often weighted least squares procedures are used to account for heteroskedasticity. We find that when the response is a non–linear function of the independent variable, the standard procedure of weighting by the inverse of the observation variances can be counter–productive. In particular, ordinary least squares may have lower asymptotic variance. We construct an adaptive estimator which has lower asymptotic variance than either OLS or standard WLS. We demonstrate our theory in a small simulation and apply these ideas to the problem of estimating the period of a periodic function using a sinusoidal model.

MSC 2010 subject classifications: Primary 62J05; secondary 62F10.
Keywords and phrases: heteroskedasticity, model misspecification, approximate models, weighted least squares, sandwich estimators, astrostatistics.

Contents

*The author thanks the Editor and two reviewers for their constructive comments.
†Long's work was supported by a faculty startup grant from Texas A&M University.

1. Introduction

Misspecified models are common. In prediction problems, simple, misspecified models may be used instead of complex models with many parameters in order to avoid overfitting. In big data problems, true models may be computationally intractable, leading to model simplifications which induce some level of misspecification. In many scientific domains there exist sets of well established models with fast computer implementations. A practitioner with a particular data set may have to choose between using one of these models (even when none are exactly appropriate) and devising, testing and implementing a new model. Pressed for time, the practitioner may use an existing misspecified model. In this work we study how to fit a misspecified linear regression model with heteroskedastic measurement error. Problems involving heteroskedastic measurement error and misspecified models are common in astronomy. We discuss an example in Section 2.

Suppose $x_i \in \mathbb{R}^p \sim F_X$ independent across i and $\sigma_i \in \mathbb{R} \sim F_\sigma$ independent across i for $1 \leq i \leq n$. Suppose

$$y_i = f(x_i) + \sigma_i \epsilon_i$$

where $\epsilon_i \sim F_\epsilon$ with $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = 1 \, \forall i$, independent across i and independent of x_i and σ_i. Define

$$\beta \equiv \underset{\beta}{\text{argmin}} \, \mathbb{E}[(f(x) - x^T\beta)^2] = \mathbb{E}[xx^T]^{-1}\mathbb{E}[xf(x)].$$

The parameter β is the slope of the best fitting least squares line. The parameter β may be of interest in several situations. For example, β minimizes mean squared error in predicting y from x among all linear functions, ie $\beta = \underset{\beta}{\text{argmin}} \, \mathbb{E}[(y - x^T\beta)^2]$. Define $g(x) = f(x) - x^T\beta$. The function g is the non–linear component of f.

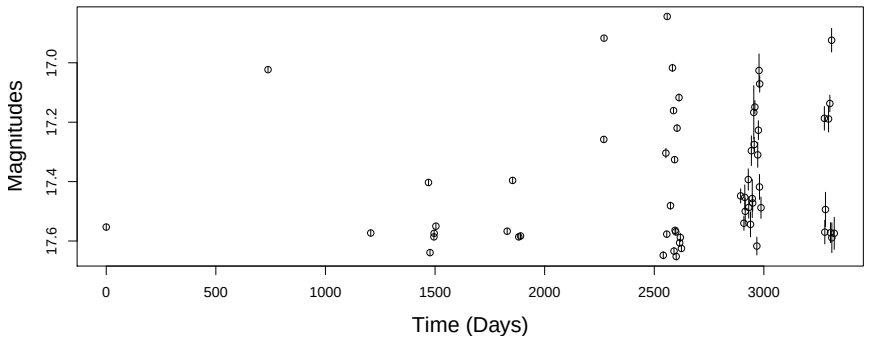
When the model is correctly specified (ie $g(x) \equiv 0$), weighted least squares (WLS) using the inverse of the observation variances as weights is asymptotically normal and has minimum asymptotic variance among all WLS estimators. In the case with model misspecification and x_i, σ_i independent, we show that WLS estimators remain asymptotically normal. However weighting by the inverse of the observation variances can result in a larger asymptotic variance than other weightings, including ordinary least squares. Using the asymptotic variance formula we determine an optimal weighting which has lower asymptotic variance than standard WLS (using the inverse of the observation variances as weights) and OLS. The optimal weighting function has the form $w(\sigma) = (\sigma^2 + \Delta)^{-1}$ where $\Delta \geq 0$ is a function of the degree of model misspecification and the design. We find adaptive estimators for w in the cases where the error variances are assumed known and where the error variances belong to one of M groups with group membership known. We also briefly consider the case where x_i and σ_i are dependent. In this setting the OLS estimator is consistent but weighted estimators are generally not consistent.

This work is organized as follows. In Section 2 we introduce a motivating problem from astronomy and offer some heuristic thinking about misspecified models and heteroskedasticity. For those readers primarily interested in the statistical theory, Section 2 can be skipped. In Section 3 we review some relevant literature and develop asymptotic results for the linear model. We present results for simulated data and the astronomy application in Section 4. We conclude in Section 5.

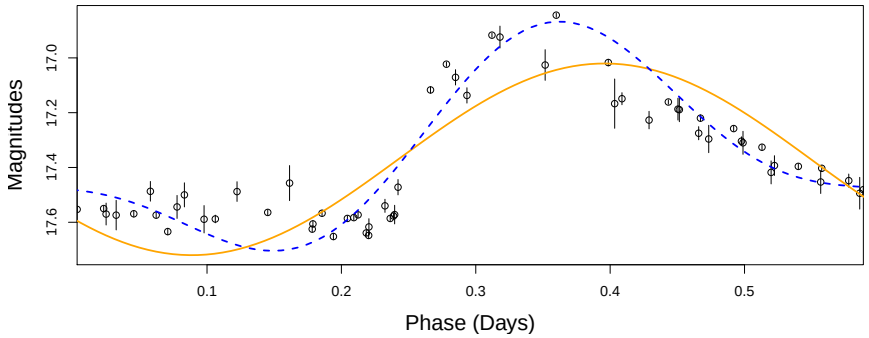
2. Misspecified Models and Heteroskedastic Error in Astronomy

Periodic variables are stars that vary in brightness periodically over time. Figure 1a shows the brightness of a single periodic variable star over time. This is known as the *light curve* of the star. Two sigma uncertainties are plotted as vertical bars around each point. Magnitude is inversely proportional to brightness, so lower magnitudes are plotted higher on the y–axis. This is a periodic variable so the changes in brightness over time are periodic. Using this data one may estimate a period for the star. When we plot the brightness measurements as time modulo period (Figure 1b), the pattern in brightness variation becomes clear. Periodic variables play an important role in several areas of astronomy including extra–galactic distance determination and estimation of the Hubble constant [26, 21]. Modern surveys, such as OGLE-III, have collected hundreds of thousands of periodic variable star light curves [28].

Accurate period estimation algorithms are necessary for creating the folded light curve (Figure 1b). A common procedure for determining the period is to perform maximum likelihood estimation using some parametric model for light curve variation. One popular model choice is a sinusoid with K harmonics. Let the data for a single periodic variable be $D = \{(t_i, y_i, \sigma_i)\}_{i=1}^n$ where y_i is the brightness at time t_i, measured with known uncertainty σ_i. Magnitude variation



(a) Unfolded Light Curve.



(b) Folded Light Curve.

Fig 1: (a) SDSS-III RR Lyrae light curve. (b) Folded light curve (x–axis is time modulo period) after estimating the period using the data in a).

is modeled as

$$y_i = \beta_0 + \sum_{k=1}^{K} a_k \sin(k\omega t_i + \phi_k) + \sigma_i \epsilon_i \tag{2.1}$$

where $\epsilon_i \sim N(0,1)$ independent across i. Here ω is the frequency, a_k is the amplitude of the k^{th} harmonic, and ϕ_k is the phase of the k^{th} harmonic. Let $a = (a_1, \ldots, a_K)$ and $\phi = (\phi_1, \ldots, \phi_K)$. Let Ω be a grid of possible frequencies. The maximum likelihood estimate for frequency is

$$\widehat{\omega} = \underset{\omega \in \Omega}{\operatorname{argmin}} \min_{a,\phi,\beta_0} \sum_{i=1}^{n} \left(\frac{y_i - \beta_0 - \sum_{k=1}^{K} a_k \sin(k\omega t_i + \phi_k)}{\sigma_i} \right)^2. \tag{2.2}$$

Generalized Lomb–Scargle (GLS) is equivalent to this estimator with $K = 1$ [32]. The analysis of variance periodogram in [23] uses this model with a fast algorithm for computing $\widehat{\omega}$.

We used estimator (2.2) with $K = 1, 2$ to determine the period of the light curve in Figure 1a. The estimates for period were essentially the same for both

$K = 1$ and $K = 2$ so in Figure 1b we folded the light curve using the $K = 1$ estimate. The solid orange line is the maximum likelihood fit for the $K = 1$ model (notice the sinusoidal shape). The blue dashed line is for the $K = 2$ model.

While the period estimates are accurate, both models are misspecified. In particular, note that the vertical lines around the brightness measurements are four standard deviations ($4\sigma_i$) in width. If the model is correct, we would expect about 95% of these intervals to contain the maximum likelihood fitted curves. For the $K = 1$ model, 10% of the intervals contain the fitted curve. For $K = 2$, 37% of observations contain the ML fitted curve. The source of model misspecification is the light curve shape which cannot be perfectly represented by a sinusoid with $K = 1, 2$ harmonics. The light curve has a long, slow decline and a sudden, sharp increase in brightness.

