Sibling Regression for Generalized Linear Models

Shiv Shankar[1] and Daniel Sheldon[1,2]

[1] University of Massachusetts, Amherst, MA 01003, USA
{sshankar,sheldon}@cs.umass.edu
[2] Mount Holyoke College, South Hadley, MA 01075, USA

Abstract. Field observations form the basis of many scientific studies, especially in ecological and social sciences. Despite efforts to conduct such surveys in a standardized way, observations can be prone to systematic measurement errors. The removal of systematic variability introduced by the observation process, if possible, can greatly increase the value of this data. Existing non-parametric techniques for correcting such errors assume linear additive noise models. This leads to biased estimates when applied to generalized linear models (GLM). We present an approach based on residual functions to address this limitation. We then demonstrate its effectiveness on synthetic data and show it reduces systematic detection variability in moth surveys.

Keywords: Sibling regression · GLM · Noise Confounding

1 Introduction

Observational data is increasingly important across a range of domains and may be affected by measurement error. Failure to account for systemic measurement error can lead to incorrect inferences. Consider a field study of moth counts for estimating the abundance of different moth species over time. Figure 1(a) shows the counts of *Semiothisa burneyata* together with 5 other species. We see that *all* species had abnormally low counts on the same day. This suggests that the low count is due to a confounder and not an actual drop in the population. The same phenomenon is also prevalent in butterfly counts (Figure 1(b)), where poor weather can limit detectability.

An abstract version of the aforementioned situation can be represented by Figure 3. X here represents ecological factors such as temperature, season etc; and Z_1, Z_2 represent the true abundance of species (such as moths). N is a corrupting noise (such as lunar phase) which affects the observable abundance θ of the organisms. Y represents an observation of the population and is modeled as a sample drawn distribution of observed abundance (for eg a Poisson distribution). Directly trying to fit a model ignoring the noise N can lead to erroneous conclusions about key factors such as effect of temperature on the population.

Distinguishing observational noise and measurement variability from true variability, often requires repeated measurements which in many cases can be

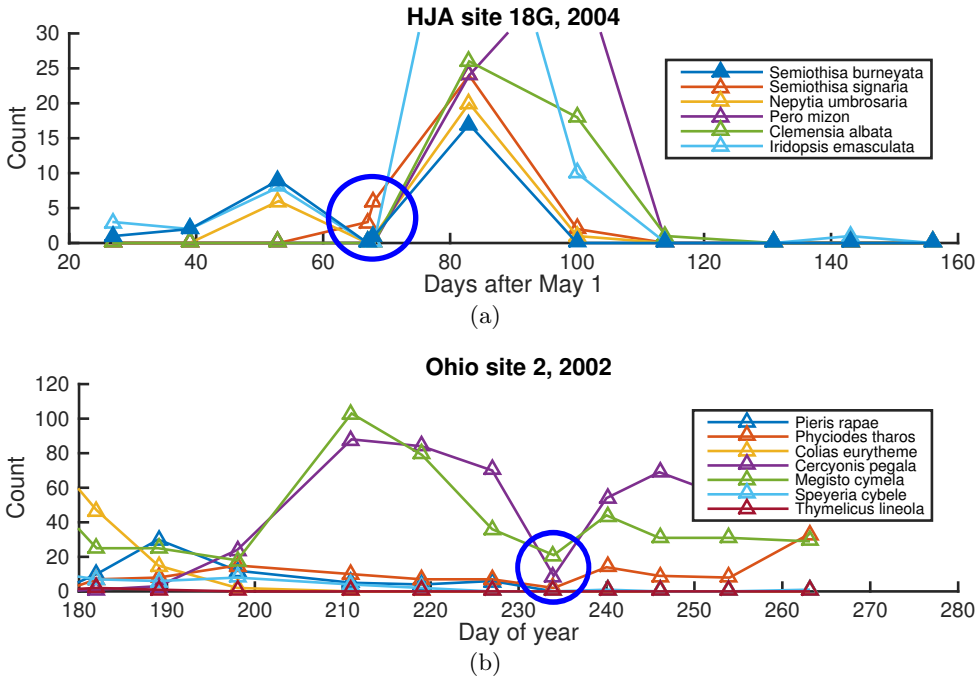

Fig. 1: Systematic detection error in moth and butterfly counts: (a) Correlated counts of other moths strongly suggest this is a detection problem. (b) Correlated detection errors in butterfly counts on day 234.

expensive, if not impossible. However, sometimes even ***in the absence of repeated measurements***, the effect of confounding noise can be estimated. Sibling regressions (Schölkopf et al., 2015) refer to one such technique that use auxiliary variables influenced by a shared noise factor to estimate the true value of the variable of interest. These techniques work without any parametric assumption about the noise distribution (Schölkopf et al., 2015; Shankar et al., 2019). However, these works assume an additive linear noise model, which limits their applications. For example in the aforementioned insect population case the effect of noise is more naturally modeled as multiplicative rather than additive.

