G-Squared Statistic for Detecting Dependence, Additive Modeling, and Calibration Concordance for Astrophysical Data

G-squared Statistic for Detecting Dependence, Additive Modeling, and Calibration Concordance for Astrophysical Data

A dissertation presented

by

Xufei Wang

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

August 2017

Dissertation advisor: Professor Jun S. Liu Xufei Wang

G-squared Statistic for Detecting Dependence, Additive Modeling, and Calibration Concordance for Astrophysical Data

Abstract

We present three topics in this thesis, G-squared statistic for independence testing as well as additive modeling, and calibration concordance by multiplicative shrinkage.

Detecting dependence is a fundamental problem. Although the Pearson correlation coefficient is effective for capturing linear dependence, it is powerless for nonlinear or heteroscedastic patterns. We introduce G-squared to test whether two univariate random variables are independent and to measure the strength of their relationship. The G-squared statistic is almost identical R-squared, for linear relationships with constant error variance, and has the intuitive meaning of the piecewise R-squared. We propose two estimators of G-squared and show their consistency. Simulations demonstrate that G-squared estimators are among the most powerful test statistics compared with several state-of-the-art methods.

We consider a nonparametric additive modeling of a reference function where the number of predictor variables can be larger than the sample size, but the number of nonzero components is comparably small. For each predictor variable, the additive component is approximated by B-spline. The G-squared estimated between each predictor and the response helps determine the knots of the B-spline. For variable selection, we apply the adaptive group least absolute shrinkage and selection operator for which we treat the spline bases of each predictor as a group; we also implement forward selection to find the subset with the minimum Bayesian information criterion value. Empirical studies show that both the approaches work well compared with two other methods.

Calibration data are often obtained by observing several sources with several instruments. Analyzing such data for proper concordance among the instruments is challenging because the physical source models are not perfectly specified and data quality varies in ways that cannot be fully quantified. We propose a log-normal hierarchical model and, for outliers, a more general log-t model. Both permit imperfection in the multiplicative mean modeling to be captured by the residual variance. Analytical solutions which take power shrinkage forms are given in special cases and Markov chain Monte Carlo algorithms are adopted for general cases. We apply our method to several data sets and demonstrate that the proposed model provides useful guidance for astrophysicists.

Contents

Listing of figures

To my family.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Prof. Jun S. Liu, for introducing me to the Statistics world and guiding me through my entire graduate life. His patience, motivation and immense knowledge have pushed me to think critically. His instruction of ‘getting your hands dirty’ has encouraged me to become a better researcher, collaborator and presenter. During these years, he has provided me with great inspiration, valuable feedback, and generous support for my research. His guidance helped me in all the time of research and writing of the dissertation. It has been my great privilege to work with him for the last five years. I also want to thank Dr. Bo Jiang for the great collaboration and discussion on G-squared for independence testing.

I would also like to thank Professor Pavlos Protopapas for introducing to the Astrophysical world. I would also like to express my sincere thanks to Professor Xiao-Li Meng for his invaluable contribution to my research in applying Statistics to astrophysical data. I am grateful to all the members of the International Astronomical Consortium for High Energy Calibration (IACHEC). I also want to thank Professor Yang Chen, David A. van Dyk, Herman L. Marshall and Vinay Kashyap for their productive collaboration and discussion on calibration concordance for astronomical instruments.

I am very thankful to have Professor Joe K. Blitzstein on my dissertation committee.

His advice on my dissertation has helped me a lot to improve the writings.

I would also like to take this moment to acknowledge all other members of the faculty in the Statistics Department, especially Professor Carl N. Morris, S. C. Samuel Kou, Natesh S. Pillai, Luke Bornn, Tirthankar Dasgupta, and Pierre E. Jacob, whose courses have shaped my understanding of Statistics. I also thank the staff of the department, Betsey Bogswell, Alice Moses, James Matejek, Madeleine Straubel, Ali Ba, and Kathleen Cloutier, for their constant help to make the journey smooth.

My time at Harvard was made enjoyable in large part due to the many friends and groups that became a part of my lift. I am grateful for time spent with my fellow colleagues, Lazhi Wang, Yang Chen, Qiuyi Han, David Jones, and Hyungsuk Tak, among many others.

Lastly, I would like to thank my family for all their love and encouragement. For my parents who raise me with a love of science and support me in all my pursuits. And most of all for my loving, supportive, encouraging, and patient husband Lawrence, whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

1

G-squared for detecting dependence

1.1 Introduction

The Pearson correlation coefficient is widely used to detect and measure the dependence of two random quantities. The square of its least-squares estimate, popularly known as the R-squared, is often used to quantify how linearly related two random variables are. However, the shortcomings of the R-squared as a measure of the strength of dependence are also significant, as discussed recently by Reshef et al.[40], which has inspired the development of many new methods for detecting dependence.

The Spearman correlation calculates the Pearson correlation coefficient between rank statistics. Although more robust than the Pearson correlation, this method still cannot capture non-monotone relationships. The alternating conditional expectation method was introduced by Breiman & Friedman[3] to approximate the maximal correlation between X and Y, i.e., to find the optimal transformations of the data, $f(X)$ and $g(Y)$, so that their correlation is maximized. The implementation of this method has its limitations because it is unfeasible to search through all possible transformations. Estimating mutual information is another popular approach because the mutual information is zero if and only if X and Y are independent. Kraskov et al.[29] proposed a method by estimating the entropy of X, Y and (X, Y) separately. The method was claimed to be numerically exact for independent cases, and effective for high dimensional variables. An energy distance-based method[53,54] and a kernel-based method[17,16] appeared separately in Statistics and machine learning literature to solve the two-sample test problem and have corresponding usage in independence tests. The two methods were recently shown to be equivalent[47]. Methods based on empirical cumulative distribution functions[22], empirical copula[15] and empirical characteristic functions[27,25] have also been proposed for detecting dependence.

Another set of approaches is based on discretizations of the random variables. Known as grid-based methods, they are primarily designed to test independence between univariate random variables. Reshef et al.[40] introduced the maximum information coefficient, which focuses on the generality and equitability of a dependence statistic; two more powerful estimators for this quantity were suggested in Reshef et al.[42]. Equitability requires that the same value of the statistic implies the same amount of dependence regardless of the type of the underlying relationship, but it is not a well-defined mathematical concept. We show that the equitability of our method is superior to all other independence testing statistics for a wide range of functional relationships. Heller et al.[21] proposed a grid based method, which utilizes the χ^2 statistic to test independence and is a distribution-free test. Doksum et al.[7] and Blyth[2] discussed the correlation curve to measure the strength of the relationship. However, a direct use of nonparametric curve estimation may rely too heavily on the smoothness of the relationship; furthermore, it also cannot deal with heteroscedastic noises.

The G-squared statistic (G^2) proposed in this chapter is derived from a regularized likelihood ratio test for piecewise linear relationships and can be viewed as an integration of continuous and discrete methods. The G-squared statistic is a function of both the conditional mean and conditional variance of one variable given the other, so it can detect general functional relationships with heteroscedastic error variances. An estimate of G^2 can be derived via the same likelihood ratio approach as R^2 when the true underlying relationship is linear. Thus, it is reasonable that G^2 is almost identical to R^2 for linear relationships. Efficient estimates of G^2 can be computed quickly using a dynamic programming method, whereas the methods of Reshef et al.[40] and Heller et al.[21] must consider grids on two variables simultaneously and hence require longer computational time. We will also show that, in terms of power, G^2 is among the the best statistics for independence testing in consideration of a wide range of functional relationships.

This chapter is organized as follows. In Section 1.2 we introduce the definition of G^2 and present two estimators, G^2_{m} and G^2_{t}. Then we study the theoretical properties of them and describe a dynamic programming algorithm. In Section 1.3 we present simulation studies to show the consistency of the estimators, as well as the power analysis and equitability study against some other popular methods. In Section 1.4 we discuss some potential future work.

1.2 Measuring dependence with G-squared

1.2.1 Defining G^2 as a generalization of R^2

The R-squared measures how well the data fit a linear regression model. Given $Y = \mu + \beta X + e$ with $e \sim \mathcal{N}(0, \sigma^2)$, the standard estimate of R-squared can be derived from a likelihood ratio test statistic for testing $\mathcal{H}_0 : \beta = 0$ against $\mathcal{H}_1 : \beta \neq 0$, i.e.,

$$R^2 = 1 - \left\{ \frac{L(\widehat{\theta})}{L_0(\widehat{\theta}_0)} \right\}^{-2/n},$$

and $L_0(\widehat{\theta}_0)$ and $L(\widehat{\theta})$ are the maximized likelihoods under $\mathcal{H}_0$ and $\mathcal{H}_1$.

Throughout this chapter, we let X and Y be univariate continuous random variables. As a working model, we assume that the relationship between X and Y can be characterized as $Y = f(X) + e$, $e \sim \mathcal{N}(0, \sigma_X^2)$ with $\sigma_X > 0$. If X and Y are independent, then $f(X) \equiv \mu$ and $\sigma_X^2 \equiv \sigma^2$. Now, let us look at a piecewise linear relationship

$$f(X) = \mu_h + \beta_h X, \quad \sigma_X^2 = \sigma_h^2, \quad c_{h-1} < X \leq c_h,$$

where c_h $(h = 0, \ldots, K)$ are called the breakpoints. While this working model allows for heteroscedasticity, it requires constant variance within each segment between two adjacent breakpoints. Testing whether X and Y are independent is equivalent to testing whether $\mu_h \equiv \mu$, $\beta_h \equiv 0$, and $\sigma_h^2 \equiv \sigma^2$. Given c_h $(h = 0, \ldots, K)$, the likelihood ratio can be written as

$$\mathrm{LR} = \exp\left(\frac{n}{2} \log \widehat{\nu}^2 - \sum_{h=1}^{K} \frac{n_h}{2} \log \widehat{\sigma}_h^2 \right),$$

where $\widehat{\nu}^2$ is the overall sample variance of Y and $\widehat{\sigma}_h^2$ is the residual variance after regressing Y on X for $X \in (c_{h-1}, c_h]$. Because R^2 is a transformation of the likelihood ratio and converges to the square of Pearson correlation coefficient, we perform the same transformation on LR. The resulting test statistic converges to a quantity related to the conditional mean and the conditional variance of Y on X. It is easy to show that, as $n \to \infty$,

$$1 - (\mathrm{LR})^{-2/n} \to 1 - \frac{\exp\left[E\{\log \mathrm{var}(Y \mid X)\}\right]}{\mathrm{var}(Y)}. \tag{1.1}$$

When $K = 1$, the relationship degenerates to a simple linear relationship and $1 - (\mathrm{LR})^{-2/n}$ is exactly R^2.

More generally, because a piecewise linear function can approximate any almost-everywhere continuous function, we can employ the same hypothesis testing framework as above to derive (1.1) for any such approximation. Thus, for any pair of random variables (X, Y), the following concept is a natural generalization of the R-squared:

$$G^2_{Y|X} \;=\; 1 - \frac{\exp\left[E\{\log \mathrm{var}(Y \mid X)\}\right]}{\mathrm{var}(Y)}, \tag{1.2}$$

in which we require that $\mathrm{var}(Y) < \infty$. Evidently, $G^2_{Y|X}$ lies between zero and one, and is equal to zero if and only if both $E(Y \mid X)$ and $\mathrm{var}(Y \mid X)$ are constant. The definition of $G^2_{Y|X}$ is closely related to the R-squared defined by segmented regression[44] discussed in Section A.2. We symmetrize $G^2_{Y|X}$ to arrive at the following quantity as the definition of the G-squared:

$$G^2 = \max(G^2_{Y|X},\ G^2_{X|Y}),$$

provided that $\mathrm{var}(X) + \mathrm{var}(Y) < \infty$. Thus, $G^2 = 0$ if and only if $E(X \mid Y)$, $E(Y \mid X)$, $\mathrm{var}(Y \mid X)$ and $\mathrm{var}(X \mid Y)$ are all constant, which is not equivalent to independence of X and Y. In practice, however, dependent cases with $G^2 = 0$ are rare.