The parameter fits of misspecified models are estimates of an approximation. In the $K = 1$ case, the parameter fits are the orange line in Figure 1b and the approximation is the sinusoid which is closest to the true light curve shape. In many cases this approximation may be useful. For example the period of the approximation may match the period of the light curve.

When fitting a misspecified model with heteroskedastic measurement error, one should choose a weighting which ensures the estimator has small variance and thus is likely close to the approximation. The use of the inverse of the observation variances as weights (in Equation (2.3)) is motivated by maximum likelihood theory under the assumption that the model is correct. However as we show in Section 3 for the linear model, these weights are generally not optimal when there is model misspecification.

As a thought experiment, consider the case where one observation has extremely small variance and other observations have much larger variance. The maximum likelihood fitted curve for this data will be very close to the observation with small variance. However the best sinusoidal approximation to the true function at this point may not be particularly close to the true function. Thus using the inverse of observation variances as weights may overweight observations with small variance in the case of model misspecification. We make these ideas precise in Section 3.3.

The choice of weights is not critical for the light curve in Figure 1a because it is well sampled ($n > 50$), so the period is easy to determine. However in many other cases light curves are more poorly sampled ($n \approx 20$), in which case weighting may affect period estimation accuracy.

2.1. Sinusoidal Fit and Linear Models

Finding the best fitting sinusoid is closely related to fitting a linear model. Using the sine angle addition formula we can rewrite the maximum likelihood estimator from Equation (2.2) as

$$\underset{\omega\in\Omega}{\operatorname{argmin}} \min_{a,\phi,\beta_0} \sum_{i=1}^{n} \left(\frac{y_i - \sum_{k=1}^{K} (a_k \cos(\phi_k) \sin(k\omega t_i) + a_k \sin(\phi_k) \cos(k\omega t_i)) - \beta_0}{\sigma_i} \right)^2$$

The sum over i can be simplified by noting the linearity of the model and reparameterizing. Let $Y = (y_1, \ldots, y_n)^T$. Let $\beta_{k1} = a_k \cos(\phi_k)$ and $\beta_{k2} = a_k \sin(\phi_k)$. Define $\beta = (\beta_0, \beta_{11}, \beta_{12}, \ldots, \beta_{K1}, \beta_{K2})^T \in \mathbb{R}^{2K+1}$. Let Σ be a $n \times n$ diagonal matrix where $\Sigma_{ii} = \sigma_i^2$. Define

$$X(\omega) = \begin{pmatrix} 1 & \sin(\omega t_1) & \cos(\omega t_1) & \ldots & \sin(K\omega t_1) & \cos(K\omega t_1) \\ 1 & \sin(\omega t_2) & \cos(\omega t_2) & \ldots & \sin(K\omega t_2) & \cos(K\omega t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \sin(\omega t_n) & \cos(\omega t_n) & \ldots & \sin(K\omega t_n) & \cos(K\omega t_n) \end{pmatrix} \in \mathbb{R}^{n \times (2K+1)}.$$

We rewrite the ML estimator as

$$\widehat{\omega} = \underset{\omega \in \Omega}{\operatorname{argmin}} \min_{\beta} (Y - X(\omega)\beta)^T \Sigma^{-1} (Y - X(\omega)\beta)$$

Every frequency ω in the grid of frequencies Ω determines a design matrix $X(\omega)$. At a particular ω, the β which minimizes the objective function is the weighted least squares estimator

$$\widehat{\beta}(\omega) = (X(\omega)\Sigma^{-1}X(\omega))^{-1}X(\omega)^T\Sigma^{-1}Y. \tag{2.3}$$

The frequency estimate may then be written as

$$\widehat{\omega} = \underset{\omega \in \Omega}{\operatorname{argmin}} \ (Y - X(\omega)\widehat{\beta}(\omega))^T\Sigma^{-1}(Y - X(\omega)\widehat{\beta}(\omega)). \tag{2.4}$$

Thus estimating frequency involves performing a weighted least squares regression (Equation (2.3)) at every frequency in the grid Ω. The motivation for the procedure is maximum likelihood. As discussed earlier, in cases where the model is misspecified, there is no theoretical support for using Σ^{-1} as the weight matrix in either Equation (2.3) or (2.4).

3. Asymptotic Theory

3.1. Problem Setup and Related Literature

Let $X \in \mathbb{R}^{n \times p}$ be the matrix with row i equal to x_i^T. Let $Y = (y_1, \ldots, y_n)^T$. Let Σ be the diagonal matrix of observation variances such that $\Sigma_{ii} = \sigma_i^2$. Let $\widehat{W}$ be a diagonal positive definite matrix. The weighted least squares estimator is

$$\widehat{\beta}(\widehat{W}) = (X^T\widehat{W}X)^{-1}X^T\widehat{W}Y.$$

In this work we seek $\widehat{W}$ which minimize error in estimating $\beta = \mathbb{E}[xx^T]^{-1}\mathbb{E}[xf(x)]$.

There is a long history of studying estimators for misspecified models, often in the context of sandwich estimators for asymptotic variances. In [10], it was shown that when the true data generating distribution θ_t is not in the model,

the MLE converges to the distribution θ_0 in the model Θ which minimizes Kullback–Liebler divergence, ie

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \mathbb{E}_{\theta_t}\left[\frac{\log f_{\theta_t}(X)}{\log f_\theta(X)}\right].$$

The asymptotic variance has a "sandwich" form which is not the inverse of the information matrix. [30] and [31] studied this behavior in the context of the linear regression model and the OLS estimator, proposing consistent estimators of the asymptotic variance. See [18] and [15] for sandwich estimators with improved finite sample performance and [27] for recent work on sandwich estimators in a Bayesian context. [2] provides a summary of sandwich estimators and proposes a bootstrap estimator for the asymptotic variance. By specializing our asymptotic theory from the weighted to the unweighted case, we rederive some of these results. However our focus is different in that we find weightings for least squares which minimize asymptotic variance, rather than estimating the asymptotic variance of unweighted procedures.

Other work has focused on correcting model misspecification, often by modeling deviations from a parametric regression function with some non–parametric model. [1] studied model misspecification when response variances are known up to a constant due to repeated measurements, ie $Var(y_i) = \sigma^2/m_i$ where m_i is known. A Gaussian prior was placed on β and the non–linear component g was modeled as being drawn from a Gaussian process. See [13] for an example with homoskedastic errors in the context of computer simulations. See [6] for an example in astronomy with known heteroskedastic errors. Our focus here is different in that instead of correcting model misspecification we consider how weighting observations affects estimation of the linear component of f.

Heteroskedasticity in the partial linear model

$$y_i = x_i^T\beta + h(z_i) + \epsilon_i.$$

is studied in [17] and [16]. Here $\mathrm{Var}(\epsilon_i) = \xi(x_i, z_i)$ for some function ξ. The parameter h is some unknown function. The response y depends on the x covariates linearly and the z covariates nonlinearly. When h is estimated poorly, weighting by the inverse of the observation variances causes parameter estimates of β to be inconsistent. In contrast, ignoring observation variances leads to consistent estimates of β. Qualitatively these conclusion are similar to our own in that they caution against using weights in the standard way.

3.2. Asymptotic Results

Our asymptotic theory makes assumptions on the form of the weight matrix.

Assumptions 1 (Weight Matrix)**.** *Suppose $\widehat{W} \in \mathbb{R}^n \times \mathbb{R}^n$ is a positive definite diagonal matrix with elements*

$$\widehat{W}_{ii} = w(\sigma_i) + n^{-1/2}\delta_{nm_i}h(\sigma_i) + n^{-1}d(\sigma_i, \delta_n)$$

where $w(\sigma) > 0$, $\mathbb{E}[w(\sigma)^4] < \infty$, h *is a bounded function,* $m_i \in \{1, \ldots, M\}$ *is a discrete random variable independent of* x_i *and* ϵ_i, δ_{nm_i} *is* $O_P(1)$ *for all* $n \in \mathbb{Z}^+$ *and* m_i, *and* $d(\sigma, \delta_n)$ *is uniformly in* σ *bounded above by an* $O_P(1)$ *random variable (ie* $\sup_\sigma |d(\sigma, \delta_n)| < \delta_n'$ *where* δ_n' *is* $O_P(1)$*).*

These assumptions include both the ordinary least squares (OLS) estimator where $\widehat{W}_{ii} = w(\sigma_i) = 1$ and the standard weighted least squares estimator where $\widehat{W}_{ii} = w(\sigma_i) = \sigma_i^{-2}$ (assuming $\mathbb{E}[\sigma^{-8}] < \infty$). In both these cases $\delta_{nm_i} = 0$ and $d = 0$ for all n, m. These additional terms are used in Sections 3.5 and 3.6 to construct adaptive estimators for the known and unknown variance cases.

Assumptions 2 (Moment Conditions)**.** *Suppose* x *and* σ *are independent, the design* $\mathbb{E}[xx^T]$ *is full rank, and* $\mathbb{E}[x_j^4 x_k^4] < \infty$ *for all* $1 \leq j, k \leq p$. *Assume* $\mathbb{E}[g(x)^4] < \infty$, $\mathbb{E}[\sigma^4] < \infty$, $\mathbb{E}[\epsilon^4] < \infty$, *and the variances are bounded below by a positive constant* $\sigma_{min}^2 \equiv \inf\{\sigma^2 : F_\sigma(\sigma) > 0\} > 0$.

The major assumption here is independence between x and σ. We address dependence in Section 3.7.