We introduce a method that extends these ideas to generalized linear models (GLMs) and general exponential family models. First, we model non-linear effect of noise by considering it as an additive variable in the natural parameters of the underlying exponential family. Secondly, instead of joint inference of the latent variables, a stagewise approach is used. This stagewise approach is justified from a generalized interpretation of sibling regression. We provide justification behind our approach and then test it on synthetic data to quantitatively demonstrate the effectiveness of our proposed technique. Finally, we apply it to the moth survey

data set used in Shankar et al. (2019) and show that it reduces measurement error more effectively than prior techniques.

2 Related Work

Estimation of observer variation can be framed as a causal effect estimation problem (Bang and Robins, 2005; Athey and Imbens, 2016). Such conditional estimates often require that all potential causes of confounding errors have been measured (Sharma, 2018). However, real-life observational studies are often incomplete, and hence these assumptions are unlikely to hold.

Natarajan et al. (2013); Menon et al. (2015) develop techniques to handle measurement error as a latent variable. Similar approaches have been used to model observer noise as class-conditional noise (Hutchinson et al., 2017; Yu et al., 2014). One concern with such approaches is that they are generally unidentifiable.

A related set of literature is on estimation with missing covariates (Jones, 1996; Little, 1992). These are generally estimated via Monte-Carlo methods (Ibrahim and Weisberg, 1992), Expectation-maximization like methods (Ibrahim et al., 1999) or by latent class analysis (Formann and Kohlmann, 1996). Another set of approaches require strong parametric assumptions about the joint distribution (Little, 1992). Like other latent variable models, these can be unidentifiable and often have multiple solutions (Horton and Laird, 1999).

MacKenzie et al. (2002) and Royle (2004) learn an explicit model of the detection process to isolate observational error in ecological surveys using repeated measurements. Various identifiability criteria have also been proposed for such models (Sólymos and Lele, 2016; Knape and Korner-Nievergelt, 2016). Lele et al. (2012) extend these techniques to the case with only single observation. However, these models are only as reliable as the assumptions made about the noise variable. Our approach on the other hand makes weaker assumptions about the form of noise.

Schölkopf et al. (2015) introduced 'Half-sibling regression'; an approach for denoising of independent variables. This approach is both identifiable and does not make assumptions about the prior distribution of noise. Shankar et al. (2019) further extended the technique to the case when the variables of interest are only conditionally independent given an observed common cause.

3 Preliminaries

Exponential Family Distributions A random variable Y is said to be from an exponential family (Kupperman, 1958) if its density can be written as

$$p(y|\theta) = h(y)\exp(\theta^T T(y) - A(\theta))$$

where θ are the (natural) parameters of the distribution, $T(y)$ is a function of the value y called the *sufficient statistic*, and $A(\theta)$ is the log normalizing constant, and $h(y)$ is a base measure.

Given an exponential family density $p(y|\theta)$, let $L(y,\theta) = -\log p(y|\theta)$ be the negative log-likelihood loss function at data point y, and let $L(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(y^{(j)},\theta)$ be the overall loss of a sample $y^{(1)},\ldots,y^{(m)}$ from the model. Also let $I(\theta) = \nabla^2 A(\theta)$ be the Fisher information.

We summarize few standard properties (e.g., see Koller and Friedman, 2009) of exponential families that we will be of use later:

1. $\nabla A(\theta) = \mathbb{E}[T(Y)]$.
2. $\nabla_\theta L(y,\theta) = A(\theta) - T(y)$
3. $\forall y : \nabla^2_\theta L(y,\theta) = \nabla^2 A(\theta) = I(\theta)$. This also implies that $\nabla^2 L(\theta) = \nabla^2 A(\theta) = I(\theta)$.

Generalized Linear Models Generalized Linear Models (GLMs) (Nelder and Wedderburn, 1972) are a generalization of linear regression models where the output variable is not necessarily Gaussian. In a GLM, the conditional distribution of Y given covariates X is an exponential family with mean $\mathbb{E}[Y|X]$ is related to a linear combination of the covariates by the link function g, and with the identity function as the sufficient statistic, i.e., $T(Y) = Y$. We focus on the special case of GLMs with *canonical link functions*, for which $\theta = X^T\beta$, i.e., the natural parameter itself is a linear function of covariates. For the canonical GLM, the following properties hold (Koller and Friedman, 2009) [3]:

1. The link function is determined by $A : \mathbb{E}[Y|X] = g^{-1}(X^T\beta) = \nabla A(X^T\beta)$
2. The conditional Fisher information matrix $I_{Y|X} = I_{\cdot|X}(\theta) = \nabla^2 A(X^T\beta)$

GLMs are commonly used in many applications. For example, logistic regression for binary classification is identical to a Bernoulli GLM. Similarly, regressions where the response variable is of a count nature, such as the number of individuals of a species, are based on Poisson GLM.

Sibling Regression Sibling regression (Schölkopf et al., 2015; Shankar et al., 2019) is a technique which detects and corrects for confounding by latent noise variable using observations of another variable influenced by the same noise variable.