1.2.2 Estimation of G^2

Without loss of generality, we focus on the estimation of $G^2_{Y|X}$; $G^2_{X|Y}$ can be estimated in the same way by flipping X and Y. When $Y = f(X) + e$ and $e \sim \mathcal{N}(0, \sigma^2_X)$ for an almost-everywhere continuous function $f(\cdot)$, we can use a piecewise linear function to approximate $f(X)$ and estimate G^2. However, in practice the number and locations of the breakpoints are unknown. We propose two estimators of $G^2_{Y|X}$, the first aiming to find the maximum penalized likelihood ratio among all possible piecewise linear approximations, and the second focusing on a Bayesian average of all approximations.

Suppose we have n sorted independent observations, (x_i, y_i) $(i = 1, \ldots, n)$, such that $x_1 < \cdots < x_n$. For the set of breakpoints, we only need to consider $c_h = x_i$. Each interval $s_h = (c_{h-1}, c_h]$ is called a slice of the observations, so that c_h $(h = 0, \ldots, K)$ divide the range of X into K non-overlapping slices. Let n_h denote the number of observations in slice h, and let $S(X)$ denote a slicing scheme of X, that is, $S(x_i) = h$ if $x_i \in s_h$, which is abbreviated as S whenever the meaning is clear. Let $|S|$ be the number of slices in S and let m_S denote the minimum size of all the slices.

To avoid overfitting when maximizing likelihood/log-likelihood ratios over both unknown parameters and all possible slicing schemes, we restrict the minimum size of each slice as $m_S \geq m$ and maximize the likelihood/log-likelihood ratio with a penalty on the number of slices. For simplicity, let $m = \lceil n^{1/2} \rceil$. Thus, we focus on the following penalized log-likelihood ratio

$$nD(Y \mid S, \lambda_0) = 2 \log \mathrm{LR}_S - \lambda_0(|S| - 1) \log n, \quad \lambda_0 > 0 \tag{1.3}$$

where LR_S is the likelihood ratio for S and $\lambda_0 \log n$ is the penalty for incurring one additional slice. From a Bayesian perspective, this is equivalent to assigning the prior distribution for the number of slices to be proportional to $n^{-\lambda_0(|S|-1)/2}$. Suppose each observation x_i $(i = 2, \dots, n-1)$ has probability

$$p_n = \frac{n^{-\lambda_0/2}}{1 + n^{-\lambda_0/2}}$$

of being the breakpoint independently. Then the probability of a slicing scheme S is

$$p_n^{|S|-1}(1-p_n)^{n-|S|-1} \propto \left(\frac{p_n}{1-p_n}\right)^{|S|-1} = n^{-\lambda_0(|S|-1)/2}.$$

When $\lambda_0 = 3$, the statistic $-nD(Y \mid S, \lambda_0)$ is equivalent to the Bayesian information criterion[46] up to a constant. Treating the slicing scheme as a nuisance parameter, we can maximize over all allowable slicing schemes to obtain that

$$D(Y \mid X, \lambda_0) = \max_{S:\, m_S \geq m} D(Y \mid S, \lambda_0).$$

Our first estimator of $G^2_{Y|X}$, which we call G^2_{m} with 'm' representing the maximum likelihood ratio, can be defined as

$$G^2_{\mathrm{m}}(Y \mid X, \lambda_0) = 1 - \exp\{-D(Y \mid X, \lambda_0)\}.$$

Thus, the overall G-squared can be estimated as

$$G^2_{\mathrm{m}}(\lambda_0) = \max\{G^2_{\mathrm{m}}(Y \mid X, \lambda_0),\ G^2_{\mathrm{m}}(X \mid Y, \lambda_0)\}.$$

By definition, $G^2_{\mathrm{m}}(\lambda_0)$ lies between 0 and 1 and $G^2_{\mathrm{m}}(\lambda_0) = R^2$ when the optimal slicing schemes for both directions have only one slice. Later, we will show that when X and Y are a bivariate normal, $G^2_{\mathrm{m}}(\lambda_0) = R^2$ almost surely for large λ_0.

Another attractive way to estimate G^2 is to integrate out the nuisance slicing scheme parameter. A full Bayesian approach would require us to compute the Bayes factor[28], which may be undesirable since we do not wish to impose too strong a modeling assumption. On the other hand, however, the Bayesian formalism may guide us to a desirable integration strategy for the slicing scheme. We thus put the problem into a Bayes framework and compute the Bayes factor for comparing the null and alternative models. The null model is only one model while the alternative is any piecewise linear model, possibly with countably infinite pieces. Let $p_0(y_1, \ldots, y_n)$ be the marginal probability of the data under the null. Let ω_S be the prior probability for slicing scheme S and let $p_S(y_1, \ldots, y_n)$ denote the marginal probability of the data under S. The Bayes factor can be written as

$$\mathrm{BF} \quad = \quad \sum_{S\colon\, m_s \geq m} \omega_S \times \frac{p_S(y_1, \ldots, y_n)}{p_0(y_1, \ldots, y_n)}. \tag{1.4}$$

The marginal probabilities are not easy to compute even with proper priors. Schwarz et al.[46] stated that if the data distribution is in the exponential family and the parameter is of dimension k, the marginal probability of the data can be approximated as

$$p(y_1, \ldots, y_n) \approx \mathrm{L} \exp\left\{-k(\log n - \log 2\pi)/2\right\}, \tag{1.5}$$

where L is the maximized likelihood. In our setup, the number of parameters k for the null model is two, and for an alternative model with a slicing scheme S is $3|S|$. Plugging expression (1.5) into both the numerator and the denominator of (1.4), we obtain

$$\mathrm{BF} \approx \sum_{S\colon\, m_s \geq m} \omega_S \mathrm{LR}_S \left(\frac{n}{2\pi}\right)^{-\frac{3|S|-2}{2}}. \tag{1.6}$$

If we take $\omega_S \propto n^{-\lambda_0(|S|-1)/2}$ ($\lambda_0 > 0$), which corresponds to the penalty term in (1.3) and is involved in defining G^2_{m}, the approximated Bayes factor can be restated as

$$\mathrm{BF}(\lambda_0) = \left\{ \sum_{S:\ m_S \geq m} n^{-\frac{\lambda_0(|S|-1)}{2}} \right\}^{-1} \sum_{S:\ m_S \geq m} \left(\frac{n}{2\pi} \right)^{-\frac{3|S|-2}{2}} \exp\left\{ \frac{n}{2} D(Y \mid S, \lambda_0) \right\}. \tag{1.7}$$

As we will discuss in Section 1.2.5, $\mathrm{BF}(\lambda_0)$ can serve as a marginal likelihood function for λ_0 and be used to find an optimal λ_0 suitable for a particular data set. This quantity also looks like an average version of G^2_{m}, but with an additional penalty. Since $\mathrm{BF}(\lambda_0)$ can take values below 1, its transformation $1 - \mathrm{BF}(\lambda_0)^{-2/n}$, as in the case where we derived R^2 via the likelihood ratio test, can take negative values, especially when X and Y are independent. It is therefore not an ideal estimator.

By removing the model size penalty term in (1.6), we obtain a modified version, which is simply a weighted average of the likelihood ratios and is guaranteed to be greater than or equal to 1:

$$\mathrm{BF}^*(\lambda_0) = \left\{ \sum_{S:\ m_S \geq m} n^{-\frac{\lambda_0(|S|-1)}{2}} \right\}^{-1} \sum_{S:\ m_S \geq m} \exp\left\{ \frac{n}{2} D(Y \mid S, \lambda_0) \right\}.$$

We can thus define a quantity like our likelihood formulation of R-squared,

$$G^2_{\mathrm{t}}(Y \mid X, \lambda_0) = 1 - \mathrm{BF}^*(\lambda_0)^{-2/n},$$

which we call the total G-squared, and define

$$G^2_{\mathrm{t}}(\lambda_0) = \max\{G^2_{\mathrm{t}}(Y \mid X, \lambda_0),\ G^2_{\mathrm{t}}(X \mid Y, \lambda_0)\}.$$

We show later that $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ are both consistent estimators of G^2.

1.2.3 Theoretical properties of the G^2 estimators

In order to show that $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ converge to G^2 as the sample size goes to infinity, we introduce the notations $\mu_X(y) = E(X \mid Y = y)$, $\mu_Y(x) = E(Y \mid X = x)$, $v^2_X(y) = \mathrm{var}(X \mid Y = y)$, $v^2_Y(x) = \mathrm{var}(Y \mid X = x)$, and the following regularity conditions:

Condition 1.1. *The random variables X and Y are bounded continuously with finite variances such that $v^2_Y(x)$, $v^2_X(y) > b^{-2}$ almost everywhere for some constant $b > 0$.*

Condition 1.2. *The functions $\mu_Y(x)$, $\mu_X(y)$, $v^2_Y(x)$ and $v^2_X(y)$ have continuous derivatives almost everywhere.*

Condition 1.3. *There exists a constant $C > 0$ such that*

$$\max\{|\mu'_X(y)|\,,\ |v'_X(y)|\} \le Cv_X(y), \quad \max\{|\mu'_Y(x)|\,,\ |v'_Y(x)|\} \le Cv_Y(x)$$

almost surely.

With these preparations, we can state our main results.

Theorem 1.1. *Under Conditions 1.1–1.3, for all $\lambda_0 > 0$,*

$$G^2_{\mathrm{m}}(Y \mid X, \lambda_0) \to G^2_{Y|X}, \quad G^2_{\mathrm{t}}(Y \mid X, \lambda_0) \to G^2_{Y|X}$$

almost surely as $n \to \infty$. Thus, $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ are consistent estimators of G^2.

A proof of the theorem is provided in Section A.3. It is expected that $G^2_{\mathrm{m}}(\lambda_0)$ should converge to G^2 just because of its construction. It is surprising that $G^2_{\mathrm{t}}(\lambda_0)$ also converges to G^2. The result, which links G^2 estimation with the likelihood ratio and Bayesian formalism, suggests that most of the information up to the second moment has been fully utilized in the two test statistics. The theorem thus supports the use of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ for testing whether X and Y are independent. The null distributions of the two statistics depend on the marginal distributions of X and Y, which can be generated empirically using permutation. One can also do a quantile-based transformation on X and Y such that their marginal distributions are standard normal; however, the G^2 based on the transformed data tends to lose some power.

When X and Y are bivariate normal, the G-squared statistic is almost the same as the R-squared when λ_0 is large enough.

Theorem 1.2. *If X and Y follow bivariate normal distribution, then for n large enough*

$$\mathrm{pr}\left\{G_{\mathrm{m}}^2(\lambda_0) = R^2\right\} \quad > \quad 1 - 3n^{-\lambda_0/3+5}.$$

So for $\lambda_0 > 18$ *and* $n \to \infty$, *we have* $G_{\mathrm{m}}^2(\lambda_0) = R^2$ *almost surely.*

The lower bound on λ_0 is not tight and can be relaxed in practice. Empirically, we have observed that $\lambda_0 \;=\; 3$ is large enough for $G_{\mathrm{m}}^2(\lambda_0)$ to be very close to R^2 in the bivariate normal setting.