Theorem 3.1. *Under Assumptions 1 and 2*

$$\sqrt{n}(\widehat{\beta}(\widehat{W}) - \beta) \xrightarrow{d} N(0, \nu(w))$$

where

$$\nu(w) = \frac{\mathbb{E}[w^2]\mathbb{E}[xx^T]^{-1}\mathbb{E}[g^2(x)xx^T]\mathbb{E}[xx^T]^{-1} + \mathbb{E}[\sigma^2 w^2]\mathbb{E}[xx^T]^{-1}}{\mathbb{E}[w]^2}. \tag{3.1}$$

See Section A.1 for a proof. If the response is linear ($g \equiv 0$) then the variance is

$$\nu(w) = \frac{\mathbb{E}[\sigma^2 w^2]}{\mathbb{E}[w]^2}\mathbb{E}[xx^T]^{-1}.$$

Setting $w(\sigma) = \sigma^{-2}$ we have $\frac{\mathbb{E}[\sigma^2 w^2]}{\mathbb{E}[w]^2} = (\mathbb{E}[\sigma^{-2}])^{-1}$. This is the standard weighted least squares estimator. This w can be shown to minimize the variance using the Cauchy Schwartz inequality. With $w(\sigma) = 1$, the asymptotic variance can be rewritten

$$\mathbb{E}[xx^T]^{-1}\mathbb{E}[(g^2(x) + \sigma^2)xx^T]\mathbb{E}[xx^T]^{-1}. \tag{3.2}$$

This is the sandwich form of the covariance for OLS derived in [30] and [31] (see [2], specifically Equations 1-3), valid even when σ and x are not independent.

3.3. OLS and Standard WLS

For notational simplicity define

$$B = \mathbb{E}[xx^T]^{-1}$$
$$A = B^T\mathbb{E}[g^2(x)xx^T]B.$$

The asymptotic variances for OLS ($\widehat{\beta}(I)$) and standard WLS ($\widehat{\beta}(\Sigma^{-1})$) are

$$\begin{aligned}\nu(I) &= A + \mathbb{E}[\sigma^2]B \\ \nu(\Sigma^{-1}) &= \frac{\mathbb{E}[\sigma^{-4}]}{\mathbb{E}[\sigma^{-2}]^2}A + \frac{1}{\mathbb{E}[\sigma^{-2}]}B.\end{aligned}$$

Each of these asymptotic variances is composed of the same two terms. The A term is caused by model misspecification while the B term is the standard asymptotic variance in the case of no model misspecification. The coefficient on A is larger for $W = \Sigma^{-1}$ because $\frac{\mathbb{E}[\sigma^{-4}]}{\mathbb{E}[\sigma^{-2}]^2} \geq 1$ by Jensen's Inequality. The coefficient on B is larger for $W = I$ because $\mathbb{E}[\sigma^2] \geq \frac{1}{\mathbb{E}[\sigma^{-2}]}$. The relative merits of OLS and standard WLS depend on the size of the coefficients and the precise values of A and B. However, qualitatively, OLS and standard WLS suffer from high asymptotic variance in opposite situations which depend on the distribution of the errors. To make matters concrete, consider error distributions of the form

$$\begin{aligned}P(\sigma = c^{-1}) &= \delta_1 \\ P(\sigma = 1) &= 1 - \delta_1 - \delta_2 \\ P(\sigma = c) &= \delta_2\end{aligned}$$

where δ_1, δ_2 are small non–negative numbers and $c > 1$ is large. Note that A and B do not depend on F_σ.

- $\delta_1 = 0, \delta_2 > 0$**:** In this situation the error standard deviation is usually 1 and occasionally some large value c. The result is large asymptotic variance for OLS. Since $\mathbb{E}[\sigma^2] > c^2\delta_2$,

  $$\nu(I) \succeq A + c^2\delta_2 B$$

  For large c this will be large. In contrast the coefficients on A and B for standard WLS can be bounded. For the coefficient on B we have $\mathbb{E}[\sigma^{-2}]^{-1} \leq (1 - \delta_2)^{-1}$. The coefficient on A with $c > 1$ is

  $$\frac{\mathbb{E}[\sigma^{-4}]}{\mathbb{E}[\sigma^{-2}]^2} = \frac{\delta_2 c^{-4} + (1 - \delta_2)}{\delta_2^2 c^{-4} + 2\delta_2 c^{-2}(1 - \delta_2) + (1 - \delta_2)^2} < \frac{1}{1 - \delta_2}.$$

  Therefore

  $$\nu(\Sigma^{-1}) \preceq (1 - \delta_2)^{-1}(A + B).$$

  In summary, standard WLS performs better than OLS when there are a small number of observations with large variance.
- $\delta_1 > 0, \delta_2 = 0$**:** In this situation the error standard deviation is usually 1 and occasionally some small value c^{-1}. For standard WLS with c large and δ_1 small, the coefficient for A is

  $$\frac{\mathbb{E}[\sigma^{-4}]}{\mathbb{E}[\sigma^{-2}]^2} = \frac{\delta_1 c^4 + (1 - \delta_1)}{\delta_1^2 c^4 + 2\delta_1 c^2(1 - \delta_1) + (1 - \delta_1)^2} \approx \frac{1}{\delta_1}.$$

Thus the asymptotic variance induced by model misspecification will be large for standard WLS. In contrast, we can bound the asymptotic variance above for OLS, independently of c and δ_1. Since $c > 1$, $\mathbb{E}[\sigma^2] < 1$ and

$$\nu(I) \preceq A + B.$$

The case where both δ_1 and δ_2 are non–zero presents problems for both OLS and standard WLS. For example if $\delta = \delta_1 = \delta_2$, both OLS and standard WLS can be made to have large asymptotic variance by setting δ small and c large. In the following section we construct an adaptive weighting which improves upon both OLS and standard WLS.

3.4. Improving on OLS and Standard WLS

Let Γ be a linear function from the set of $p \times p$ matrices to $\mathbb{R}$ such that $\Gamma(C) > 0$ whenever C is positive definite. We seek some weighting $w = w(\sigma)$ for which $\Gamma(\nu(w))$ (recall that ν is the asymptotic variance) is lower than OLS and standard WLS. Natural choices for Γ include the trace (minimize the sum of variances of the parameter estimates) and the $\Gamma(C) = C_{jj}$ (minimize the variance of one of the parameter estimates).

Theorem 3.2. *Under Assumptions 1 and 2, every function in the set*

$$\underset{w(\sigma)}{\operatorname{argmin}}\ \Gamma(\nu(w))$$

is proportional to

$$w_{min}(\sigma) = (\sigma^2 + \Gamma(A)\Gamma(B)^{-1})^{-1} \tag{3.3}$$

with probability 1.

Section A.2 contains a proof. The proportionality is due to the fact that the estimator is invariant to multiplicative scaling of the weights.

Corollary 3.1. *Under Assumptions 2,*

$$\Gamma(\nu(w_{min})) \leq \min(\Gamma(\nu(I)), \Gamma(\nu(\Sigma^{-1})))$$

with strict inequality if $\mathbb{E}[g^2(x)xx^T]$ is positive definite and the distribution of σ is not a point mass.

A proof is contained in Section A.3. Thus if we can construct a weight matrix $\widehat{W}$ which satisfies Assumptions 1 with $w(\sigma) = w_{min}(\sigma)$, then by the preceding theorem the associated weighted estimator will have lower asymptotic variance then either OLS or standard WLS. We now construct such a weighting in the case of known and unknown error variances.

3.5. Known Error Variances

With the σ_i known we only need to estimate A and B in w_{min} in Equation (3.3). Let $\Delta = \Gamma(A)\Gamma(B)^{-1}$. Let

$$\widehat{B} = \left(\frac{1}{n}X^T X\right)^{-1}.$$

Let $\widehat{\beta}(\widehat{W})$ be a root n consistent estimator of β (eg $\widehat{W} = I$ is root n consistent by Theorem 3.1) and let

$$\widehat{g}(x_i)^2 = (y_i - x_i^T\widehat{\beta}(\widehat{W}))^2 - \sigma_i^2.$$

Let

$$\widehat{A} = \widehat{B}^T\left(\sum \sigma_i^{-4}\right)^{-1}\left(\sum x_i x_i^T \widehat{g}(x_i)^2\sigma_i^{-4}\right)\widehat{B}.$$

Then we have

$$\widehat{\Delta} = \max(\Gamma(\widehat{A})\Gamma(\widehat{B})^{-1}, 0). \tag{3.4}$$

The estimated optimal weighting matrix is the diagonal matrix $\widehat{W}_{min}$ with diagonal elements

$$\widehat{W}_{min,ii} = \frac{1}{\sigma_i^2 + \widehat{\Delta}}. \tag{3.5}$$

A few notes on this estimator:

- The term $x_i x_i^T\widehat{g}(x_i)^2$ is an estimate of $x_i x_i^T g(x_i)^2$. These estimates are weighted by σ_i^{-4}. The term $(\sum \sigma_i^{-4})^{-1}$ normalizes the weights. This weighting is motivated by the fact that

$$\begin{aligned} x_i x_i^T\widehat{g}(x_i)^2 &= x_i x_i^T((y_i - x_i^T\widehat{\beta}(\widehat{W}))^2 - \sigma_i^2) \\ &= x_i x_i^T((y_i - x_i^T\beta)^2 - \sigma_i^2) + O(n^{-1/2}) \\ &= x_i x_i^T((g(x_i) + \sigma_i\epsilon_i)^2 - \sigma_i^2) + O(n^{-1/2}). \end{aligned}$$

  Analysis of the first order term shows

$$\mathbb{E}[x_i x_i^T((g(x_i) + \sigma_i\epsilon_i)^2 - \sigma_i^2)|x_i, \sigma_i] = x_i x_i^T g^2(x_i)$$

  and

$$\begin{aligned} &\mathrm{Var}(x_i x_i^T((g(x_i) + \sigma_i\epsilon_i)^2 - \sigma_i^2)|x_i, \sigma_i)_{jk} \\ &= x_{ij}^2 x_{ik}^2(\sigma_i^4(\mathbb{E}[\epsilon^4] - 1) + 4g(x_i)^2\sigma_i^2 + 4g(x_i)\sigma_i^3\mathbb{E}[\epsilon^3]) \end{aligned}$$

  Thus by weighting the estimates by σ_i^{-4}, we can somewhat account for the different variances. Unfortunately since the variance depends on g, $\mathbb{E}[\epsilon^3]$, and $\mathbb{E}[\epsilon^4]$ which are unknown, it is not possible to weight by exactly the inverse of the variances. Other weightings are possible and in general adaptivity will hold.