The causal models depicted in Figure 2 capture the essential use case of sibling regression techniques. Here, Z_1, Z_2 represent the unobserved variables of interest, which we would like to estimate using the observed variables Y_1, Y_2; however the observations are confounded by the common unobserved noise N. Schölkopf et al. (2015) use the model illustrated in Figure 2(a) as the basis of their *half-sibling regression* approach to denoise measurements of stellar brightness. The ***a priori*** independence of Z_1, Z_2 implies that any correlation between Y_1, Y_2 is an artifact of the noise process.

The half-sibling estimator of Z_1 from (Schölkopf et al., 2015) is

$$\hat{Z}_1 = Y_1 - \mathbb{E}[Y_1|Y_2] + \mathbb{E}[Y_1] \tag{1}$$

[3] These properties can be obtained by applying the aforementioned exponential family properties.

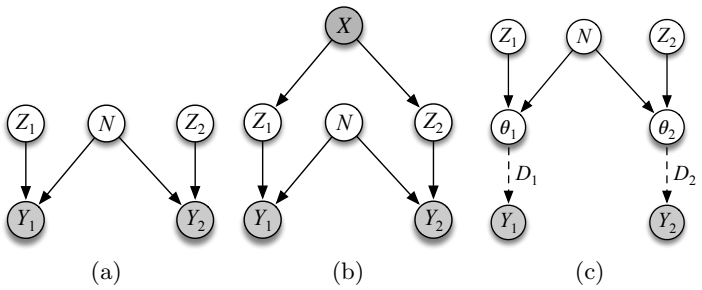

Fig. 2: (a) Half-sibling regression. (b) Three-quarter sibling regression. (c) Generalized linear sibling model without covariates.

This equation can be interpreted as removing the part of Y_1 that can be explained by Y_2, and then adding back the mean $\mathbb{E}[Y_1]$. Schölkopf et al. (2015) did not include the final $\mathbb{E}[Y_1]$ term, so our estimator differs from the original by a constant; we include this term so that $\mathbb{E}[\hat{Z}_1] = \mathbb{E}[Y_1]$. In practice, the expectations required are estimated using a regression model to predict Y_1 from Y_2, which gives rise to the name "sibling regression".

Three-quarter sibling regression (Shankar et al., 2019) generalizes the idea of half-sibling regression to the case when Z_1 and Z_2 are not independent but are conditionally independent given an observed covariate X. In the application considered by Shankar et al. (2019), these variables are population counts of different species in a given survey; the assumed dependence structure is shown in Figure 2(b). This estimator has a similar form to the half sibling version, except the expectations now condition on X:

$$\hat{Z}_{1|X} = Y_1 - \mathbb{E}[Y_1|X, Y_2] + \mathbb{E}[Y_1|X] \tag{2}$$

4 Sibling Regression for Generalized Linear Models

Model We now formally specify the model of interest. We will focus first on the half-sibling style model with no common cause X, and revisit the more general case later. Figure 2(c) shows the assumed independence structure. As in the previous models, the true variables of interest are Z_1 and Z_2, and the observed variables are Y_1 and Y_2. The variable N is confounding noise that influences observations of both variables. Unlike half-sibling regression, exponential-family sampling distributions D_1 and D_2 mediate the relationship between the hidden and observed variables; the variables θ_1 and θ_2 are the parameters of these distributions. The dotted arrow indicate that the relation between Y_i and θ_i is via sampling from an exponential family distribution, and not a direct functional dependency. On the other hand θ_1, θ_2 are deterministic functions of (Z_1, N) and (Z_2, N), respectively. We assume the noise acts additively on the natural

parameter of the exponential family. Mathematically the model is

$$Y_i \sim D_i(\theta_i), \quad \theta_i = Z_i + N \quad i \in \{1, 2\}. \tag{3}$$

More generally, the noise term N can be replaced by a non-linear function $\psi_i(N)$ mediating the relationship between the noise mechanism and the resultant additive noise; however, in general ψ_i will not be estimable so we prefer to directly model the noise as additive.

The key idea behind sibling regression techniques is to find a "signature" of the latent noise variable using observations of another variable. In this section we will motivate our approach by reformulating prior sibling regression methods in terms of *residuals*.

Lemma 1. *Let $R_i = Y_i - \mathbb{E}[Y_i]$ be the* residual *(deviation from the mean) of Y_i in the half-sibling model, and let $R_{i|X} = Y_i - \mathbb{E}[Y_i|X]$ be the residual relative to the conditional mean in the three-quarter sibling model. The estimators of Eqs.* (1) *and* (2) *can be rewritten as*

$$\hat{Z}_1 = Y_1 - \mathbb{E}[R_1|R_2], \quad \hat{Z}_{1|X} = Y_1 - \mathbb{E}\big[R_{1|X} \mid R_{2|X}\big].$$

Proof. For the half-sibling case, write

$$\begin{aligned}
\hat{Z}_1 &= Y_1 - \mathbb{E}[Y_1 \mid Y_2] + \mathbb{E}[Y_1] \\
&= Y_1 - \mathbb{E}\big[Y_1 - \mathbb{E}[Y_1] \mid Y_2\big] \\
&= Y_1 - \mathbb{E}\big[Y_1 - \mathbb{E}[Y_1] \mid Y_2 - \mathbb{E}[Y_2]\big] \\
&= Y_1 - \mathbb{E}[R_1 \mid R_2]
\end{aligned}$$

The three-quarter case is similar.