1.2.4 Dynamic programming algorithm for computing G_{m}^2 and G_{t}^2

The brute force calculation of either G_{m}^2 or G_{t}^2 has a computational complexity of $O(2^n)$ and is prohibitive in practice. Fortunately, we have found a dynamic programming scheme for computing both quantities with a time complexity of $O(n^2)$. The algorithms for computing $G_{\mathrm{m}}^2(Y \mid X, \lambda_0)$ and $G_{\mathrm{t}}^2(Y \mid X, \lambda_0)$ are roughly the same except for one operation, i.e., maximization versus summation, and can be summarized by the following steps:

Step 1.1 (Data preparation)**.** *Arrange the observed pairs* (x_i, y_i) $(i = 1, \ldots, n)$ *according* *to the sorted x_is from low to high. Then normalize y_i such that*

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} y_i^2 = n.$$

Step 1.2 (Main algorithm)**.** *Define $m = \lceil n^{1/2} \rceil$ as the smallest slice size, $\lambda = -\lambda_0 \log(n)/2$ and $\alpha = e^{\lambda}$. Initialize three sequences: $(M_i,\ B_i,\ T_i)$ $(i = 1, \ldots, n)$ with $M_1 = 0$ and $B_1 = T_1 = 1$. For $i = m, \ldots, n$, recursively fill in entries of the tables with*

$$M_i = \max_{k \in K_i} \left(\lambda + M_k + l_{k:i}\right), \quad B_i = \sum_{k \in K_i} \alpha B_k, \quad T_i = \sum_{k \in K_i} \alpha T_k L_{k:i},$$

where $K_i = \{1\} \cup \{k : k = m+1, \ldots, i-m+1\}$, $l_{k:i} = -(i-k)\log(\widehat{\sigma}^2_{k:i})/2$ and $L_{k:i} = \exp\{l_{k:i}\}$, with $\widehat{\sigma}^2_{k:i}$ as the residual variance of regressing y on x for observations (x_j, y_j) $(j = k, \ldots, i)$.

Step 1.3. *The result is*

$$G^2_{\mathrm{m}} = 1 - \exp\left\{-\frac{2}{n}(M_n - \lambda)\right\}, \quad G^2_{\mathrm{t}} = 1 - (T_n/B_n)^{-2/n}.$$

Here, M_i $(i = m, \ldots, n)$ stores the partial maximized likelihood ratio up to the ordered observation (x_k, y_k) $(k = 1, \ldots, i)$, B_i $(i = m, \ldots, n)$ stores the partial normalizing constant, and T_i $(i = m, \ldots, n)$ stores the partial sum of the likelihood ratios. When n is extremely large, we can speed up the algorithm by considering fewer slice schemes. For example, we can divide X into chunks of size m by rank and consider only slicing schemes between the chunks. For this method, the computational complexity is $O(n)$. We can compute $G^2_{\mathrm{m}}(X \mid Y, \lambda_0)$ and $G^2_{\mathrm{t}}(X \mid Y, \lambda_0)$ similarly to get $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$. Empirically, the algorithm is faster than many other powerful methods, as shown in Section A.1.

1.2.5 An empirical Bayes strategy for selecting λ_0

Although the choice of the penalty parameter λ_0 is not critical for the general use of G^2, we typically use $\lambda_0 = 3$ for G^2_{m} and G^2_{t} because $D(Y \mid X, 3)$ is equivalent to the Bayesian information criterion. Fine-tuning λ_0 can improve the estimation of G^2. We thus propose a data-driven strategy for choosing λ_0 adaptively. $\mathrm{BF}(\lambda_0)$ in (1.7) can be viewed as an approximation to $\mathrm{pr}(y_1, \ldots, y_n \mid \lambda_0)$ up to a normalizing constant. We thus can use the maximum likelihood principle to choose the λ_0 that maximizes $\mathrm{BF}(\lambda_0)$. We then use the chosen λ_0 (called λ_0^*) to compute $G^2_{\mathrm{m}}(\lambda_0^*)$ and $G^2_{\mathrm{t}}(\lambda_0^*)$ as estimators of G^2. In practice, we evaluate $\mathrm{BF}(\lambda_0)$ for a set of discrete λ_0 values, e.g., $\{0.5l\}_{l=1}^{8}$, and pick the one that maximizes $\mathrm{BF}(\lambda_0)$. $\mathrm{BF}(\lambda_0)$ can be computed efficiently via a dynamic programming algorithm similar to that described in Section 1.2.4. As an illustration, we consider the sampling distributions of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ with $\lambda_0 = 0.5,\ 1.5,\ 2.5$ and 3.5 for

Example 1.1. *$X \sim \mathcal{N}(0,1)$, $Y = X + e$ and $e \sim \mathcal{N}(0, \sigma^2)$;*

Example 1.2. *$X \sim \mathcal{N}(0,1)$, $Y = \sin(4\pi x)/0.7 + e$ and $e \sim \mathcal{N}(0, \sigma^2)$.*

We simulate $n = 225$ data points. For each model, we set $\sigma = 1$ so that $G^2_{Y|X} = 0.5$ and perform 1,000 replications. Figure 1.1 shows histograms of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ with different λ_0 values. The results demonstrate that, for relationships that can be approximated well by a linear function, a larger λ_0 is preferred because it penalizes the number of slices more heavily and the resulting sampling distributions are less biased. On the other hand, for complicated relationships such as the trigonometric function, a smaller λ_0 is preferable because it allows more slices, which can help capture fluctuations in the functional relationship. The figure also shows that the empirical Bayes selection of λ_0 works very well, leading to a proper choice of λ_0 for each simulated data set from both examples and resulting in the most accurate estimates of G^2. Now we let $\sigma = 9.95$ so that $G^2_{Y|X} = 0.01$.

Figure 1.2 presents the same analysis as Fig. 1.1 but here X and Y are almost independent. A larger λ_0 is preferable for both models; a small λ_0 tends to use more slices than necessary and overfits the relationship. The data-driven λ_0 still gives the most accurate estimates of the $G^2_{Y|X}$. Consistency of the data-driven estimators is proven in Section A.3.

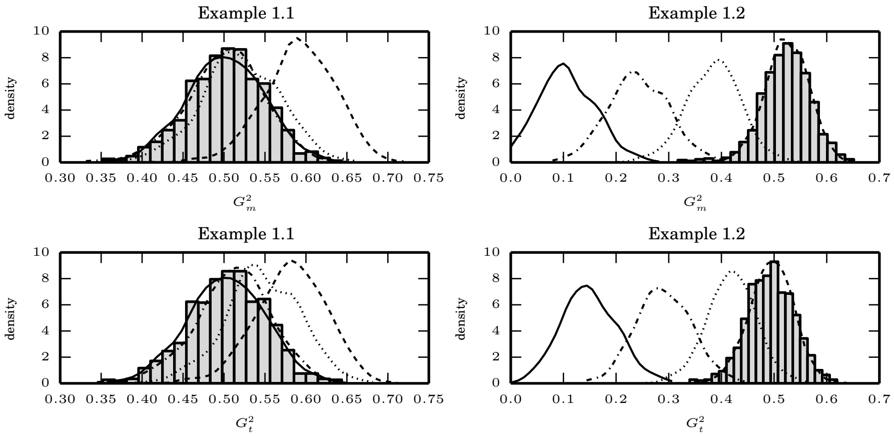

Figure 1.1: Sampling distributions of G^2_{m} and G^2_{t} under the two models in Section 1.2.5 with $G^2_{Y|X} = 0.5$ for $\lambda_0 = 0.5$ (dashes), 1.5 (dots), 2.5 (dot-dashes) and 3.5 (solid). The density function in each case is estimated by the histogram. The sampling distributions of G^2_{m} and G^2_{t} with the empirical Bayes selection of λ_0 are in gray shadow and overlaid on top of other density functions.

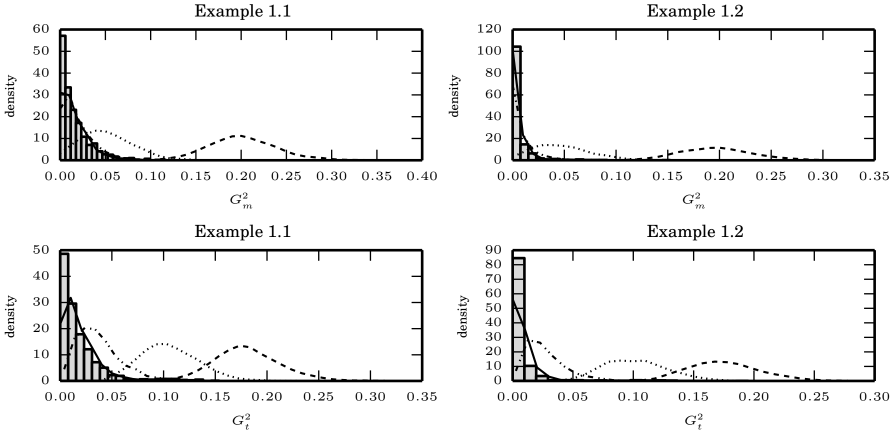

Figure 1.2: Sampling distributions of G^2_{m} and G^2_{t} under the two models in Section 1.2.5 with $G^2_{Y|X} = 0.01$. The legends are the same as in Fig. 1.1.

1.3 Simulation studies

1.3.1 Consistency of G^2_{m} and G^2_{t}

For a general relationship, the true value of G^2 is nontrivial to compute. However, we can calculate $G^2_{Y|X}$ for some special examples and evaluate the sum of squared errors of the estimators. Especially when $\sigma_X \equiv \sigma$,

$$G^2_{Y|X} = \frac{\mathrm{var}\{f(X)\}}{\mathrm{var}\{f(X)\} + \sigma^2}.$$

With $X \sim U(0,1)$, we consider Examples 1.3–1.10

Example 1.3. *$Y = X + e$ and $e \sim \mathcal{N}(0,1)$;*

Example 1.4. *$Y = X + e$ and $e \sim \mathcal{N}(0,\sigma_X^2)$;*

Example 1.5. *$Y = X^2/\sqrt{2} + e$ and $e \sim \mathcal{N}(0,1)$;*

Example 1.6. *$Y = X^2/\sqrt{2} + e$ and $e \sim \mathcal{N}(0,\sigma_X^2)$;*

Example 1.7. *$Y = X + e$ and $e \sim \sqrt{3}U(-1,1)$;*

Example 1.8. *$Y = X + e$ and $e \sim \sqrt{3}\sigma_X U(-1,1)$;*

Example 1.9. *$Y = X^2/\sqrt{2} + e$ and $e \sim \sqrt{3}U(-1,1)$;*

Example 1.10. *$Y = X^2/\sqrt{2} + e$ and $e \sim \sqrt{3}\sigma_X U(-1,1)$.*

For Examples 1.3, 1.5, 1.7 and 1.9, $G^2_{Y|X} = 0.5$; for Examples 1.4, 1.6, 1.8 and 1.10, $\sigma_X = \exp\{-|X|/2\}$ and $G^2_{Y|X} = 0.7$. We simulate $1,000$ replications for each model and sample size combination, and use $\lambda_0 = 3$ for G^2_{m} and G^2_{t}. Table 1.1 shows the sum of squared errors of $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$ for the different models as n varies.

We find that the sum of squared errors decrease roughly in the order of n^{-1} for both estimators and that G^2_{t} appears slightly more accurate. The sum of squared errors are similar when the function relationships are the same, regardless of the error type. This confirms that the estimation accuracies of G^2_{m} and G^2_{t} are not sensitive to the Gaussian assumption.