- Since A and B are positive semi–definite, $\Gamma(A)\Gamma(B)^{-1} \geq 0$. Thus for estimating Δ, we use the maximum of a plug–in estimator and 0 (Equation (3.4)).

Theorem 3.3. *Under Assumptions 2, $\widehat{W}_{min}$ from Equation (3.5) satisfies Assumptions 1 with $w(\sigma) = w_{min}(\sigma)$.*

See Section A.4 for a proof. Theorem 3.3 shows it is possible to construct better estimators than both OLS and standard WLS. In practice, it may be best to iteratively update estimates of $\widehat{W}_{min}$ starting with a known root n consistent estimator such as $\widehat{W} = I$. We take this approach in our numerical simulations in Section 4.1.

For the purposes of making confidence regions we need estimators of the asymptotic variance. Above we developed consistent estimators for A and B. We take a plug–in approach to estimating the asymptotic variance for a particular weighting W. Specifically

$$\widehat{\nu}_1(\widehat{W}) = \frac{n(1^T\widehat{W}^2 1)\widehat{A} + n(1^T\widehat{W}\Sigma\widehat{W}1)\widehat{B}}{(1^T\widehat{W}1)^2}. \tag{3.6}$$

We also define the oracle $\widehat{\nu}_{OR}(\widehat{W})$ which is the same as $\widehat{\nu}_1$ but uses A and B rather than $\widehat{A}$ and $\widehat{B}$. While $\widehat{\nu}_{OR}$ cannot be used in practice, it is useful for evaluating the performance of $\widehat{\nu}_1$ in simulations.

Finally suppose the error variance is known up to a constant, i.e. $\sigma_i^2 = k\tau_i^2$ where τ_i^2 is known but k and σ_i^2 are unknown. In the case without model misspecification, one can simply use weights τ_i^{-2} since the weighted estimator is invariant up to rescaling of the weights. The situation is more complicated when model misspecification is present. Simulations and informal mathematical derivations (not included in this work) suggest that replacing the σ_i with τ_i in Equation (3.5) results in weights that are suboptimal. In particular, when $k > 1$ (underestimated errors), the resulting weights are closer to OLS than optimal while if $k < 1$ (overestimated errors), the resulting weights are closer to standard WLS than optimal.

3.6. Unknown Error Variances

Suppose for observation i we observe $m_i \in \{1, \ldots, M\}$, the group membership of observation i. Observations in group m have the same (unknown) variance $\sigma_m^2 > 0$. See [8], [5], and [9] for work on grouped error models in the case where the response is linear.

The m_i are assumed independent of x_i and ϵ_i, with probability mass function f_m (supported on $1, \ldots, M$). While the σ_m for $m = 1, \ldots, M$ are fixed unknown parameters, the probability mass function f_m induces the probability distribution function F_σ on σ. So we can define

$$\mathbb{E}[h(\sigma)] = \sum_{m=1}^{M} h(\sigma_m) f_m(m)$$

for any function h.

Theorem 3.1 shows that even if the σ_m were known, standard weighted least squares is not generally optimal for estimating β in this model. It is not possible to estimate w_{min} as proposed in Section 3.5 because that method requires knowledge of σ_m. However we can re–express the optimal weight function as

$$\begin{aligned} w_{min}(m) &= \frac{1}{\sigma_m^2 + \frac{\Gamma(B^T\mathbb{E}[g^2(x)xx^T]B)}{\Gamma(B)}} \\ &= \frac{\Gamma(B)}{\Gamma(B^T\mathbb{E}[(g^2(x)+\sigma_m^2)xx^T]B)} \\ &= \frac{\Gamma(B)}{\Gamma(B^TC_mB)} \end{aligned}$$

where the last equality defines C_m. Note that σ_m is a fixed unknown parameter, not a random variable. One can estimate B with $\widehat{B} = (n^{-1}X^TX)^{-1}$ and C_m with

$$\widehat{C}_m = \frac{1}{\sum_{i=1}^n \mathbb{1}_{m_i=m}} \sum_{i=1}^n (y_i - x_i^T\widehat{\beta}(\widehat{W}))^2 x_i x_i^T \mathbb{1}_{m_i=m}$$

where $\widehat{\beta}(\widehat{W})$ is a root n consistent estimator of β (for example $\widehat{W} = I$ suffices by Theorem 3.1). The estimated weight matrix $\widehat{W}_{min}$ is diagonal with

$$\widehat{W}_{min,ii} = \frac{\Gamma(\widehat{B})}{\Gamma(\widehat{B}^T\widehat{C}_{m_i}\widehat{B})}. \tag{3.7}$$

Theorem 3.4. *Under Assumptions 2, $\widehat{W}_{min}$ from Equation (3.7) satisfies Assumptions 1 with $w(\sigma) = w_{min}(m)$.*

See Section A.5 for a proof. Thus in the case of unknown errors it is possible to construct an estimator which outperforms standard WLS and OLS. As is the case with known errors, one can iteratively update $\widehat{W}_{min}$, starting with some (possibly inefficient) root n consistent estimate of β.

For estimating the asymptotic variance we cannot use Equation (3.6) because that method required an estimate of A, a quantity for which we do not have an estimate in the unknown error variance setting. Instead note that the asymptotic variance of Equation (3.1) may be rewritten

$$\nu(W) = \frac{B\mathbb{E}[(g^2(x)+\sigma^2)w^2xx^T]B}{\mathbb{E}[w]^2} = \frac{B\mathbb{E}[(y-x^T\beta)^2w^2xx^T]B}{\mathbb{E}[w]^2}.$$

Thus a natural estimator for the asymptotic variance is

$$\widehat{\nu}_2(\widehat{W}) = \frac{n\widehat{B}\left(\sum_{i=1}^n (y_i - x_i^T\widehat{\beta}(\widehat{W}))^2\widehat{W}_{ii}^2 x_i x_i^T\right)\widehat{B}}{(1^T\widehat{W}1)^2}. \tag{3.8}$$

3.7. Dependent Errors

Suppose one drops the independence assumption between x and σ. This will be the case whenever the error variance is a function of x, a common assumption in the heteroskedasticity literature [3, 4, 12]. We require the weight matrix W to be diagonal positive definite with diagonal elements $W_{ii} = w(\sigma_i)$, some function of the error variance. The estimator for β is

$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y.$$

Recalling we write w for $w(\sigma)$, we have the following result.

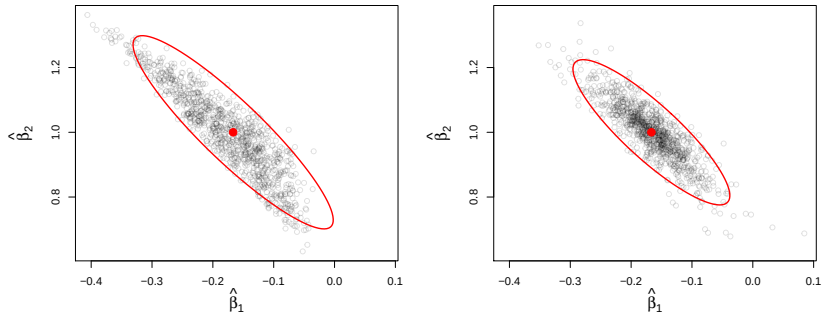
Theorem 3.5. *Assuming* $\mathbb{E}[xx^T w]$, $\mathbb{E}[wxf(x)]$, *and* $\mathbb{E}[xw\sigma]$ *exist and* $\mathbb{E}[xx^T]$ *is positive definite,*

$$\widehat{\beta}(W) \to_{a.s.} \mathbb{E}[xx^T w]^{-1}\mathbb{E}[wxf(x)]. \tag{3.9}$$

See Section A.6 for a proof. If x and σ are independent then the r.h.s is $\mathbb{E}[xx^T]^{-1}\mathbb{E}[xf(x)]$ and the estimator is consistent (as demonstrated by Theorem 3.1). Interestingly the estimator is also consistent if one lets $w(\sigma) = 1$ (OLS), regardless of the dependence structure between x and σ. However weighted estimators will not generally be consistent (including standard WLS). This observation suggests the OLS estimator may be preferred in the case of dependent errors. We show an example of this situation in the simulations of Section 4.1.