This formulation provides a concise interpretation of sibling regressions as regressing the residuals of Y_1 on those of Y_2.

4.1 Extension to GLM

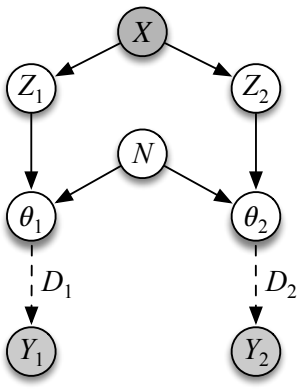

Fig. 3: Generalized sibling model with covariates

Problem Statement The problem is to obtain estimates of Z_1 and Z_2 given m independent observations $(y_1^{(1)}, y_2^{(1)}), \ldots (y_1^{(m)}, y_2^{(m)})$ of Y_1 and Y_2. The general idea is, as in prior sibling regressions, to model and remove the signature of the noise variable. Schölkopf et al. (2015); Shankar et al. (2019) solve this problem in the special case where D_i is Dirac. We wish to extend this to the case of other sampling distributions.

By symmetry, it suffices to focus on estimating only one variable, so we will henceforth consider Z_1 to be the estimation target. We allow the possibility that Z_2 is multivariate to model the case when there are many siblings that each contain a trace of the noise variable.

Inspiring from the residual form of sibling-regression 1, we derive a residual which can be used to approximate the confounding noise in the GLM case.

Computing a residual requires defining a reference, in analogy to the conditional global mean $\mathbb{E}[Y_1|X]$ used in Equation 2. We will use the maximum-likelihood estimate under the "global" model $Y_1 \sim D_1(X^T\beta)$ as the reference. Let $\hat{\beta}$ be the estimated regression coefficients for the relationship between the covariates X and variable Y_1. The corresponding model predictions $\hat{Y}_1$ are then given by $g^{-1}(X^T\hat{\beta})$ where g is the link function. We define $\mathcal{R}$ as:

$$\mathcal{R}(Y) = I_{Y|X}^{-1}(Y - \hat{Y}) \tag{4}$$

where $I_{Y|X}$ is the conditional Fisher information.

Proposition 1.

$$\begin{aligned}\mathbb{E}[\mathcal{R}(Y_1)|X, N] &= Z_1 + N - X^T\hat{\beta} + O(|Z_1 + N - X^T\hat{\beta}|^2) \\ &\approx Z_1 + N - X^T\hat{\beta}\end{aligned}$$

Proof. We temporarily drop the subscript so that $(Y|Z, N) \sim D(\theta)$ where $\theta = Z + N$ is the natural parameter of the exponential family distribution D corresponding to the specified GLM.

$$\begin{aligned}\mathbb{E}[\mathcal{R}(Y)|X, N] =& \mathbb{E}\big[I_{Y|X}^{-1}(Y - \hat{Y})|X, N\big] \\ =& \mathbb{E}\big[[I_{Y|X}(\hat{\beta})]^{-1}(Y - \nabla A(X^T\hat{\beta})|X, N\big] \\ =& [I_{Y|X}(\hat{\beta})]^{-1}(\mathbb{E}\big[Y|X, N\big] - \nabla A(X^T\hat{\beta})) \\ =& [I_{Y|X}(\hat{\beta})]^{-1}(\nabla A(Z + N) - \nabla A(X^T\hat{\beta})) \\ =& [I_{Y|X}(\hat{\beta})]^{-1}[\nabla^2 A(X^T\hat{\beta})(Z + N - X^T\hat{\beta}) \\ & \quad + O(|Z + N - X^T\hat{\beta}|^2)] \\ =& [(Z + N - X^T\hat{\beta}) + O(|Z + N - X^T\hat{\beta}|^2)]\end{aligned}$$

The second line used the definition of canonical link function. The third uses linearity of expectation. The fourth line uses the first property of exponential families from Section 3. The fifth line applies the Taylor theorem to ∇A about $X^T\hat{\beta}$. The last two lines use our previous definitions that $I_{Y|X} = \nabla^2 A(\theta)$ and $\theta = X^T\beta + N$. Restoring the subscripts on all variables except N, which is shared, gives the result.

In simpler terms, $\mathcal{R}(Y_i)$ can provide a useful approximation for $Z_i + N - X^T\hat{\beta}_i$. Moreover, if we assume that X largely explains Z_i, then the above expression is dominated by N. We would then like to use $\mathcal{R}$ to estimate and correct for the noise.

Proposition 2. *Let the approximation in Proposition 1 be exact, i.e.,* $\mathbb{E}[\mathcal{R}(Y_i)|X, N] = Z_i + N - X^T\hat{\beta}_i$*, then* $\mathbb{E}[\mathcal{R}(Y_1)|\mathcal{R}(Y_2), X] - \mathbb{E}[\mathcal{R}(Y_1)|X] = \mathbb{E}[N|X, \mathcal{R}(Y_2)]$ *upto a constant.*

The derivation is analogous to the one presented by Shankar et al. (2019) and is given in Appendix **??**. While the higher order residual terms in Claim 1 generally cannot be ignored, Claim 2 informally justifies how we can estimate N by regressing $\mathcal{R}(Y_1)$ on $\mathcal{R}(Y_2)$.