Table 1.1: Sum of squared errors for G^2_{m} and G^2_{t} with increasing n

		G^2_{m}			G^2_{t}	
n	100	225	400	100	225	400
Example 1.3	5.11	2.37	1.35	4.99	2.39	1.35
Example 1.4	4.56	2.56	1.42	3.56	1.88	1.05
Example 1.5	19.27	9.30	5.17	13.15	6.41	3.67
Example 1.6	16.45	7.55	4.16	11.53	5.37	3.04
Example 1.7	4.87	2.29	1.49	5.56	2.76	1.93
Example 1.8	4.10	2.43	1.49	3.12	1.77	1.08
Example 1.9	20.29	9.05	5.38	13.45	6.13	3.78
Example 1.10	17.29	8.98	4.82	11.73	6.42	3.46

1.3.2 Power analysis

Now we compare the powers of different independence testing methods for various relationships. Here, we again fix $\lambda_0 = 3$ for both G^2_{m} and G^2_{t}. Other methods we test include the alternating conditional expectation[3](ACE), Genest's test[15], Pearson correlation (COR), distance correlation[53](DCOR), the method of Heller et al.[21](DDP), the characteristic function method[27], Hoeffding's test[22], the mutual information method[29] and two methods, MIC$_e$ and TIC$_e$, based on the maximum information criterion[40]. We follow the procedure for computing the powers of different methods as described in previous studies of Reshef et al.[41] and a 2012 online note by N. Simon and R. Tibshirani.

For different relationships $f(X)$ and different values of noise levels σ^2, we simulate $(X,\ Y)$ with the following model:

$$X \sim U(0,1), \quad Y = f(X) + e, \quad e \sim \mathcal{N}(0, \sigma^2),$$

where $\mathrm{var}\{f(X)\} = 1$. Thus $G^2_{Y|X} = (1 + \sigma^2)^{-1}$ is a monotone function of the signal-to-noise ratio and it is of interest for us to observe how the performances of different methods deteriorate as the signal strength weakens for various relationships. We use permutations to generate the null distribution and to set the rejection region for all testing methods in all examples.

Figure 1.3 shows the power comparisons for eight relationships. We set the sample size $n = 225$ and perform 1,000 replications for each relationship and $G^2_{Y|X}$ value. For a clear presentation, we only plot COR, DCOR, DDP, TIC$_e$, G^2_{m} and G^2_{t}, and put results for other methods in Section A.4. For any method with tuning parameters, we choose the one that results in the highest average power over all the examples. Due to computational concerns, we choose $K = 3$ for DDP. It is seen that G^2_{m} and G^2_{t} perform robustly, and are always among the most powerful methods, with G^2_{t} performing slightly more powerful than G^2_{m} in almost all examples. They outperform other methods in cases such as the high frequency sine, triangle and piecewise constant functions, where piecewise linear approximation is more appropriate than other approaches. For monotonic examples such as linear and radical relationships, G^2_{m} and G^2_{t} have slightly lower power than COR, DCOR and DDP, but are still highly competitive.

We also study the performances of these methods for $n = 50,\ 100$ and 400, and find that G^2_{m} and G^2_{t} still show high power regardless of n although their advantages are much less obvious when n is small. More details are in Section A.4.

1.3.3 Equitability

Intuitively, equitability[40] reflects the ‘robustness’ of a statistic that describes the dependence between two random variables, to the underlying relationship. For example, Pearson correlation is not an equitable statistic because it is zero for (X, Y) in Example 1.2, no matter how small σ is. An ideal equitable statistic can imply the same amount of depen-

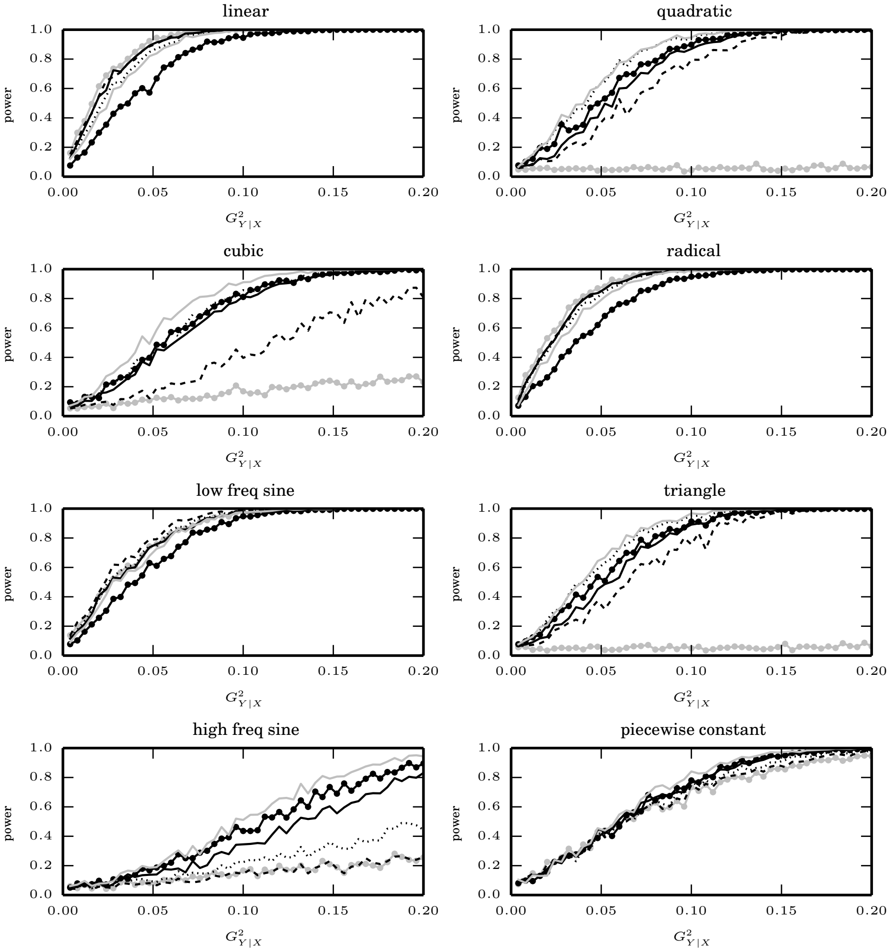

Figure 1.3: The powers of G^2_{m} (black solid), G^2_{t} (grey solid), COR (grey markers), DCOR (black dashes), DDP (black dots) and TIC$_e$ (black markers) for testing independence between X and Y when the underlying true relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant, respectively. The x-axis is $G^2_{Y|X}$, a monotone function of the signal-to-noise ratio, and the y-axis is the power. We choose $n = 225$ and perform 1,000 replications for each relationship and $G^2_{Y|X}$.

dence, regardless of the type of relationship. In other words, equitable statistics can be used to gauge the degree of dependence. Reshef et al.[43] gave two equivalent definitions for the equitability of a statistic that measures dependence. They used $\Psi = \mathrm{cor}^2\{Y, f(X)\}$ to define the degree of dependence when the dependence of Y on X can be described by a function. When $\mathrm{var}(Y \mid X)$ is a constant, $\Psi \equiv G^2_{Y|X}$. For a perfectly equitable statistic, its sampling distribution should be almost identical for different relationships with the same Ψ. But the existence of such a statistic for any well-defined large class of relationships remains unclear.

We repeat the equitability study in Reshef et al.[40]. Figure 1.4 shows the 95% confidence bands for G^2_{m} and G^2_{t}, compared with ACE, COR, DCOR and MIC$_e$ for $X \sim \mathcal{N}(0, 1)$ and the relationships in Example 1.3–1.6. We choose different values of Ψ with $n = 225$ and conduct 1,000 replications for each case. The plots show that G^2_{m} and G^2_{t} increase along with Ψ for all relationships, as expected, and that the confidence bands obtained under different relationships have a similar size and location for the same Ψ. The confidence bands are also comparably narrow. The MIC$_e$ displays good equitability, though slightly worse than G^2_{m} and G^2_{t}, while the other three statistics do poorly for non-monotone relationships. ACE tends to have wider confidence bands for Examples 1.5 and 1.6 than the other methods, while COR and DCOR have non-overlapping confidence bands for different relationships when Ψ is moderately large. In other words, COR and DCOR can yield drastically different values for two relationships with the same Ψ. This phenomenon is as expected, since it is known that these two statistics do not perform well for non-monotone relationships.

An alternative strategy to study equitability of a statistic is to test $\mathcal{H}_0 : \Psi = x_0$ against $\mathcal{H}_1 : \Psi = x_1$ $(x_0 < x_1)$ for a broad set of relationships using the statistic. The more powerful a test statistic for all types of relationships, the better its equitability. For each aforementioned method, we perform right-tailed tests with the type-I error fixed at $\alpha =$

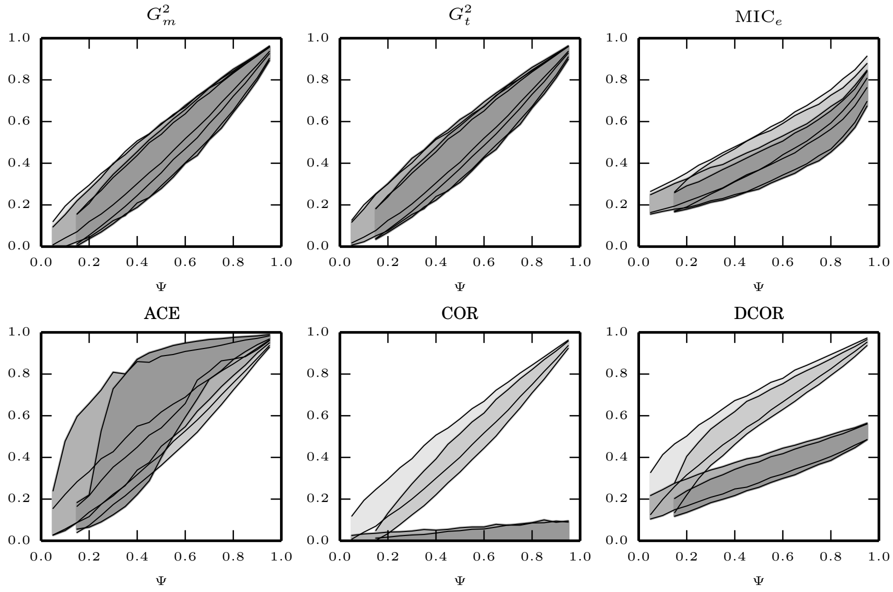

Figure 1.4: The plots from the top left to the bottom right are the 95% confidence bands of Ψ for the 6 indicated methods. We choose $n = 225$ and perform 1,000 replications for each relationship and each value of Ψ for Examples 1.3–1.6. The shadow is the lightest for Example 1.3 and darkest for Example 1.6. Ψ is a monotone function of the signal-to-noise ratio when the error variance is constant. The y-axis shows the values of the corresponding statistic.

0.05 and different combinations of (x_0, x_1) $(0 < x_0 < x_1 < 1)$. Given a fixed sample size, a perfectly equitable statistic should yield the same power for all kinds of relationships so that it is able to reflect the degree of dependence by a single value. Most statistics can perform well only for a small class of relationships. We use a heat map to demonstrate the average power of a test statistic with different pairs of (x_0, x_1) in Fig. 1.5. Each dot in the plot represents the average power of a test statistic over a class of relationships; the darker the color, the higher the power. We simulate (X, Y) with the following model

$$X \sim U(0,1),\ Y = f(X) + e,\ e \sim \mathcal{N}(0, \sigma^2).$$

The twenty chosen relationships, which are inspired by Reshef et al.[41] are in Section A.4. We carry out the testing for $(x_0, x_1) = (i/50, j/50)$ $(i < j = 1, \ldots, 49)$ with $n = 225$

and conduct 1,000 replications. For any method with a tuning parameter, we choose the parameter that results in the greatest average power. We observe that G^2_{m}, G^2_{t} and MIC$_e$ have the best equitability, followed by ACE and TIC$_e$. The average powers for G^2_{m}, G^2_{t} and MIC$_e$ over the entire range of (x_0, x_1) are all 0.6, although G^2_{m} and G^2_{t} are slightly better for larger x_0's. Besides, using our empirical Bayes method to select λ_0, the equitability of G^2_{m} and G^2_{t} can be further improved. In comparison, all the remaining methods are not as equitable.