4. Numerical Experiments

4.1. Simulation



(a) $\widehat{W} = \Sigma^{-1}$ (b) $\widehat{W} = I$

(c) $\widehat{W} = (\Sigma + \widehat{\Delta})^{-1}$ (d) $\widehat{W}_{min,ii} = \frac{\Gamma(\widehat{B})}{\Gamma(\widehat{B}^T \widehat{C}_{m_i} \widehat{B})}$

Fig 2: Parameter estimates using (a) standard WLS (b) OLS (c) estimated weights assuming the σ_i are known (d) estimated weights using only the group membership of the variances. The red ellipses are the asymptotic variances of the various methods.

	WLS	OLS	$(\Sigma + \widehat{\Delta})^{-1}$	$\frac{\Gamma(\widehat{B})}{\Gamma(\widehat{B}^T \widehat{C}_{m_i} \widehat{B})}$
$\widehat{\nu}_1$	0.536	0.945	0.807	—
$\widehat{\nu}_2$	0.393	0.96	0.843	0.759
$\widehat{\nu}_{OR}$	0.925	0.945	0.956	—

TABLE 1
Fraction of times β is in 95% confidence region.

We conduct a small simulation study to demonstrate some of the ideas pre-

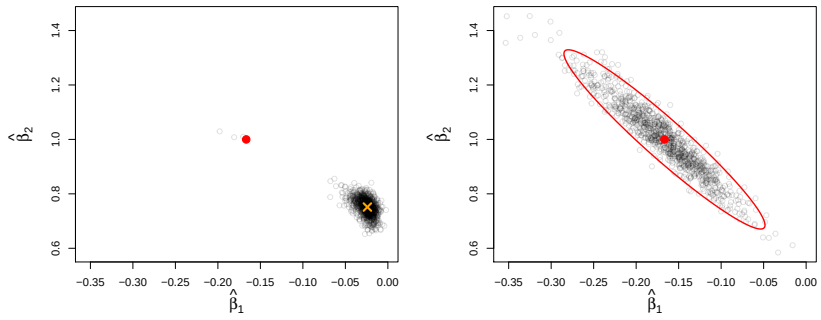
sented in the last section.[1] Consider modeling the function $f(x) = x^2$ using linear regression with an intercept term. Let $x \sim Unif(0, 1)$. The best linear approximation to f is $\beta_1 + \beta_2 x$ where $\beta_1 = -1/6$ and $\beta_2 = 1$. We first suppose σ is drawn independently from x from a discrete probability distribution such that $P(\sigma = 0.01) = P(\sigma = 1) = 0.05$ and $P(\sigma = 0.1) = 0.9$. Since σ has support on a finite set of values, we can consider the cases where σ_i is known (Section 3.5) and where only the group m_i of observation i is known (Section 3.6). We let Γ be the trace of the matrix.

We generate samples of size $n = 100$, $N = 1000$ times and make scatterplots of the parameter estimates using weights $W = \Sigma^{-1}$ (standard WLS), $W = I$ (OLS), $\widehat{W}_{min} = (\Sigma + \widehat{\Delta})^{-1}$, and $\widehat{W}_{min,ii} = \frac{\Gamma(\widehat{B})}{\Gamma(\widehat{B}^T \widehat{C}_{m_i} \widehat{B})}$. The OLS estimator does not require any knowledge about the σ_i. The fourth estimator uses only the group m_i of observation i. For the two adaptive estimators, we use $\widehat{\beta}(I)$ as an initial root n consistent estimator of β and iterate twice to obtain the weights.

Results are shown in Figure 2. The red ellipses are the asymptotic variances. The results show that OLS outperforms standard WLS. Estimating the optimal weighting with or without knowledge of the variances outperforms both OLS and standard WLS. Exact knowledge of the weights (c) somewhat outperforms only knowing the group membership of the variances (d).

We construct 95% confidence regions using estimates of the asymptotic variance and determine the fraction of times (out of the N simulations) that the true parameters are in the confidence regions. Recall that in Section 3.5 we proposed $\widehat{\nu}_1$ (Equation (3.6)) as well as the oracle $\widehat{\nu}_{OR}$ for estimating the asymptotic variance when the error variances are known. In Section 3.6 we proposed $\widehat{\nu}_2$ (Equation (3.8)) when the error variances are unknown. The estimator $\widehat{\nu}_2$ can also be used when the error variances are known. We use all three of these methods for constructing confidence regions for standard WLS, OLS, and $\widehat{W} = (\Sigma + \widehat{\Delta})^{-1}$. For $\widehat{W}_{ii} = \frac{\Gamma(\widehat{B})}{\Gamma(\widehat{B}^T \widehat{C}_{m_i} \widehat{B})}$ we use only $\widehat{\nu}_2$ because $\widehat{\nu}_1$ requires knowledge of Σ. Table 1 contains the results. While for OLS the nominal coverage probability is approximately attained, the other methods are anti–conservative for $\widehat{\nu}_1$ and $\widehat{\nu}_2$. Estimates for WLS are especially poor. The performance of the oracle is rather good, suggesting that the problem lies in estimating A and B.

[1] Code to reproduce the work in this section can be accessed at http://stat.tamu.edu/~jlong/hetero.zip or by contacting the author.



(a) $\widehat{W} = \Sigma^{-1}$ (b) $\widehat{W} = I$

Fig 3: Parameter estimates using (a) standard WLS and (b) OLS when there is dependence between x and σ. We see that WLS is no longer consistent. The red point in each plot is the true parameter values. The orange $\times$ in the left plot is the value to which standard WLS is converging (r.h.s. of Equation (3.9)). The red ellipse is the OLS sandwich asymptotic variance for the dependent case, Equation (3.2).

To illustrate the importance of the σ, x independence assumption, we now consider the case where σ is a function of x. Specifically,

$$\sigma = \begin{cases} 0.01 & : x < 0.05 \\ 0.1 & : 0.05 \leq x \leq 0.95 \\ 1 & : x > 0.95 \end{cases}$$

All other parameters in the simulation are the same as before. Note that the marginal distribution of σ is the same as the first simulation. We know from Section 3.7 that weighted estimators may no longer be consistent. In Figure 3 we show a scatter plot of parameter estimates using standard WLS and OLS. We see that the WLS estimator has low variance but is highly biased. The OLS estimator is strongly preferred.

4.2. Analysis of Astronomy Data

[25] identified 483 RR Lyrae periodic variable stars in Stripe 82 of the Sloan Digital Sky Survey III. We obtained 450 of these light curves from a publicly available data base [11].[2] Figure 1a shows one of these light curves. These light curves are well observed ($n > 50$), so it is fairly easy to estimate periods. For example, [25] used a method based on the Supersmoother algorithm of [7]. However there is interest in astronomy in developing period estimation algorithms that work well on poorly sampled light curves [29, 19, 14, 24]. Well sampled

[2]We use only the g–band data for determining periods.

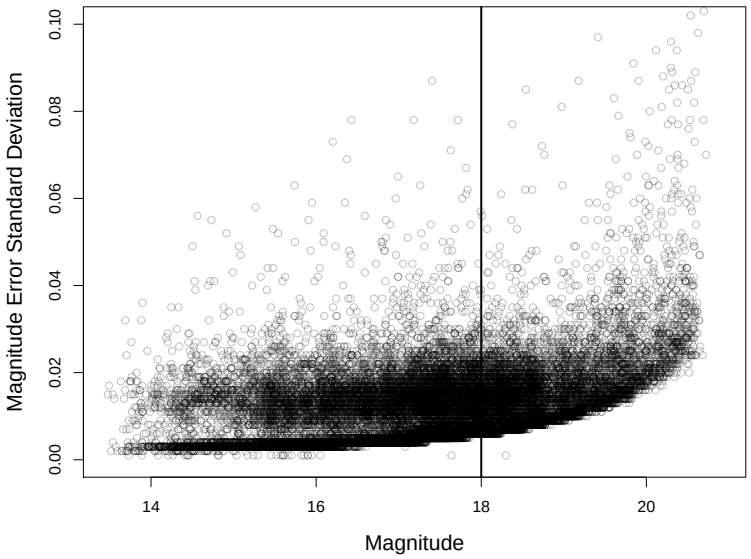


Fig 4: Magnitude error versus magnitude scatterplot. As magnitude increases (observation is less bright), the uncertainty rises. We use only stars where all photometric measurements are less than 18 magnitudes. In this region magnitude error and magnitude are approximately independent.

light curves offer an opportunity to test period estimation algorithms because ground truth is known and they can be artificially downsampled to create realistic simulations of poorly sampled light curves.

As discussed in Section 2, each light curve can be represented as $\{(t_i, y_i, \sigma_i)\}_{i=1}^n$ where t_i is the time of the y_i brightness measurement made with uncertainty σ_i. In Figure 4 we plot magnitude error (σ_i) against magnitude (y_i) for all observations of all 450 light curves. For higher magnitudes (less bright observations), the observation uncertainty is larger. In an attempt to ensure independence between σ and x assumed by our asymptotic theory, we use only the bright stars in which all magnitudes are below 18 (left of the vertical black line in Figure 4). In this region, magnitude and magnitude error are approximately independent. This reduces the sample to 238 stars. We also ran our methods on the larger set of stars. Qualitatively, the results which follow are similar.