Note that $\mathcal{R}(Y_i)$ is a *random variable* and that the above expression is for the expectation of $\mathcal{R}$ which may not be exactly known (due to limited number of samples drawn from each conditional distribution) . However we observe that a) Z_1, Z_2 are independent conditional on the observed X and b) the noise N is independent of X and Z and is common between $\mathcal{R}(Y_1), \mathcal{R}(Y_2)$. This suggests that in presence of multiple auxiliary variables (let Y_{-i} denote all Y variables except Y_i) , we can improve our estimate of N by combining information from all of them. Since Z_i's are conditionally independent (given X) of each other, we can expect their variation to cancel each other on averaging. The noise variable on the other hand being common will not cancel. Appendix **??** provides a justification for this statement.

Once we have an estimate $\hat{N}$ of the noise N, we can now estimate Z_1. However unlike standard sibling regression (Figures 2(a), 2(b)), where the observed Y_1 was direct measurement of Z_1, in our case the relationship gets mediated via θ_1. If we had the true values of θ_1 one can directly use Equation 2 to obtain Z_1. One can try to obtain θ_1 from Y_1; but since Y_1 includes sampling error affecting our estimates of Z_1. Instead we rely once again on the fact that the model is a GLM, which are efficient and unbiased when all covariates are available. We re-estimate our original GLM fit but now with additional covariate $\hat{N}$ which is a proxy for N.

Implementation Based on the above insights, we propose Algorithm 1, labelled Sibling GLM (or SGLM) for obtaining denoised estimates from GLM models. In practice, the conditional expectations required are obtained by fitting regressors for the target quantity from the conditioning variables . As such we denote them by $\hat{E}[\cdot|\cdot]$): to distinguish them from true expectation values. Since, per Claim 1, the true expected value is approximately linear in the conditioning variables, in our experiments we used ordinary least squares regression to estimate the conditional expectations in Step 3.

5 Experiments

In this section, we experimentally demonstrate the ability of our approach to reduce estimation errors in our proposed setting, first using simulations and semi-synthetic examples and then with a moth survey data set. Furthermore in our experiments we found the correlation between $\mathcal{R}(Y_i)$ and X to be small, and therefore simplified Step 3 and 4 in Algorithm 1 simplify to $N = \hat{\mathbb{E}}[\mathcal{R}(Y_1)|\mathcal{R}(Y_2)]$

Algorithm 1 SGLM Algorithm

Input: $(X^{(k)}, Y_1^{(k)}, Y_2^{(k)})_{k=1,\ldots,n}$
Output: Estimates of latent variable $\hat{Z}_1$
Training: denote fitted regression models by $\hat{E}[\cdot|\cdot]$):

1. Compute $\hat{\mathbb{E}}[Y_1|X], \hat{\mathbb{E}}[Y_2|X]$ by training suitable GLM
2. Compute $\mathcal{R}(Y_1), \mathcal{R}(Y_2)$ as given by Equation 4
3. Fit regression models for $\mathcal{R}(Y_1)$ using $\mathcal{R}(Y_2), X$ as predictors to obtain estimators of $\hat{\mathbb{E}}[\mathcal{R}(Y_1)|\mathcal{R}(Y_2), X], \hat{\mathbb{E}}[\mathcal{R}(Y_1)|X]$
4. Create $\hat{N}$ such that its k^{th} value $\hat{N}^{(k)} = \hat{\mathbb{E}}[\mathcal{R}(Y_1)|\mathcal{R}(Y_2)^{(k)}, X^{(k)}] - \hat{\mathbb{E}}[\mathcal{R}(Y_1)|X^{(k)}]$ $\forall k \in [1, n]$
5. Estimate $\hat{Z}_1$ by fitting GLM models with n as an additional covariate i.e $\hat{E}[Y_1|X, n]$

5.1 Synthetic Experiment

We first conduct simulation experiments where, by design, the true value of Z_1 is known. We can then quantitatively measure the performance of our method across different ranges of available auxiliary information contained within Y_{-1}.

Description We borrow the approach of Schölkopf et al. (2015), generalized to standard GLM distributions. We run these experiments for Poisson and Gamma distributions. This simulation was conducted by generating 120 observations of 20 different Poisson variables. Each individual observation was obtained via a noise-corrupted Poisson or Gamma distribution where, for each observation, the noise affected all variables simultaneously as described below. Specifically, each variable Y_i at a single observation time is obtained via a generative process dependent on $X \in \mathbb{R}$ and noise variable $N \in \mathbb{R}$ as:

$$Y_i \sim D(\underbrace{w_X^{(i)} X}_{Z_i} + w_N^{(i)} N + \epsilon)$$

The variables X and N are drawn uniformly from $[-1, 1]$. Similarly, the coefficient $w_N^{(i)}$ is drawn from a uniform distribution on $[-1, 1]$, while $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is independent noise. Finally, $w_X^{(i)}$ is drawn from the standard conjugate prior for the distribution D.