1.4 Discussions and future works

G-squared can be viewed as a direct generalization of R-squared. While maintaining the same interpretability as R-squared, G-squared is also a powerful measure of dependence for general relationships. Instead of resorting to curve-fitting methods to estimate the underlying relationship, we employ piecewise linear approximations with penalties and dynamic programming algorithms. Furthermore, one can approximate a relationship between two variables with piecewise polynomials or other flexible basis functions, with perhaps additional penalty terms to control the complexity. In the next chapter, we generalize this idea and use the G^2_{m} estimator to select knots in spline curve fitting.

Right now, the distributions of G^2_{m} and G^2_{t} for two independent random variables are still unknown. Simulations show when (X, Y) follow independent normal, the distribution of G^2_{m} is close to the distribution of R^2 {Beta($\frac{1}{2}$, $\frac{n-2}{2}$)}, but with a heavier right tail. We can transform X and Y to normal distributions and use Beta($\frac{1}{2}$, $\frac{n-2}{2}$) as a reference for computing the p-value. This approach can save computation time compared with the permutation test but the power is much lower when the signal is weak. It's worthwhile studying the distributions of G^2_{m} or G^2_{t} for independent relationship. Another potential work is

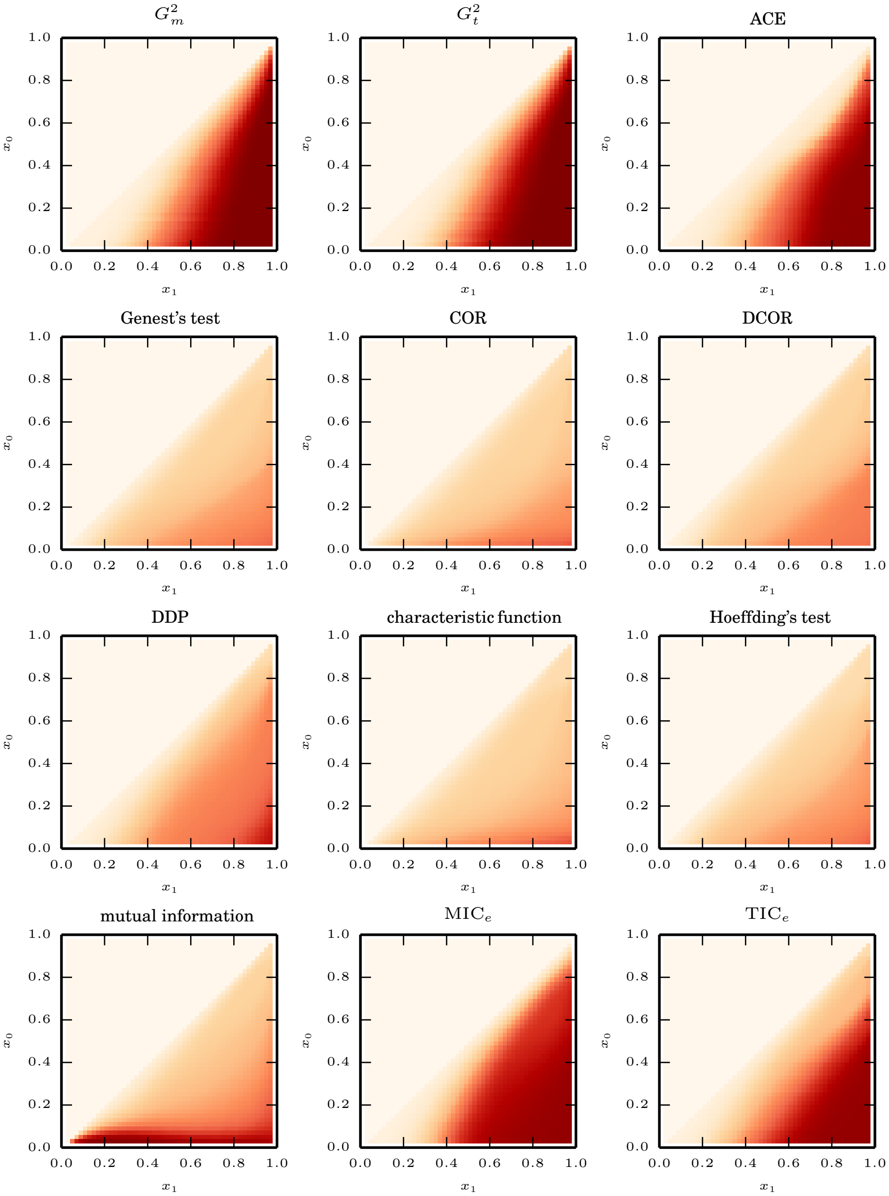

Figure 1.5: Heat maps for the equitability of different methods. Each red dot corresponding to (x_1, x_0) represents the power of the method for testing $\mathcal{H}_0 : \Psi = x_0$ against $\mathcal{H}_1 : \Psi = x_1$, averaging over a class of relationships. The darker a dot, the higher the average power. We choose sample size $n = 225$ and perform 1,000 replications for each relationship and pair of (x_1, x_0).

that we can generalize the definition of G-squared in (1.2) as

$$1 - \frac{g^{-1}(E[g\{\mathrm{var}(Y \mid X)\}])}{\mathrm{var}(Y)},$$

where g is an increasing concave function. For G-squared, we choose $g = \log$ and we can study the statistic with other possible functions in the future.

Remark This chapter is based on a published paper by Wang et al.[56]

2

G-squared for additive modeling

2.1 Introduction

The problem of estimating the relationship between a response variable and multiple predictor variables emerged long time ago from many practical problems. Let Y be the response variable of which the distribution depends on the predictor variables $X_1, \ldots, X_p$, such that

$$Y = f(X_1, \ldots, X_p) + e, \quad e \sim [0, \sigma^2].$$

The function f is called the reference function. The classic linear regression is a special case which assumes that f is a linear combination of the predictors. When the reference function is complex, even if there is only one predictor, the problem becomes challenging.

Another special case of the problem is the additive model introduced by Hastie & Tibshirani[19], Stone[51] and Stone[52]. In additive model, the function f is the summation of p univariate functions, each of which is a function of a distinct predictor variable. To be precise, the dependence of Y on $X_1, \ldots, X_p$ is

$$Y = \sum_{j=1}^{p} f_j(X_j) + e, \quad e \sim [0, \sigma^2].$$

First, let us discuss the problem when there is only one predictor, denoted as X.

In Statistics and machine learning literature, two main approaches are used for univariate curve fitting: kernel regression and spline method. The kernel regression, proposed by Nadaraya[36] and Watson[57], assumed that the realization of $f(x)$ is a weighted average of f in a neighborhood of x. Silverman[48] showed that spline smoothing corresponds approximately to smoothing by a kernel method with bandwidth depending on the local density of design points. Spline method approximates f with a continuous piecewise polynomial function. The places where the pieces connect are called knots. Friedman & Silverman[12] and Friedman[11] utilized truncated linear functions for curve fitting. Other commonly used spline is the basis spline (B-spline) function[5,6], which has the minimal support with respect to a given degree and partition. For example, twice continuous differentiable cubic splines with equidistant knots are commonly used. The number and locations of the knots are key to the curve fitting and pre-determined knots can always be questioned with counter examples. Many data-driven methods were proposed to tackle this problem. Smoothing spline[10] took every observation of X as the knots and penalized the squared L_2 norm of the secondary derivative of the fitted curve. Eilers & Marx[9] and Wand & Ormerod[55] came up with P-spline and O-spline respectively, both of which had very similar penalty as smoothing splines and gave approximations of the squared L_2 norm of the secondary derivatives. These methods can even take more knots than the number of observations.

In this chapter, we propose a partition scheme based on the dependence between X and Y. In Chapter 1 we introduce the G-squared statistic to evaluate the dependence between two univariate random variables and provide two estimators. The $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ estimator gives a piecewise linear, though sometimes discontinuous, estimation of the underlying relationship. We will use the knots from this partition and perform spline fitting.

For additive modeling, we will first find the knots for each predictor variable, create the spline bases and regress the response on all the bases. Unfortunately, additive modeling has good performance when the number of variables (p), is much smaller than the sample size (n). In recent years, practical problems inspired the study of large-p-small-n cases, where the response depends only on a small proportion of the predictors, but the number of predictors are much larger than the sample size. Lin et al.[30] penalized the sum of the reproducing kernel Hilbert space norms of each component, to select variables and fit a nonparametric estimation of each component. This method can also consider interactions between variables. Ravikumar et al.[39] chose the L_2 norm of each component as the penalty, which can be treated as a function version of the group least absolute shrinkage and selection operator[60] (LASSO). Huang et al.[24] utilized adaptive group LASSO, as a generalization of the adaptive LASSO[61], to select nonzero components. In this chapter, we propose two approaches for components selection: the adaptive group LASSO penalty and Bayesian information criterion (BIC). When p is comparably small, we suggest the BIC approach because it has smaller integrated squared error; when p is comparably large, we suggest the adaptive group LASSO approach because it has a more consistent variable selection result.

The rest of this chapter is organized as follows. In Section 2.2 we propose the spline curve fitting with the knots by the G^2_{m} estimator. In Section 2.3 we discuss two marginal curve fitting methods and the two approaches for additive modeling. In Section 2.4 we present some simulation studies, both for curve fitting and additive modeling, and apply the methods to the Boston housing data. Section 3.4.2 concludes this chapter and proposes future research works.

2.2 G-squared for curve fitting

2.2.1 G^2_{m} estimator for curve fitting

Suppose we have two random variables X and Y such that $Y = f(X) + e$ where $e \sim [0, \sigma^2]$. When estimating $G^2_{Y|X}$ with $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$, we obtain a piecewise linear estimator of $f(X)$. If the optimal slicing has more than one slice, the estimator is discontinuous at the breakpoints. Let $\hat{f}$ be the piecewise linear function fitted by $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$. To show that $\hat{f}$ converges to f, we need the following conditions:

Condition 2.1. *The random variables X and Y are bounded continuously with finite nonzero variances.*

Condition 2.2. $f(X)$ *has continuous and bounded derivatives almost everywhere.*

With these preparations, we can state that

Theorem 2.1. *Under Conditions 2.1 and 2.2, for all* $\lambda_0 > 0$*,*

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{f}(X_i)\}^2 \to \sigma^2 \text{ and } \frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - \hat{f}(X_i)\}^2 \to 0 \text{ as } n \to \infty.$$

The theorem, proved in Section B.1, shows that the mean squared error (MSE) will converge to zero as the sample size increases. Another interesting quantity to describe consistency besides MSE is the integrated squared error (ISE), defined as

$$\text{ISE} = E\{f(X) - \hat{f}(X)\}^2.$$

This quantity can be viewed as the expectation of the out-of-sample MSE. It is also of interest to see whether the integrated squared error will converge to zero as the sample size increases. We will compare ISE in the simulation studies. After estimating $G^2_{Y|X}$ by the G^2_{m} estimator, if there are more than one slice, we can use the breakpoints as knots to perform spline fitting. In this chapter, we use cubic spline throughout the examples. Simulation studies show that this method is more robust to different f compared with the plain spline fitting with equidistant knots.