In order to simulate challenging period recovery settings, we downsample each of these light curves to have $n = 10, 20, 30, 40$. We estimate periods using sinusoidal models with $K = 1, 2, 3$ harmonics. For each model we consider three methods for incorporating the error variances. In the first two methods, we weight by the the inverse of the observations variances (Σ^{-1}) as suggested by maximum likelihood for correctly specified models and the identity matrix (I). Since this is not a linear model, it is not possible to directly use the weighting

idea proposed in Section 3.5. We propose a modification for the light curve scenario. We first fit the model using identity weights and determine a best fit period. We then determine the optimal weighting at this period following the procedure of Section 3.5. Recall from Section 2 that at a fixed period, the sinusoidal models are linear. Using the new weights, we then refit the model and estimate the period. A period estimate is considered correct if it is within 1% of the true value.

	$K=1$			$K=2$			$K=3$		
	Σ^{-1}	I	Δ	Σ^{-1}	I	Δ	Σ^{-1}	I	Δ
10	0.09	0.16	0.15	0.13	0.11	0.11	0.03	0.03	0.03
20	0.46	0.58	0.59	0.63	0.68	0.69	0.69	0.77	0.77
30	0.64	0.78	0.79	0.71	0.82	0.83	0.82	0.86	0.85
40	0.75	0.79	0.79	0.80	0.85	0.85	0.87	0.92	0.92

TABLE 2

Fraction of periods estimated correctly using different weightings for models with $K = 1, 2, 3$ harmonics. Ignoring the observation uncertainties (I) in the fitting is superior to using them (Σ^{-1}). The strategy for determining an optimal weight function (Δ) does not provide much improvement over ignoring the weights. More complex models ($K = 3$) perform worse than simple models ($K = 1$) when there is limited data ($n = 10$), but better when the functions are better sampled ($n = 40$). The standard errors on these accuracies is no larger than $\sqrt{0.5(1-0.5)/238} \approx 0.032$.

The fraction of periods estimated correctly are contained in Table 2. In nearly all cases ignoring observation uncertainties (I) outperforms using the inverse of the observation variances as weights (Σ^{-1}). The improvement is greatest for the $K = 1$ model and least for the $K = 3$ model, possibly due to the decreasing model misspecification as the number of harmonics increases. The very poor performance of the $K = 3$ models with 10 magnitude measurements is due to overfitting. With $K = 3$, there are 8 parameters which is too complex a model for 10 observations. Optimizing the observation weights does not appear to improve performance over not using weights. This is potentially due to the fact that the model is highly misspecified (see Figure 1b).

5. Discussion

5.1. Other Problems in Astronomy

Heteroskedastic measurement error is ubiquitous in astronomy problems. In many cases some degree of model misspecification is present. In this work, we focused on the problem of estimating periods of light curves. Other problems include:

- [22] observe the brightness of galaxies through several photometric filters. Variances on the brightness measurements are heteroskedastic. The brightness measurements for each galaxy are matched to a set of templates. Assuming a normal measurement error model, maximum likelihood would suggest weighting the difference between observed brightness and template brightness by the inverse of the observation variance. In personal

communication, [22] stated that galaxy templates contained some level of misspecification. [22] addressed this issue by inflating observation variances, using weights of $(\sigma^2 + \Delta)^{-1}$ instead of σ^{-2}. The choice of $\Delta > 0$ was based on qualitative analysis of model fits. Section 3.3 provides a theoretical justification for this practice.

- [20] models spectra of galaxies as linear combinations of simple stellar populations (SSP) and non–linear distortions. While parameters which define an SSP are continuous, a discrete set of SSPs are selected as prototypes and the galaxies are modeled as linear combinations of the prototypes. This is done for computational efficiency and to avoid overfitting. However prototype selection introduces some degree of model misspecification as the prototypes may not be able to perfectly reconstruct all galaxy spectra. Galaxy spectra are observed with heteroskedastic measurement error and the inverse of the observation variances are used as weights when fitting the model (see Equation 2.2 in [20]).

5.2. Conclusions

We have shown that WLS estimators can perform poorly when the response is not a linear function of the predictors because observations with small variance have too much influence on the fit. In the misspecified model setting, OLS suffers from the usual problem that observations with large variance induce large asymptotic variance in the parameter estimates. For cases in which some observations have very small variance and other observations have very large variance, procedures which optimize the weights may achieve significant performance improvements as shown in the simulation in Section 4.1.

This work primarily focused on the case where x and σ are independent. However results from Section 3.7 showed that when independence fails, weighted estimators will typically be biased. This additional complication makes OLS more attractive relative to weighted procedures.

For practitioners we recommend caution in using the inverse of the observation variances as weights when model misspecification is present. As a check, practitioners could fit models twice, with and without weights, and compare performance based on some metric. More sophisticated methods, such as specifically tuning weights for optimal performance may be attempted. Our asymptotic theory provides guidance on how to do this in the case of the linear model.

Appendix A: Technical Notes

A.1. Proof of Theorem 3.1

Let $g(X) \in \mathbb{R}^n$ be the function g applied to the rows of X. We sometimes write w for $w(\sigma)$. We have

$$\begin{aligned}\widehat{\beta}(\widehat{W}) &= (X^T\widehat{W}X)^{-1}X^T\widehat{W}Y \\ &= (X^T\widehat{W}X)^{-1}X^T\widehat{W}(X\beta + g(X) + \Sigma^{1/2}\epsilon) \\ &= \beta + \underbrace{((1/n)X^T\widehat{W}X)^{-1}}_{\equiv q}\underbrace{(1/n)X^T\widehat{W}(g(X) + \Sigma^{1/2}\epsilon)}_{\equiv z}.\end{aligned}$$

In part 1 we show that

$$q \xrightarrow{P} \mathbb{E}[xx^T]^{-1}\mathbb{E}[w]^{-1}.$$

In part 2 we show that

$$\sqrt{n}z \xrightarrow{d} N(0, \mathbb{E}[w^2]\mathbb{E}[g^2xx^T] + \mathbb{E}[\sigma^2w^2]\mathbb{E}[xx^T]).$$

Thus by Slutsky's Theorem

$$\begin{aligned}&\sqrt{n}(\widehat{\beta}(\widehat{W}) - \beta) \\ &= q\sqrt{n}z \\ &\xrightarrow{d} N\left(0, \mathbb{E}[w]^{-2}(\mathbb{E}[w^2]\mathbb{E}[xx^T]^{-1}\mathbb{E}[g^2(x)xx^T]\mathbb{E}[xx^T]^{-1} + \mathbb{E}[\sigma^2w^2]\mathbb{E}[xx^T]^{-1})\right)\end{aligned}$$

1. **Show $q \xrightarrow{P} \mathbb{E}[xx^T]^{-1}\mathbb{E}[w]^{-1}$:** Recall that by Assumptions 1

$$\widehat{W}_{ii} = w(\sigma_i) + n^{-1/2}\delta_{nm_i}h(\sigma_i) + n^{-1}d(\sigma_i, \delta_n)$$

where h is a bounded function, δ_{nm_i} are $O_P(1)$, and the d is uniformly (in σ) bounded by an $O_P(1)$ random variable.

$$\begin{aligned}q^{-1} &= (1/n)X^T\widehat{W}X \\ &= \frac{1}{n}\sum x_ix_i^T\widehat{W}_{ii} \\ &= \frac{1}{n}\sum x_ix_i^Tw(\sigma_i) + \underbrace{\frac{1}{n^{3/2}}\sum x_ix_i^Th(\sigma_i)\delta_{nm_i}}_{\equiv R_1} + \underbrace{\frac{1}{n^2}\sum x_ix_i^Td(\sigma_i, \delta_n)}_{\equiv R_2}\end{aligned}$$

We show that $R_1, R_2 \xrightarrow{P} 0$. Noting that $\mathbb{E}[|x_{ij}x_{ik}h(\sigma_i)\mathbb{1}_{m_i=m}|] < \infty$ because h is bounded and the x have second moments we have

$$|R_{1jk}| = n^{-1/2}\left|\sum_{m=1}^{M}\delta_{nm}\left(n^{-1}\sum_{i=1}^{n}x_{ij}x_{ik}h(\sigma_i)\mathbb{1}_{m_i=m}\right)\right| \xrightarrow{P} 0.$$

Using the fact that $|d(\sigma_i, \delta_n)| < \delta_n'$ where δ_n' is $O_P(1)$ we have

$$|R_{2jk}| \leq n^{-1}\delta_n' \left(\frac{1}{n} \sum_{i=1}^{n} |x_{ij}x_{ik}| \right) \xrightarrow{P} 0.$$

Thus

$$q^{-1} \xrightarrow{P} \mathbb{E}[xx^T w] = \mathbb{E}[xx^T]\mathbb{E}[w]$$

where the last equality follows from the facts that σ and x are independent. The desired result follows from the continuous mapping theorem.