Results We conducted these simulations and measured the error in the estimated Z versus the true Z. Due to inherent variability caused by sampling, the error will not go down to zero. We present below the results on Poisson regression, while other results can be found in the appendix.

Figure 4 shows the estimation error for Poisson regression as a function of the dimension n of Y_{-1}, i.e., the number of auxiliary variables available for denoising. In Figure 4(a) we plot the mean square error against the true value of Z_1 aggregated across the runs. Clearly increasing n reduces error. This is expected, as the noise N can be better estimated using more auxiliary variables. This in turn leads to lower error in estimates. Due to the effect of N, the standard GLM estimates are biased. The simulation results bear this out, where we get

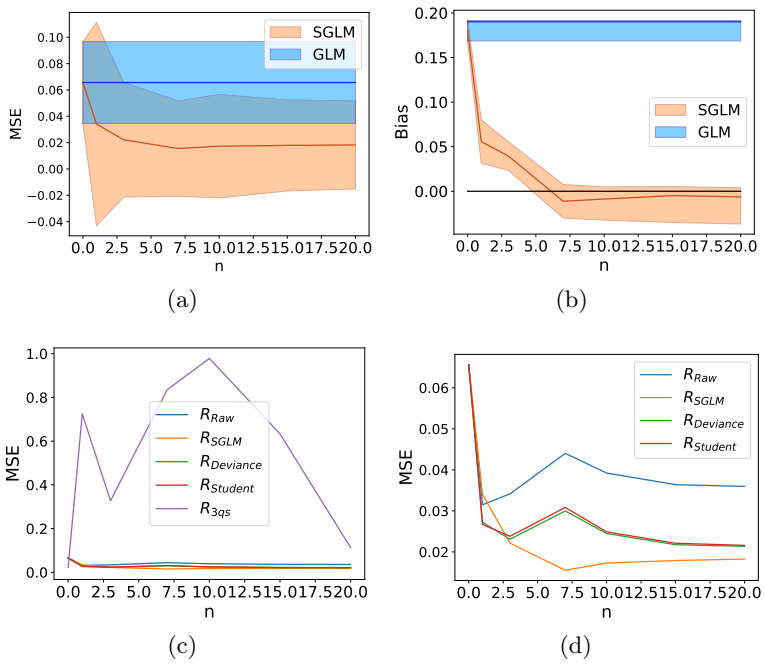

Fig. 4: Results on synthetic data. Mean Squared error (a), bias (b), and residual comparison (c,d) vs dimension of $|Y_2|$ on Poisson regressions

more than *10%* bias in Poisson regression estimates. On the other hand, by being able to correct for the noise variable on a observation basis, our approach gives bias less than *3%*. We plot in Figure 4(b) the bias of these estimates for Poisson. Once again as expected, increasing n reduces the bias.

We also experiment with other possible versions of residuals $\mathcal{R}$ [4] including residual deviance and student residuals. In Figure 4(c) we have plotted the results of these methods against the method of Shankar et al. (2019) (by fitting linear models on transformed observations). As is evident from the figure, data transformation, while a common practice, leads to substantially larger errors. Figure 4(d) presents the effect of changing the residual definition. Under our interpretation of sibling regression (Claim 1), any version of residual would be acceptable. This intuition is borne out it in these results as all the residuals perform reasonably. However from the figure, our proposed residual definition is the most effective.

[4] details in the Appendix

5.2 Discover Life Moth Observations

Our next set of experiments use moth count surveys conducted under the Discover Life Project [5]. The survey protocol uses screens illuminated by artificial lights to attract moths and then records information of each specimen. This data has been collected over multiple years at different locations on a regular basis.

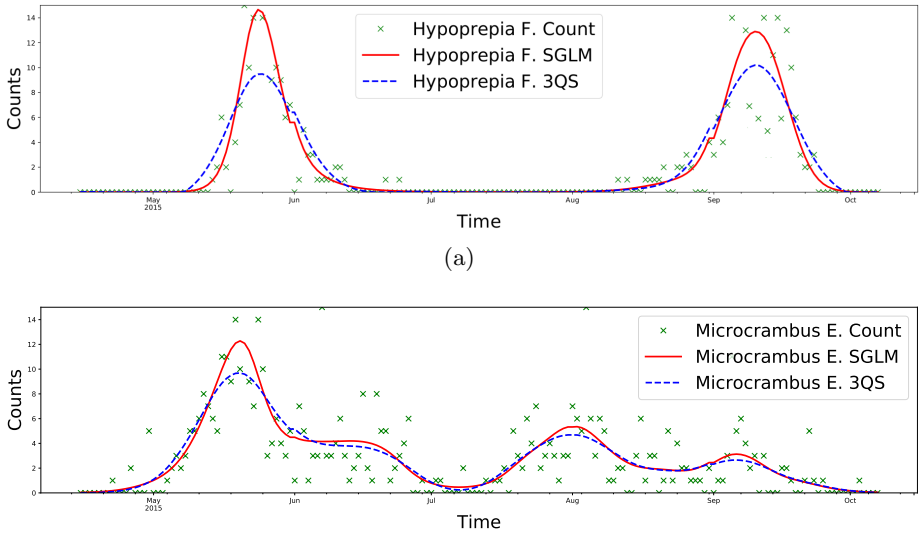

Fig. 5: Seasonal patterns of a) *Hypoprepia fucosa* and b) *Microcrambus elegans* as estimated by 3QS regression and our method alongside the observed counts. Note the higher peaks and better overall fit of our method.