2.2.2 Choice of λ_0

As discussed in Section 1.2.5, the choice of λ_0 can determine the estimation of $G^2_{Y|X}$. We suggest $\lambda_0 = 3$ because the corresponding penalized log-likelihood resembles BIC. We also propose a data-drive strategy, which can help improve the accuracy of estimating G^2. This strategy tries to balance the signal strength and curve complexity. Suppose there is one oracle λ_0 that yields the minimum ISE, called the oracle ISE. Simulation studies in Section 2.4.2 shows that the ISE by λ_0^* (defined in Section 1.2.5) is close to the oracle ISE. In the rest of this chapter, if not stated explicitly, we use λ_0^* for curve fitting and additive modeling.

2.3 G-squared for additive modeling

Suppose the reference function between Y and $X_1, \ldots, X_p$ is

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + e, \quad Ef_j(X_j) = 0, \quad \mathrm{var}(e) = \sigma^2.$$

We assume that only a small subset of f_js are nonzero. For additive modeling, we need to select the nonzero components and fit the corresponding curves. In this section, we present the two approaches for variable selection in additive modeling. Before any selection step, we first estimate G^2_{m} between each predictor and the response, and then create spline bases of each predictor according to the knots by G^2_{m}. If no need for variable selection, we regress the response on the bases. Otherwise, we propose 1) the adaptive group LASSO to select variables, in which each group consists of the spline bases of each variable; 2) BIC to select variables, in which we choose or drop an entire group together. Simulation studies in Section 2.4.4 show that the first approach has better consistency with respect to variable selection and the second approach has smaller prediction error.

2.3.1 Marginal curve fitting

In additive modeling, we calculate G^2_{m} between each predictor and the response marginally, and take the knots under each optimal slicing scheme. The B-spline is well defined for bounded variables, so for simplicity we assume that each predictor is between 0 and 1. Now let us have a look at the spline bases used for additive modeling. For each variable X_j, if the optimal slicing scheme is only one slice, the spline base is $\psi_{j,1}(x) \propto x$ such that

$$E\{\psi_{j1}(X_j)\} = 0, \quad \mathrm{var}\{\psi_{j1}(X_j)\} = 1.$$

If the optimal slicing scheme has more than one slice, suppose the knots are c_{j_t} $(t = 1, \ldots, K_j{-}m)$ where m is the degree of the spline, the spline bases are denoted as $\psi^m_{jk}(x)$ $(k = 1, \ldots K_j)$ such that

$$E\{\psi^m_{jk}(X_j)\} = 0, \quad \text{var}\{\psi^m_{jk}(X_j)\} = 1, \quad \text{cov}\{\psi^m_{jk_1}\psi^m_{jk_2}(X_j)\} = 0\ (k_1 \neq k_2).$$

We treat the bases of each predictor variables as a group and use these groups for the adaptive group LASSO or the BIC model selection.

When some of the marginal G^2_{m}s are too small, the marginal curve fitting can select only one slice and fail to fit the relationship properly. We introduce an adaptive way for the marginal fitting. Before describing the procedure, let us look at a simple and intuitive example. Suppose there are only two independent predictor variables, X_1 and X_2, then

$$G^2_{Y|X_j} = \frac{\text{var}\{f_j(X_j)\}}{\text{var}\{f_1(X_1)\} + \text{var}\{f_2(X_2)\} + \sigma^2}\ (j = 1, 2).$$

If we can estimate $f_1(X_1)$ precisely, then let $R = Y - f_1(X)$ and

$$G^2_{R|X_2} = \frac{\text{var}\{f_2(X_2)\}}{\text{var}\{f_2(X_2)\} + \sigma^2}.$$

It is obvious $G^2_{R|X_2} \geq G^2_{Y|X_2}$ and there is no doubt the estimation of $f_2(X_2)$ is easier by $G^2_{R|X_2}$ than $G^2_{Y|X_2}$. This inspires us to explore knots for each predictor adaptively: we can fit G^2_{m} on the residual variable instead of on the original response variable. The advantage is obvious: the G^2 between the predictor and the residual variable should be larger than that between the predictor and the response, so the relationship is easier to capture. In other words, in each step, we calculate the G^2_{m} between the residual variable and the remaining predictors, pick the largest one and subtract the piecewise linear fitted values from the residual variable. Algorithms 1 and 2 present the steps for the simple and adaptive marginal curve fitting. The shortcoming of the adaptive method is that the procedure can introduce new noises, and the computation time is $O(p^2)$. Simulation studies show that when the number of predictors is comparably small and when the predictors are independent, the adaptive marginal fitting can reduce ISE. When the number of predictors is large or when the predictors are dependent, we prefer the simple marginal fitting.

Algorithm 1: Simple marginal curve fitting

```
for $j = 1, \dots, p$ do
    Compute spline bases with $G_{\mathrm{m}}^2(Y \mid X_j, \lambda_0^*)$;
end
```

Algorithm 2: Adaptive marginal curve fitting

```
Initialize $R = Y, I = \{1, \dots, p\}$;
for $j = 1, \dots, p$ do
    for $k \in I$ do
        Let $k_j = \mathrm{argmax}_{k \in S} G_{\mathrm{m}}^2(R \mid X_k, \lambda_0^*)$ and $\hat{f}_{k_j}$ be the corresponding piecewise
          linear fitted values;
        Compute spline bases with $G_{\mathrm{m}}^2(Y \mid X_{k_j}, \lambda_0^*)$;
        Let $R = R - \hat{f}_{k_j}$ and $I = I \setminus \{k_j\}$;
    end
end
```

2.3.2 Adaptive group LASSO for additive modeling

First, we describe the group LASSO. For linear regression, variable selection with response Y and predictors X_j's, if we have the prior knowledge that some of the predictors should be selected or dropped together, we can perform group LASSO[60] instead of plain LASSO. Suppose that X can be grouped by the non-overlapping groups $g \in \mathcal{G}$. Let

$I_g = \{j : j \in g\}$ be the indices of the predictors belonging to group g. Suppose the predictors inside each group are standard and orthogonal, that is $X_{I_g}^T X_{I_g} = I_{|I_g|}$. The group LASSO minimizes the following penalty function:

$$\|Y - \mu - \sum_{g \in \mathcal{G}} X_{I_g} \beta_{I_g}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \sqrt{|I_g|} \|\beta_{I_g}\|_2. \tag{2.1}$$

Here, the L_2 norm of a random variable Z is defined as $\|Z\|_2^2 = E(Z^2)$. Under this setting, $\|\beta_{I_g}\|_2 = \|X_{I_g}\beta_{I_g}\|_2$, which is the L_2 norm of the fitted values of each group. $\sqrt{|I_g|}$ penalizes the group size. The group LASSO guarantees that the predictors inside each group are selected or dropped together. In the additive modeling, we can treat each spline base $\psi_{j,k}^m(X_j)$ as a new variable and the spline bases for a same predictor belong to a same group. The penalty function is as follows

$$\|Y - \mu - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \psi_{jk}^m(X_j)\beta_{jk}\|^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K_j} \beta_{jk}^2}. \tag{2.2}$$

Let$\tilde{f}_j = \sum_{k=1}^{K_j} \psi_{jk}^m(X_j)\beta_{jk}$, (2.2) is equivalent to

$$\|Y - \mu - \sum_{j=1}^{p} \tilde{f}_j(X_j)\|^2 + \lambda \sum_{j=1}^{p} \|\tilde{f}_j(X_j)\|_2.$$

The difference between (2.1) and (2.2) is that we do not penalize the size of each group. This is because when calculating G_{m}^2, we have already considered the number of slices.

For real observations, suppose the realization of $(Y, X_1, \ldots, X_p)$ are $(y_i, x_{1i}, \ldots, x_{pi})$ $(i = 1, \ldots, n)$. After we find the knots of each variable and calculate the spline bases, we linearly transform the splines bases, $\psi^m_{jk}(x_{ji})$ to $\varphi^m_{jk}(x_{ji})$, such that

$$\frac{1}{n}\sum_{i=1}^{n}\varphi^m_{jk}(x_{ji})=0, \quad \frac{1}{n}\sum_{i=1}^{n}\varphi^m_{jk_1}(x_{ji})\varphi^m_{jk_2}(x_{ji})=I_{\{k_1=k_2\}}\ (k_1,\ k_2=1,\ldots,p).$$

Then we minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\bar{y}-\sum_{j=1}^{p}\sum_{k=1}^{K_j}\varphi^m_{j,k}(x_{j,i})\beta_{j,k}\right\}^2+\lambda\sum_{j=1}^{p}\sqrt{\sum_{k=1}^{K_j}\beta^2_{j,k}}. \tag{2.3}$$

Adaptive LASSO is introduced by Zou[61] for a more consistent variable selection. For our adaptive group LASSO step, we first perform group LASSO and get the estimated $\hat{\beta}_{jk}$'s. Then we perform a second step group LASSO to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\sum_{j=1}^{p}\sum_{k=1}^{K_j}\varphi^m_{jk}(x_{ji})\beta_{jk}\right\}^2+\lambda\sum_{j=1}^{p}\omega_j\sqrt{\sum_{k=1}^{K_j}\beta^2_{jk}}, \quad \omega_j^{-1}=\sqrt{\sum_{k=1}^{K_j}\hat{\beta}^2_{jk}}.$$

If some predictor is not selected in the first step group LASSO, the weight for the second step group LASSO equals infinity, which means this predictor will not be selected in the second group LASSO. Intuitively, if one component is selected in the first step but its L_2 norm is extremely small, this component can possibly be a false positive selection. In the second step, its weight is very large so that this component can be easily dropped.

2.3.3 Bayesian information criterion

The Bayesian information criterion is defined as

$$n\log\left[\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\sum_{j\in J}\sum_{k=1}^{K_j}\varphi^m_{j,k}(x_{j,i})\beta_{j,k}\right\}^2\right]+\log(n)d_J, \tag{2.4}$$

where J is the set of the selected predictors and d_J is the degree of freedom. Intuitively, d_J is between $|J|$ and $\sum_{j\in J} K_j$, so we choose $\sum_{j\in J} K_j$ in this chapter. The optimal variable selection is J^* that minimizes (2.4). We perform forward selection to find the optimal J^* and only use BIC when the number of predictors is small.

2.4 Empirical studies

2.4.1 G-squared for curve fitting

In this section, we compare the plain spline fitting and the spline fitting with the knots by G^2_{m}, which we call the G^2_{m} estimator (denoted as $\tilde{f}$) in this section. We consider the following relationships in Fig. 2.1:

Example 2.1. *$f_1(x) = x/0.23$,*

Example 2.2. *$f_2(x) = (x-0.5)^2/0.075$,*

Example 2.3. *$f_3(x) = ||x-0.5|-0.25|/0.07$,*

Example 2.4. *$f_4(x) = \sin(4\pi x^3)/0.58$.*

The relationships are normalized so when $X \sim U(0,1)$, $\mathrm{var}\{f_j(x)\} = 1$ $(j = 1,\ldots,4)$. Let use consider $Y = f_j(X) + e$ with $e \sim \mathcal{N}(0,1)$ and $n = 225$. For the plain spline fitting, when we choose N equidistant knots, they are $j/(N+1)$ $(j = 1,\ldots,N)$. To estimate ISE, we sample $n_0 = 10,000$ independent new observations, x^0_i $(i = 1,\ldots n_0)$, from $U(0,1)$ and estimate the ISE by the

$$\frac{1}{n_0}\sum_{i=1}^{n_0}\{\tilde{f}_j(x^0_i) - f_j(x^0_i)\}^2.$$

Tables 2.1 shows the average ISE for the four examples by different strategies. We choose $\lambda_0 = 3$ for the G^2_{m} estimator. For the plain spline fitting, if the locations of the knots are exactly the change-points of the curve, like $N = 1$ for f_2 and $N = 3$ for f_3, the spline fitting yields the minimum ISEs and is slightly smaller than the G^2_{m} estimator. However, when the locations of the knots do not match the real change-points, like $N = 2$ for f_3, the integrated squared error increases drastically. For the real problem, it is unrealistic to assume that the number of change-points is known or the change-points are equidistant. Besides, when the relationship is linear (f_1) or when the changes points are not equidistant (f_4), ISE by the plain spline fitting is larger than the G^2_{m} estimator. We suggest the G^2_{m} estimator for curve fitting because it is robust to the underlying relationships, and can adjust the number and locations of the knots based on the data.