2. **Show $\sqrt{n}z \xrightarrow{d} N(0, \mathbb{E}[w^2]\mathbb{E}[g^2 xx^T] + \mathbb{E}[\sigma^2 w^2]\mathbb{E}[xx^T])$:**

$$\begin{aligned}
\sqrt{n}z &= n^{-1/2} \sum_{i=1}^{n} (g(x_i) + \sigma_i \epsilon_i) \widehat{W}_{ii} x_i \\
&= n^{-1/2} \sum_{i=1}^{n} \underbrace{(g(x_i) + \sigma_i \epsilon_i) w(\sigma_i) x_i}_{a_i} + \underbrace{n^{-1} \sum_{i=1}^{n} (g(x_i) + \sigma_i \epsilon_i) x_i \delta_{nm_i} h(\sigma_i)}_{R_3} \\
&\quad + \underbrace{n^{-3/2} \sum_{i=1}^{n} (g(x_i) + \sigma_i \epsilon_i) d(\sigma_i, \delta_n) x_i}_{R_4}
\end{aligned}$$

$\mathbb{E}[a_i] = \mathbb{E}[(g(x_i) + \sigma_i \epsilon_i) w(\sigma_i) x_i] = 0$ because $\mathbb{E}[g(x_i)x_i] = 0$ and ϵ_i is independent of all other terms and mean 0. We have

$$\begin{aligned}
Cov(a_i)_{jk} &= \mathbb{E}[a_{ij} a_{ik}] \\
&= \mathbb{E}[(g(x) + \sigma\epsilon)^2 w^2 x_j x_k] \\
&= \mathbb{E}[g^2(x) w^2 x_j x_k] + 2\mathbb{E}[g(x)\sigma\epsilon w^2 x_j x_k] + \mathbb{E}[\sigma^2 \epsilon^2 w^2 x_j x_k] \\
&= \mathbb{E}[w^2]\mathbb{E}[g^2(x) x_j x_k] + \mathbb{E}[\sigma^2 w^2]\mathbb{E}[x_j x_k].
\end{aligned}$$

So $Cov(a_i) = \mathbb{E}[w^2]\mathbb{E}[g^2 xx^T] + \mathbb{E}[\sigma^2 w^2]\mathbb{E}[xx^T]$. The desired result now follows from the CLT and showing that $R_3, R_4 \xrightarrow{P} 0$. Note that

$$\begin{aligned}
&\mathbb{E}[(g(x_i) + \sigma_i \epsilon_i) x_i h(\sigma_i) \mathbb{1}_{m_i = m}] \\
&= \mathbb{E}[g(x_i) x_i]\mathbb{E}[h(\sigma_i)\mathbb{1}_{m_i = m}] + \mathbb{E}[\sigma_i \epsilon_i x_i h(\sigma_i) \mathbb{1}_{m_i = m}] \\
&= 0.
\end{aligned}$$

Thus

$$R_3 = \sum_{m=1}^{M} \left(\delta_{nm} n^{-1} \sum_{i=1}^{n} (g(x_i) + \sigma_i \epsilon_i) x_i h(\sigma_i) \mathbb{1}_{m_i = m} \right) \xrightarrow{P} 0$$

because the terms inside the i summand are i.i.d. with expectation 0. Finally recalling that the $d(\sigma_i, \delta_n)$ is bounded above by δ_n' which is uniform $O_P(1)$, we have

$$|R_4| \leq n^{-1/2} \delta_n' \frac{1}{n} \sum_{i=1}^{n} |(g(x_i) + \sigma_i \epsilon_i) x_i| \xrightarrow{P} 0.$$

A.2. Proof of Theorem 3.2

Since $w > 0$, by Cauchy Schwartz

$$\Gamma(\nu(w)) = \frac{\mathbb{E}[w^2(\Gamma(A) + \sigma^2\Gamma(B))]}{\mathbb{E}[w]^2} \geq \mathbb{E}[(\Gamma(A) + \sigma^2\Gamma(B))^{-1}]^{-1}$$

with equality iff

$$w(\sigma) \propto \frac{1}{\Gamma(A) + \sigma^2\Gamma(B)} \propto (\sigma^2 + \Gamma(A)\Gamma(B)^{-1})^{-1}$$

with probability 1.

A.3. Proof of Corollary 3.1

We must show

$$\Gamma(\nu(w_{min})) \leq \min(\Gamma(\nu(I)), \Gamma(\nu(\Sigma^{-1})))$$

with strict inequality if $\mathbb{E}[g^2(x)xx^T]$ is positive definite and the distribution of σ is not a point mass. The inequality follows from Theorem 3.2. By Theorem 3.2, the inequality is strict whenever the functions $w(\sigma) = 1$ and $w(\sigma) = \sigma^{-2}$ are not proportional to $w_{min}(\sigma) = (\sigma^2 + \Gamma(A)\Gamma(B)^{-1})^{-1}$ with probability 1. Since $B \succ 0$ and $A = B^T\mathbb{E}[xx^Tg(x)^2]B \succ 0$, $\Gamma(A)\Gamma(B)^{-1} > 0$. So if σ is not constant with probability 1, $P(w_{min}(\sigma) = c) < 1$ for any c. Therefore w_{min} is not proportional to $w(\sigma) = 1$ with probability 1. Similarly, for w_{min} to be proportional to $w(\sigma) = \sigma^{-2}$, there must exist a c such that

$$1 = P(\sigma^2 + \Gamma(A)\Gamma(B)^{-1} = c\sigma^2) = P(\Gamma(A)\Gamma(B)^{-1} = \sigma^2(c-1)).$$

However since the constant $\Gamma(A)\Gamma(B)^{-1} > 0$ and σ is not a point mass, such a c does not exist.

A.4. Proof of Theorem 3.3

Let $\Delta = \Gamma(A)\Gamma(B)^{-1}$. In part 1 we show that

$$\widehat{\Delta} = \Delta + n^{-1/2}\delta_n$$

where δ_n is $O_P(1)$. In part 2 we show that

$$\frac{1}{\sigma_i^2 + \widehat{\Delta}} = \underbrace{\frac{1}{\sigma_i^2 + \Delta}}_{\equiv w(\sigma_i)} + n^{-1/2}\delta_n h(\sigma_i) + n^{-1}d(\sigma_i, \delta_n)$$

where δ_n is $O_P(1)$, $d(\sigma_i, \delta_n)$ is bounded uniformly by an $O_P(1)$ random variable, and h is a bounded function. Thus the weight matrix $\widehat{W}$ with diagonal elements $\widehat{W}_{ii} = (\sigma_i^2 + \widehat{\Delta})^{-1}$ satisfies Assumptions 1 with $w(\sigma) = w_{min}(\sigma)$.

1. Recall $B = \mathbb{E}[xx^T]^{-1}$. Let δ_n be $O_P(1)$ which changes definition at each appearance. Define $\widehat{B}^{-1} = n^{-1}X^TX$. By the delta method we have

$$\widehat{B} = B + n^{-1/2}\delta_n \tag{A.1}$$

and

$$\Gamma(\widehat{B}) = \Gamma(B) + n^{-1/2}\delta_n. \tag{A.2}$$

By assumption $\widehat{\beta}(\widehat{W}) = \beta + n^{-1/2}\delta_n$, thus

$$\begin{aligned}
&\left(\sum \sigma_i^{-4}\right)^{-1} \sum \sigma_i^{-4} x_i x_i^T \widehat{g}(x_i)^2 \\
&= \left(\sum \sigma_i^{-4}\right)^{-1} \sum \sigma_i^{-4} x_i x_i^T ((y_i - x_i^T \widehat{\beta}(\widehat{W}))^2 - \sigma_i^2) \\
&= \left(\sum \sigma_i^{-4}\right)^{-1} \sum \sigma_i^{-4} x_i x_i^T ((y_i - x_i^T \beta)^2 - \sigma_i^2) + n^{-1/2}\delta_n \\
&= \frac{\mathbb{E}[\sigma^{-4}]}{n^{-1}\sum \sigma_i^{-4}} \frac{1}{n} \sum \mathbb{E}[\sigma^{-4}]^{-1} \sigma_i^{-4} x_i x_i^T ((y_i - x_i^T \beta)^2 - \sigma_i^2) + n^{-1/2}\delta_n
\end{aligned}$$

Note that $\mathbb{E}[\sigma^{-4}](n^{-1}\sum \sigma_i^{-4})^{-1} \xrightarrow{P} 1$. Further note that $\mathbb{E}[\sigma^{-4}]^{-1}\sigma_i^{-4} x_i x_i^T ((y_i - x_i^T\beta)^2 - \sigma_i^2)$ are i.i.d. with expectation $\mathbb{E}[xx^T g(x)^2]$. Thus by the CLT and Slutsky's Theorem

$$\left(\sum \sigma_i^{-4}\right)^{-1} \sum \sigma_i^{-4} x_i x_i^T \widehat{g}(x_i)^2 = \mathbb{E}[xx^T g(x)^2] + n^{-1/2}\delta_n. \tag{A.3}$$

Since $\widehat{A} = \widehat{B}^T \left(\sum \sigma_i^{-4}\right)^{-1} \left(\sum \sigma_i^{-4} x_i x_i^T \widehat{g}(x_i)^2\right) \widehat{B}$, by Equations (A.1) and (A.3) we have

$$\widehat{A} = A + n^{-1/2}\delta_n.$$

which implies

$$\Gamma(\widehat{A}) = \Gamma(A) + n^{-1/2}\delta_n.$$

Combining this result with Equation (A.2) we have

$$\Gamma(\widehat{A})\Gamma(\widehat{B})^{-1} = \underbrace{\Gamma(A)\Gamma(B)^{-1}}_{\equiv \Delta} + n^{-1/2}\delta_n.$$

Since A and B are p.s.d., $\Delta \geq 0$. Therefore

$$|\Delta - \underbrace{\max(\Gamma(\widehat{A})\Gamma(\widehat{B})^{-1}, 0)}_{\equiv \widehat{\Delta}}| \leq |\Delta - \Gamma(\widehat{A})\Gamma(\widehat{B})^{-1}|.$$