A common systematic confounder in such studies is moonlight. Moth counts are often low on full moon nights as the light of the moon reduces the number of moths attracted to the observation screens. In their paper, Shankar et al. (2019) present the 'three-quarter sibling' (3QS) estimator and use it to denoise moth population counts. However, to apply the model they used least-squares regression on transformed count variables. Such transformations can potentially induce significant errors in estimation. The more appropriate technique would be to build models via a Poisson generalized additive model (GAM). In this experiment, we use our technique to directly learn the underlying Poisson model.

Description We follow the methodology and data used by Shankar et al. (2019). We choose moth counts from the Blue Heron Drive site for 2013 through 2018. We then follow the same cross-validation like procedure, holding each year out as a test fold, while using all other years for training. However, instead of transforming the variables, we directly estimated a Poisson GAM model with

[5] https://www.discoverlife.org/moth

the pyGAM (Servén and Brummitt, 2018) package. Next, we compute Z_i with our SGLM-algorithm. This procedure is repeated for all folds and all species. We compare the MSE obtained by our estimates against the 3QS estimates. Note here that due to the absence of ground truth, the prediction error is being used as a proxy to assess the quality of the model. The hypothesis is that correcting for systematic errors such as the ones induced by the moon will help to generalize better across years.

Results First we compare the residuals as used in Shankar et al. (2019) (by fitting linear models on transformed observations) against the residuals as obtained by our method, in terms of correlation with lunar brightness. A higher (magnitude) correlation indicates that the residuals are a better proxy for this unobserved confounder. The results are in Table 1. For comparison, we also provide correlations obtained by simply using the difference between the model prediction and observed values. Clearly, our method is most effective at capturing the effect of lunar brightness on the counts.

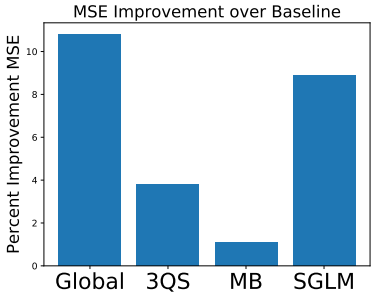

Fig. 6: Average percent improvement in predictive MSE relative to a GAM fitted to the raw counts

Species	Moonlight Correlation R_{SGLM}	R_{3QS}	R_{raw}
Melanolophia C.	**-0.55**	-0.21	-0.42
Hypagyrtis E.	**-0.66**	-0.42	-0.62
Hypoprepia F.	**-0.65**	-0.36	-0.59

Table 1: Correlation with lunar brightness of different residuals. $\hat{R}_{\text{SGLM}}$ is our residual, $\hat{R}_{\text{3QS}}$ are residuals from 3QS estimator and R_{Raw} is the raw difference between prediction and observation.

Next, we compare the decrease in prediction error obtained by our method against the methods tested by Shankar et al. (2019). These results are presented in Figure 6. The mean squared error (MSE) is computed only on data from the test-year with moon brightness zero. "Global" is an oracle model shown for comparison. It is fit on multiple years to smooth out both sources of year-to-year variability (intrinsic and moon phase). "MB" is a model that includes moon brightness as a feature to model detection variability. From the figure, we can see that our method not only improves substantially over the 3QS estimator (9% vs 4%) but is comparable with the global model fit on multiple years (9% vs 10%).[6] This is partly because the transformed linear model is unsuited for these variables. Our technique on the other hand can directly handle a broader class of conditional models and more diverse data types.

Finally, to present the difference in behavior of the two approaches, we plot the estimated moth population curves in Figure 5. These curves are plotted for

[6] The global model is included as a rough guide for the best possible generalization performance, even though it does not solve the task of denoising data within each year.

two different species with the dotted line representing the 3QS regression model, while the bold line represents our method. The actual observed counts are also plotted as points. One can clearly see the impact of the transformation, which produces flatter curves. For example, the height of the peaks for *Hypoprepia fucosa* end up significantly lower than the observed counts. On the other hand, our method predicts higher and narrower peaks, which better match the observed values.

6 Ethical Impact

Applications Our method is more focused towards ecological survey data applications. Such surveys provide information useful for setting conservation policies. However there are other potential domains of application. Measurement noise is ubiquituous in experimental studies, and applied scientists often used different schemes to protect against confounding (Genbäck and de Luna, 2019) and measurement effects (Zhang et al., 2018). As such our method may be useful applied to domains such as drug reporting (Adams et al., 2019) and epidemiology (Robins and Morgenstern, 1987).