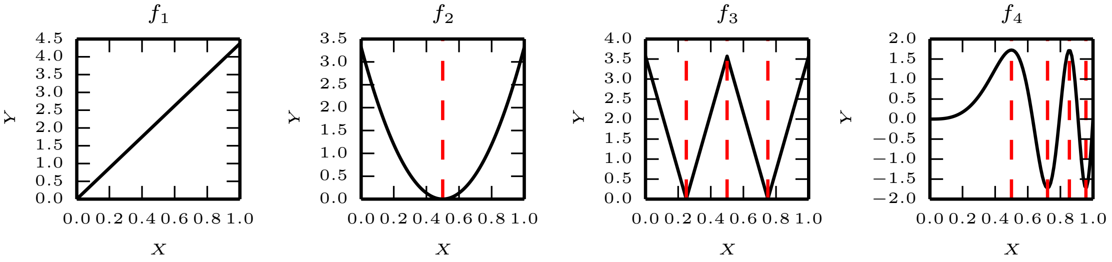

Figure 2.1: The curves for f_j $(j = 1, \dots, 4)$. The dashes indicate the change-points.

Table 2.1: The average ISEs for $f_j(x)$ $(j = 1, \dots, 4)$ with plain spline fitting and the G^2_{m} estimator

	f_1	f_2	f_3	f_4
$N = 1$	**0.023 (0.015)**	**0.023 (0.015)**	0.236 (0.021)	0.761 (0.031)
$N = 2$	0.028 (0.017)	0.027 (0.016)	0.382 (0.027)	0.629 (0.028)
$N = 3$	0.033 (0.019)	0.032 (0.018)	**0.044 (0.017)**	**0.364 (0.060)**
$N = 4$	0.038 (0.021)	0.037 (0.019)	0.085 (0.020)	0.496 (0.110)
G^2_{m}	**0.011 (0.014)**	**0.027 (0.017)**	**0.065 (0.035)**	**0.262 (0.150)**

2.4.2 Optimal λ_0 for curve fitting

Next, let us study the choice of λ_0 for the following four relationships:

Example 2.5. $g_1(x) = x/0.23.$

Example 2.6. $g_2(x) = \sin(\pi x)/0.3,$

Example 2.7. $g_3(x) = \sin(2\pi x)/0.7,$

Example 2.8. $g_4(x) = \sin(3\pi x)/0.7,$

The relationships are normalized so when $X \sim U(0, 1)$, $\text{var}\{g_j(x)\} = 1$ $(j = 1, \ldots 4)$. Let use consider $Y = g_j(X) + e$ with $e \sim \mathcal{N}(0, 1)$ and $n = 225$. We choose λ_0 from $\{0.1 \times 1.1^t\}_{t=1}^{60}$. For each pair of (X, Y), we fit the G^2_{m} estimator with each λ_0 and then fit the G^2_{m} estimator with λ^*_0 chosen from the above 60 candidates, as discussed in Section 1.2.5 and 2.2.2. In Fig. 2.2, the bold lines represent the ISEs fitted by different λ_0s, and the dashes are the integrated squared errors by the data-driven λ^*_0. The ISE by λ^*_0 is quite close to the oracle ISE and the ratio between them is less than 1.13. An interesting phenomenon is that for each g_j, there is a region of λ_0 that can yield almost the same minimum ISE. This is because for λ_0's in this region, the slicing schemes for a pair of (X, Y) are always the same. Figure 2.3 shows the histograms of the chosen λ^*_0 and most of the values are smaller than 4. In the following sections, when we use the data-driven strategy to choose λ_0, we choose from $\{0.5l\}_{l=1}^{8}$.

For a better presentation of our methods, we use some abbreviations for different variations. The simple marginal fitting is denoted as MGS-; the adaptive marginal fitting is denoted as AMGS-. The adaptive group LASSO approach is denoted as -AGL; the BIC approach is denoted as -BIC. For example, a method with adaptive marginal fitting and adaptive group LASSO approach is called AMGS-AGL.

2.4.3 Simple and adaptive marginal fitting

In this section, we repeat the example in Lin et al.[30] and let $X_j = (W_j + tU)/(1 + t)$ where $U,\ W_j \sim U(0, 1)$ $(j = 1, \ldots, p)$. Therefore $\text{corr}(X_j, X_k) = t^2/(1 + t^2)$. The relationships

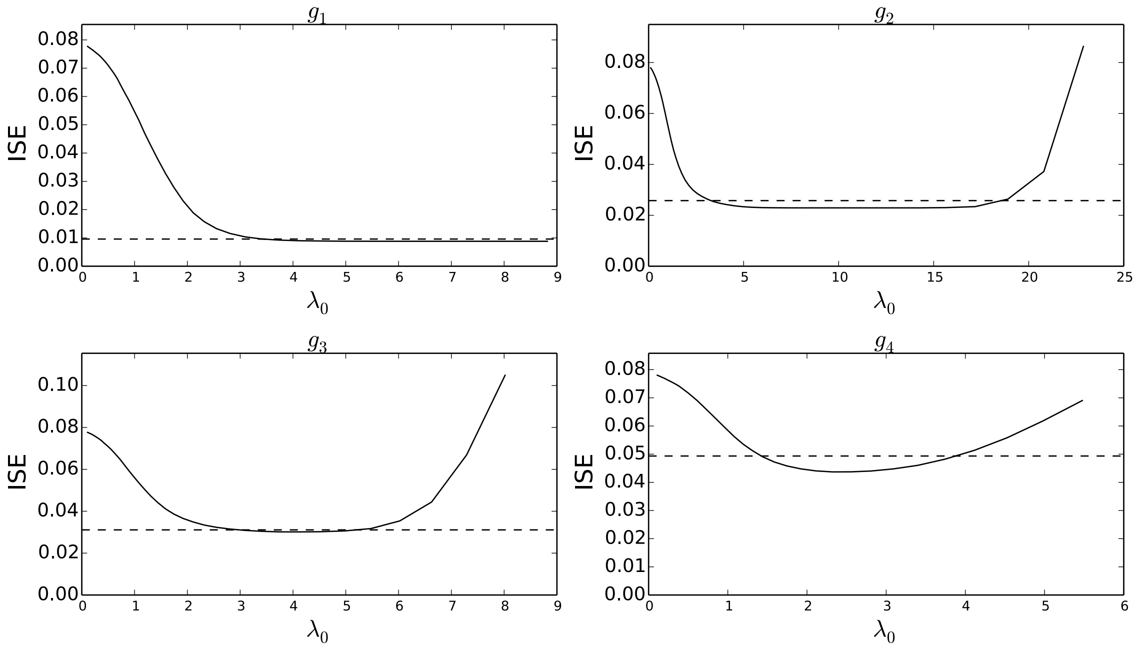

Figure 2.2: The ISEs by different λ_0's and λ_0^*'s for g_j $(j = 1, \dots, 4)$

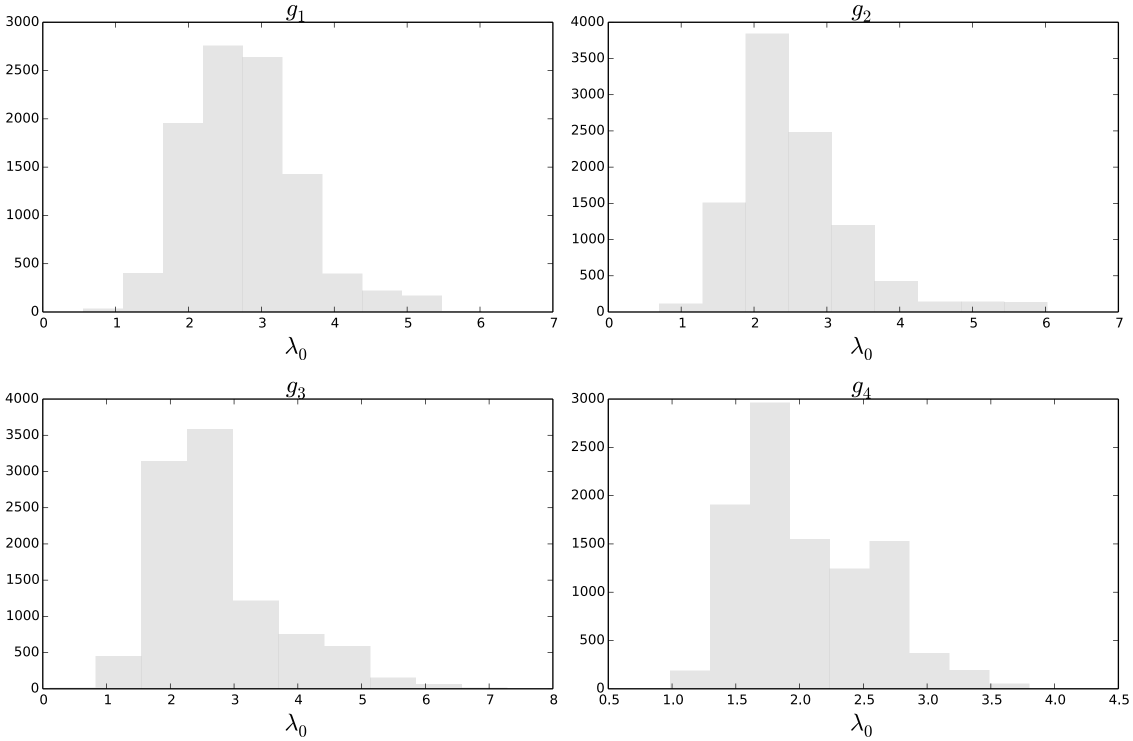

Figure 2.3: The histogram of λ_0^*'s for g_j $(j = 1, \dots, 4)$

are

$$\begin{aligned} h_1(x) &= 5x, \quad h_2(x) = 3(2x-1)^2, \quad h_3(x) = \frac{4\sin(2\pi x)}{2-\sin(2\pi x)}, \\ h_4(x) &= 6\{0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) \\ &\quad +0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)\}. \end{aligned}$$

Let $e \sim \mathcal{N}(0,1)$ and $\sigma = 1.32$ so that the signal to noise ratio is 9. When $t = 0$, the predictors are independent, and

$$G^2_{Y|X_1} = 0.12, \quad G^2_{Y|X_2} = 0.05, \quad G^2_{Y|X_3} = 0.19, \quad G^2_{Y|X_4} = 0.54.$$

If we can pick X_4, X_3, X_1, X_2 and remove $h_4(X_4)$, $h_3(X_3)$, $h_1(X_1)$ sequentially, the adjusted G-squared will become

$$G^2_{Y|X_4} = 0.54, \quad G^2_{R_1|X_3} = 0.41, \quad G^2_{R_2|X_1} = 0.45, \quad G^2_{R_3|X_2} = 0.31,$$

where $R_1 = Y - h_4(X_4)$, $R_2 = R_1 - h_3(X_3)$ and $R_3 = R_2 - h_1(X_1)$. The original marginal G-squared for X_2 is extremely small compared with the other variables, however, the 'adjusted' G-squared of X_2 increases by six times.