Thus

$$\widehat{\Delta} = \Delta + n^{-1/2}\delta_n.$$

2. From part 1, using the fact that $(1-x)^{-1} = 1 + x + x^2(1-x)^{-1}$, we have

$$\begin{aligned}
\frac{1}{\sigma_i^2 + \widehat{\Delta}} &= \frac{1}{\sigma_i^2 + \Delta + n^{-1/2}\delta_n} \\
&= \left(\frac{1}{\sigma_i^2 + \Delta}\right)\left(\frac{1}{1 - \left(\frac{-n^{-1/2}\delta_n}{\sigma_i^2 + \Delta}\right)}\right) \\
&= \left(\frac{1}{\sigma_i^2 + \Delta}\right)\left(1 - \frac{n^{-1/2}\delta_n}{\sigma_i^2 + \Delta} + \frac{\frac{n^{-1}\delta_n^2}{(\sigma_i^2+\Delta)^2}}{1 + \frac{n^{-1/2}\delta_n}{\sigma_i^2+\Delta}}\right) \\
&= \frac{1}{\sigma_i^2 + \Delta} - n^{-1/2}\delta_n \underbrace{\frac{1}{(\sigma_i^2 + \Delta)^2}}_{\equiv h(\sigma_i)} + n^{-1} \underbrace{\frac{\delta_n^2(\sigma_i^2 + \Delta)^{-2}}{(\sigma_i^2 + \Delta) + n^{-1/2}\delta_n}}_{\equiv d(\sigma_i, \delta_n)}.
\end{aligned}$$

The function h is bounded because the σ_i are bounded below by a positive constant and $\Delta \geq 0$. Note that since $\sigma_i \geq \sigma_{min} > 0$ we have

$$d(\sigma_i, \delta_n) \leq \frac{\frac{\delta_n^2}{\sigma_{min}^4}}{\sigma_{min}^2 + n^{-1/2}\delta_n}$$

where the right hand side is $O_P(1)$.

A.5. Proof of Theorem 3.4

Let δ_n, δ_{nm} be $O_P(1)$ which change definition at each appearance. From Equations (A.1) and (A.2) in Proof A.4 we have

$$\begin{aligned}
\widehat{B} &= B + n^{-1/2}\delta_n \\
\Gamma(\widehat{B}) &= \Gamma(B) + n^{-1/2}\delta_n.
\end{aligned}$$

We have

$$\begin{aligned}
\widehat{C}_m &= \frac{1}{\sum_{i=1}^n \mathbb{1}_{m_i=m}} \sum_{i=1}^n (y_i - x_i^T\widehat{\beta}(\widehat{W}))^2 x_i x_i^T \mathbb{1}_{m_i=m} \\
&= \frac{n f_m(m)}{\sum_{i=1}^n \mathbb{1}_{m_i=m}} \left(\frac{1}{n}\sum_{i=1}^n \frac{(y_i - x_i^T\beta)^2 x_i x_i^T \mathbb{1}_{m_i=m}}{f_m(m)}\right) + n^{-1/2}\delta_{nm} \\
&= C_m + n^{-1/2}\delta_{nm}
\end{aligned}$$

where the last equality follow from the facts that the terms inside the sum are i.i.d with expectation $C_m = \mathbb{E}[(g^2(x) + \sigma_m^2)xx^T]$ and $\frac{n f_m(m)}{\sum_{i=1}^n \mathbb{1}_{m_i=m}} \to_P 1$. Thus we have

$$\widehat{W}_{min,ii} = w_{min}(m_i) + \delta_{nm_i} n^{-1/2}$$

which satisfies the form of Assumptions 1.

A.6. Proof of Theorem 3.5

$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y = \left(\frac{1}{n}\sum x_i x_i^T w(\sigma_i)\right)^{-1} \left(\frac{1}{n}\sum x_i w(\sigma_i) y_i\right)$$

By the SLLN and the continuous mapping theorem

$$\left(\frac{1}{n}\sum x_i x_i^T w(\sigma_i)\right)^{-1} \to_{as} \mathbb{E}[x x^T w(\sigma)]^{-1}.$$

Note that

$$\frac{1}{n}\sum x_i w(\sigma_i) y_i = \frac{1}{n}\sum x_i w(\sigma_i) f(x_i) + \frac{1}{n}\sum x_i w(\sigma_i)\epsilon_i \sigma_i.$$

The summands in second term on the r.h.s. are i.i.d. with expectation 0. Therefore

$$\frac{1}{n}\sum x_i w(\sigma_i) y_i \to_{as} \mathbb{E}[x w(\sigma) f(x)].$$

References

[1] B. Blight and L. Ott. A bayesian approach to model inadequacy for polynomial regression. *Biometrika*, 62(1):79–88, 1975.

[2] A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhan, and L. Zhao. Models as approximations: How random predictors and model violations invalidate classical inference in regression. *arXiv preprint arXiv:1404.1578*, 2014.

[3] R. J. Carroll. Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, pages 1224–1233, 1982.

[4] R. J. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, pages 429–441, 1982.

[5] J. Chen and J. Shao. Iterative weighted least squares estimators. *The Annals of Statistics*, pages 1071–1092, 1993.

[6] I. Czekala, S. M. Andrews, K. S. Mandel, D. W. Hogg, and G. M. Green. Constructing a flexible likelihood function for spectroscopic inference. *The Astrophysical Journal*, 812(2):128, 2015.

[7] J. H. Friedman. A variable span smoother. Technical report, DTIC Document, 1984.

[8] W. A. Fuller and J. Rao. Estimation for a linear regression model with unknown diagonal covariance matrix. *The Annals of Statistics*, pages 1149–1158, 1978.

[9] P. M. Hooper. Iterative weighted least squares estimation in heteroscedastic linear models. *Journal of the American Statistical Association*, 88(421): 179–184, 1993.

[10] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

[11] Ž. Ivezić, J. A. Smith, G. Miknaitis, H. Lin, D. Tucker, R. H. Lupton, J. E. Gunn, G. R. Knapp, M. A. Strauss, B. Sesar, et al. Sloan digital sky survey standard star catalog for stripe 82: The dawn of industrial 1% optical photometry. *The Astronomical Journal*, 134(3):973, 2007.

[12] J. Jobson and W. Fuller. Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association*, 75(369):176–181, 1980.

[13] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 425–464, 2001.

[14] J. P. Long, E. C. Chi, and R. G. Baraniuk. Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data. *arXiv preprint arXiv:1412.6520*, 2014.

[15] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3): 217–224, 2000.

[16] Y. Ma and L. Zhu. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):305–322, 2013.

[17] Y. Ma, J.-M. Chiou, and N. Wang. Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, 93(1):75–84, 2006.

[18] J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

[19] N. Mondrik, J. P. Long, and J. L. Marshall. A multiband generalization of the analysis of variance period estimation algorithm and the effect of inter-band observing cadence on period recovery rate. *arXiv preprint arXiv:1508.04772*, 2015.

[20] J. W. Richards, A. B. Lee, C. M. Schafer, P. E. Freeman, et al. Prototype selection for parameter estimation in complex models. *The Annals of Applied Statistics*, 6(1):383–408, 2012.

[21] A. G. Riess, L. Macri, S. Casertano, H. Lampeitl, H. C. Ferguson, A. V. Filippenko, S. W. Jha, W. Li, and R. Chornock. A 3% solution: determination of the hubble constant with the hubble space telescope and wide field camera 3. *The Astrophysical Journal*, 730(2):119, 2011.

[22] B. Salmon, C. Papovich, S. L. Finkelstein, V. Tilvi, K. Finlator, P. Behroozi, T. Dahlen, R. Davé, A. Dekel, M. Dickinson, et al. The relation between star formation rate and stellar mass for galaxies at 3.5 z 6.5 in candels. *The Astrophysical Journal*, 799(2):183, 2015.

[23] A. Schwarzenberg-Czerny. Fast and statistically optimal period search in uneven sampled observations. *The Astrophysical Journal Letters*, 460(2): L107, 1996.

[24] B. Sesar, Ž. Ivezić, R. H. Lupton, M. Jurić, J. E. Gunn, G. R. Knapp, N. De Lee, J. A. Smith, G. Miknaitis, H. Lin, et al. Exploring the variable sky with the sloan digital sky survey. *The Astronomical Journal*, 134(6):

2236, 2007.

[25] B. Sesar, Ž. Ivezić, S. H. Grammer, D. P. Morgan, A. C. Becker, M. Jurić, N. De Lee, J. Annis, T. C. Beers, X. Fan, et al. Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82. *The Astrophysical Journal*, 708(1):717, 2010.

[26] B. J. Shappee and K. Stanek. A new cepheid distance to the giant spiral m101 based on image subtraction of hubble space telescope/advanced camera for surveys observations. *The Astrophysical Journal*, 733(2):124, 2011.

[27] A. A. Szpiro, K. M. Rice, and T. Lumley. Model-robust regression and a bayesian" sandwich" estimator. *The Annals of Applied Statistics*, pages 2099–2113, 2010.

[28] A. Udalski, M. Szymanski, I. Soszynski, and R. Poleski. The optical gravitational lensing experiment. final reductions of the ogle-iii data. *Acta Astronomica*, 58:69–87, 2008.

[29] J. T. VanderPlas and Z. Ivezic. Periodograms for multiband astronomical time series. *arXiv preprint arXiv:1502.01344*, 2015.

[30] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

[31] H. White. Using least squares to approximate unknown regression functions. *International Economic Review*, pages 149–170, 1980.

[32] M. Zechmeister and M. Kürster. The generalised lomb-scargle periodogram. a new formalism for the floating-mean and keplerian periodograms. *Astronomy and Astrophysics*, 496(2):577–584, 2009.