Implications Our method provides an approach to handle the presence of unobserved confounding noise. The unobserved confounder however need not be a nuisance variable. Empirical datasets often exhibit different biases due to non-nuisance confounders (Davidson et al., 2019), which can lead to unfairness with respect race, gender and other protected attributes (Olteanu et al., 2016; Chouldechova et al., 2018). In some cases a protected attribute itself might be a confounder, in which case our approach can have implications for developing fairer models. Since our method partially recovers the unobserved confounder (noise), it can lead to identification or disclosure of protected attributes even when such data has been hidden or unavailable. This could lead to issues regarding privacy and security. The proposed method does not handle such issues, and adequate measures may be warranted for deploying this method..

7 Conclusion

Our paper has two primary contributions: a) reinterpreting sibling regression in residual form which enabled the generalization to GLMs and b) presenting a residual definition which corresponds to the case of noise in natural parameters. Based on these we designed a practical approach and demonstrated its potential on an environmental application.

A future line of work would be to develop goodness-of-fit tests for these models. A second question could be to generalize this chain of reasoning to complex non-linear dependencies. Finally since this method partially recovers the unobserved confounder (noise), it can potentially lead to identification of protected attributes even when such data has been hidden. As such another venue of future research is in the direction of how sibling regression can affect fairness and security of models.

Bibliography

R. Adams, Y. Ji, X. Wang, and S. Saria. Learning models from data with measurement error: Tackling underreporting. *arXiv:1901.09060*, 2019.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353–7360, 2016.

H. Bang and J. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962—-973, 2005.

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 134–148. PMLR, 2018.

T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Workshop on Abusive Language Online*, 2019.

A. K. Formann and T. Kohlmann. Latent class analysis in medical research. *Statistical methods in medical research*, 5(2):179–211, 1996.

M. Genbäck and X. de Luna. Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics*, 75(2):506–515, Mar 2019.

N. J. Horton and N. M. Laird. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1): 37–50, 1999.

R. A. Hutchinson, L. He, and S. C. Emerson. Species distribution modeling of citizen science data as a classification problem with class-conditional noise. In *AAAI*, pages 4516–4523, 2017.

J. G. Ibrahim and S. Weisberg. Incomplete data in generalized linear models with continuous covariates. *Australian Journal of Statistics*, 34(3), 1992.

J. G. Ibrahim, S. R. Lipsitz, and M.-H. Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1), 1999.

M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

J. Knape and F. Korner-Nievergelt. On assumptions behind estimates of abundance from counts at multiple sites. *Methods in Ecology and Evolution*, 7(2): 206–209, 2016.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Technique*. MIT Press, 2009.

M. Kupperman. Probabilities of hypotheses and information-statistics in sampling from exponential-class populations. *The Annals of Mathematical Statistics*, 29 (2):571–575, 1958. ISSN 00034851. URL http://www.jstor.org/stable/2237349.

S. R. Lele, M. Moreno, and E. Bayne. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5(1):22–31, 2012.

R. J. Little. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

D. I. MacKenzie, J. D. Nichols, G. B. Lachman, S. Droege, A. Royle, and C. A. Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

A. Menon, B. van Rooyen, C. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. *J. Mach. Learn. Res.*, 16, 2015.

N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*. 2013.

J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370—-384, 1972.

M. Torkamani, S. Shankar, P. Wallis and A. Rooshenas Learning compact neural networks using ordinary differential equations as activation functions In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019

A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *CoRR*, 2016.

J. Robins and H. Morgenstern. The foundations of confounding in epidemiology. *Computers and Mathematics with Applications*, 14:869–916, 1987.

J. A. Royle. N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics*, 60(1):108–115, 2004.

S. Shankar, S. Sarawagi Labeled memory networks for online model adaptation volume 32, number 1 In *Proceedings of the AAAI Conference on Artificial Intelligence, 2018*

B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

D. Servén and C. Brummitt. pygam: Generalized additive models in python, Mar. 2018. URL https://doi.org/10.5281/zenodo.1208723.

S. Shankar, D. Sheldon, T. Sun, J. Pickering, and T. Dietterich. Three-quarter sibling regression for denoising observational data. pages 5960–5966, 08 2019. https://doi.org/10.24963/ijcai.2019/826.

A. Sharma. Necessary and probably sufficient test for finding valid instrumental variables. *CoRR*, abs/1812.01412, 2018.

P. Sólymos and S. R. Lele. Revisiting resource selection probability functions and single-visit methods: clarification and extensions. *Methods in Ecology and Evolution*, 7(2):196–205, 2016.

H. White. *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press, 1994. https://doi.org/10.1017/CCOL0521252806.

J. Yu, R. A. Hutchinson, and W.-K. Wong. A latent variable model for discovering bird species commonly misidentified by citizen scientists. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Y. Zhang, D. Jenkins, S. Manimaran, and W. Johnson. Alternative empirical bayes models for adjusting for batch effects in genomic studies. *BMC Bioinformatics*, 19, 2018.