We choose $p = 10$, $n = 100,\ 225,\ 400$ and fit with MGS-AGL, MGS-BIC, AMGS-AGL, AMGS-BIC for 1,000 repetitions. For the adaptive group LASSO, we pick λ by five-fold cross validation. Besides, we compare our method with SPAM by Ravikumar et al.[39] and the adaptive group LASSO method (AGL) by Huang et al.[24]. For SPAM and AGL, we use 5 knots and the locations are the $k/6$ $(k = 1, \ldots, 5)$ quantiles for each predictor. Table 2.2 presents the ISEs with $t = 0$. The adaptive marginal curve fitting, especially with -BIC, has smaller ISE, compared with SPAM and AGL. Table 2.3 shows the false negative and positive selections. We find that the adaptive marginal curve fitting with -AGL reduces both selections compared with the simple marginal curve fitting. This is because in the simple marginal curve fitting, G^2_{m} estimator always fits one slice for X_2 and the linear correlation between $h_2(X_2)$ and Y happens to be zero, so X_2 is not always selected. The adaptive marginal curve fitting can fit more than one slice for X_2 so the influence of this variable on Y can be detected. When the predictors are dependent, the marginal G^2 is not easy to compute and it is not straight forward to 'remove' the impact of one single predictor, so we suggest the adaptive marginal curve fitting when the predictors are independent with small p.

Table 2.2: The average ISEs with $t = 0$ for h_j $(j = 1, \ldots, 4)$

n	100	225	400
MGS-AGL	2.46 (1.32)	1.23 (0.77)	0.63 (0.42)
MGS-BIC	2.04 (1.03)	1.00 (0.68)	0.49 (0.38)
AMGS-AGL	1.50 (1.07)	0.62 (0.25)	0.42 (0.16)
AMGS-BIC	**1.26 (1.13)**	**0.45 (0.21)**	**0.29 (0.15)**
SPAM	2.00 (0.77)	0.95 (0.38)	0.63 (0.20)
AGL	1.49 (0.61)	0.59 (0.17)	0.39 (0.10)

Table 2.3: The average of false negative and positive selections with $t = 0$ for h_j $(j = 1, \ldots, 4)$

	false negative			false positive		
n	100	225	400	100	225	400
MGS-AGL	0.66	0.41	0.18	0.02	0	0
MGS-BIC	0.61	0.39	0.17	0.30	0.15	0.08
AMGS-AGL	0.27	0	0	0.04	0	0
AMGS-BIC	0.24	0	0	0.31	0.16	0.10
SPAM	0.01	0	0	3.64	4.05	4.29
AGL	0.24	0	0	0.09	0.01	0

2.4.4 A variable selection example

In this section, we use the same predictors as in Section 2.4.3 and the following relationships:

$$
\begin{aligned}
k_1(x) &= 2.25\sin(4\pi x^2), \quad k_2(x) = 5x, \quad k_3(x) = 5(2x-1)^2, \\
k_4(x) &= -187.5(x-0.2)(x-0.3)(x-0.5)(x-0.9).
\end{aligned}
$$

When $X \sim U(0,1)$, $\text{var}\{k_j(X)\} = 1$ $(j = 1, \ldots, 4)$. Let $e \sim \mathcal{N}(0,1)$ so that the signal to noise ratio is 9. Figure 2.4 shows an example with $n = 100, p = 10$ and $t = 0$. The dots are the observations, the bold lines are the fitted curves by AMGS-BIC and the gray lines are the real curves. Each curve is standardized for a better illustration. It is obvious that the fitted curves almost lie on the true curves, which means that the fitting method can capture the relationship between the predictors and the response almost exactly.

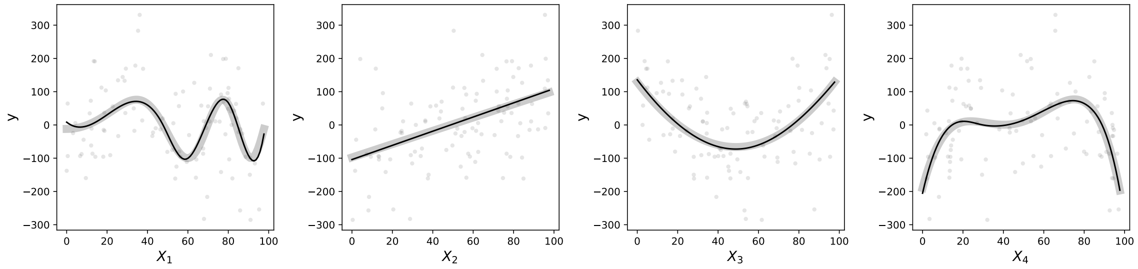

Figure 2.4: The example in Section 2.4.4 and the fitted curves by AMGS-BIC with $n = 100, p = 10$ and $t = 0$. The dots are the observations, the bold lines are the fitted curves and the gray lines are the real curves.

For a better study of our method, we choose the number of sample size as $n = 100,\ 225,\ 400$ and the number of predictors as $p = 10,\ 20,\ 50,\ 100,\ 200$. For each combination of (n,p), we perform 1,000 repetitions. In the main text, we only consider the uncorrelated cases with $t = 0$. We present the results for the correlated cases in Section B.2.

2.4.4.1 Small p

In this part, we only consider $p = 10$ and 20. We implement AMGS-AGL and AMGS-BIC together with SPAM and AGL. Table 2.4 tells us that among all the methods, AMGS-BIC has the smallest ISE. Compared with -AGL method, -BIC method has smaller bias when the method picks the correct variables. The table shows that with larger sample size, our methods performs better.

Table 2.4: The average ISEs (small p) with $t = 0$ for k_j $(j = 1, 2, 3, 4)$

		$p = 10$	
n	100	225	400
AMGS-AGL	1.14 (2.56)	0.39 (0.22)	0.23 (0.14)
AMGS-BIC	1.10 (3.46)	**0.29 (0.20)**	**0.16 (0.12)**
SPAM	1.75 (0.41)	1.07 (0.20)	0.85 (0.12)
AGL	**0.85 (0.39)**	0.35 (0.11)	0.22 (0.07)
		$p = 20$	
n	100	225	400
AMGS-AGL	1.23 (1.12)	0.38 (0.22)	0.23 (0.12)
AMGS-BIC	1.31 (1.52)	**0.30 (0.19)**	**0.16 (0.11)**
SPAM	1.95 (0.47)	1.13 (0.20)	0.89 (0.12)
AGL	**0.84 (0.36)**	0.34 (0.11)	0.22 (0.07)

Now let us discuss the variable selection performance. We present the average number of false negative and false positive selections of each method in Tables 2.5. The results show that SPAM has high false positive selection and the variable selection of our methods, especially AMGS-AGL, are more consistent. We can use -BIC for smaller ISE and -AGL for consistent variable selection result.

2.4.4.2 Large p

Now we consider $p = 50,$ 100 and 200 and implement MGS-AGL together with SPAM and AGL. Similarly, we compare the average ISEs as well as false negative and positive selections of predictors. Table 2.6 tells us that the average ISEs of AMG-AGL are slightly

Table 2.5: The average of false negative and positive selections (small p) with $t = 0$ for k_j $(j = 1, 2, 3, 4)$

false negative	$p = 10$			$p = 20$		
n	100	225	400	100	225	400
AMGS-AGL	0.06	0	0	0.11	0	0
AMGS-BIC	0.07	0	0	0.14	0	0
SPAM	0	0	0	0	0	0
AGL	0	0	0	0	0	0
false positive	$p = 10$			$p = 20$		
n	100	225	400	100	225	400
AMGS-AGL	0.03	0	0	0.10	0	0
AMGS-BIC	0.32	0.17	0.11	0.77	0.47	0.27
SPAM	3.75	3.90	4.25	7.34	7.70	8.25
AGL	0.08	0.01	0	0.11	0.03	0

larger than those of AGL. As the sample size grows, the average ISEs approach those of AGL. Tables 2.7 shows the average false negative and false positive selections of the predictor variables and MGS-AGL has the best consistency performance.

Table 2.6: The average ISEs (large p) with $t = 0$ for k_j $(j = 1, 2, 3, 4)$

		$p = 50$	
n	100	225	400
MGS-AGL	1.63 (1.25)	0.56 (0.46)	0.32 (0.21)
SPAM	2.19 (0.54)	1.21 (0.21)	0.92 (0.12)
AGL	**0.87 (0.35)**	**0.33 (0.11)**	**0.21 (0.07)**
		$p = 100$	
n	100	225	400
MGS-AGL	1.62 (1.43)	0.58 (0.45)	0.32 (0.22)
SPAM	2.39 (0.57)	1.27 (0.21)	0.95 (0.13)
AGL	**0.88 (0.45)**	**0.33 (0.11)**	**0.22 (0.07)**
		$p = 200$	
n	100	225	400
MGS-AGL	1.68 (1.35)	0.57 (0.45)	0.33 (0.22)
SPAM	2.68 (0.69)	1.34 (0.23)	0.98 (0.13)
AGL	**1.01 (0.74)**	**0.33 (0.11)**	**0.21 (0.07)**

To conclude, our methods perform better with larger sample size ($n \geq 225$). When the number of predictors is not large, if the variables are independent, we suggest the AMGS-

Table 2.7: The average of false negative and positive selections (large p) with $t = 0$ for k_j $(j = 1, 2, 3, 4)$

false negative	$p = 50$			$p = 100$			$p = 200$		
n	100	225	400	100	225	400	100	225	400
MGS-AGL	0.20	0	0	0.21	0	0	0.22	0	0
SPAM	0	0	0	0	0	0	0	0	0
AGL	0	0	0	0	0	0	0.04	0	0
false positive	$p = 50$			$p = 100$			$p = 200$		
n	100	225	400	100	225	400	100	225	400
MGS-AGL	0.24	0.02	0	0.40	0.07	0	0.77	0.11	0.01
SPAM	12.91	14.18	14.65	17.44	19.34	19.65	21.31	25.36	26.19
AGL	0.15	0.16	0.04	0.20	0.32	0.06	0.20	0.39	0.11

BIC for a smaller prediction error; otherwise we suggest MGS-BIC. When the number of predictors is large, we suggest MGS-AGL for a consistent variable selection.

2.4.5 Boston housing data

We consider the Boston housing price data from Harrison & Rubinfeld[18] with $n = 506$ observations for the census districts of the Boston metropolitan area. The data is available in the R-package 'lmbench'. We choose ten continuous predictors to predict 'medv'. We perform AMGS-BIC and select six relevant predictors, 'nox', 'rm', 'dis', 'tax', 'ptratio' and 'lstat'. The irrelevant predictors are 'crim', 'indus', 'age' and 'b'. We use $\lambda_0 = 3$ here because all the predictors have very large marginal G^2_{m}. Figure 2.5 shows the fitted curves for the selected variables. The gray dots are the real observations and the black lines are the fitted curves. The predictors are linearly transformed to the interval $[0, 1]$.

2.5 Conclusions

In this chapter, we use the G^2_{m} estimator to perform curve fitting and additive modeling. In fact, the G^2_{m} estimator already fits a piecewise linear curve between the two random variables. When the true relationship is linear, the G^2_{m} estimator can easily identify it. The knots by the G^2_{m} estimator can be treated as the knots for spline curve fitting. Simulation

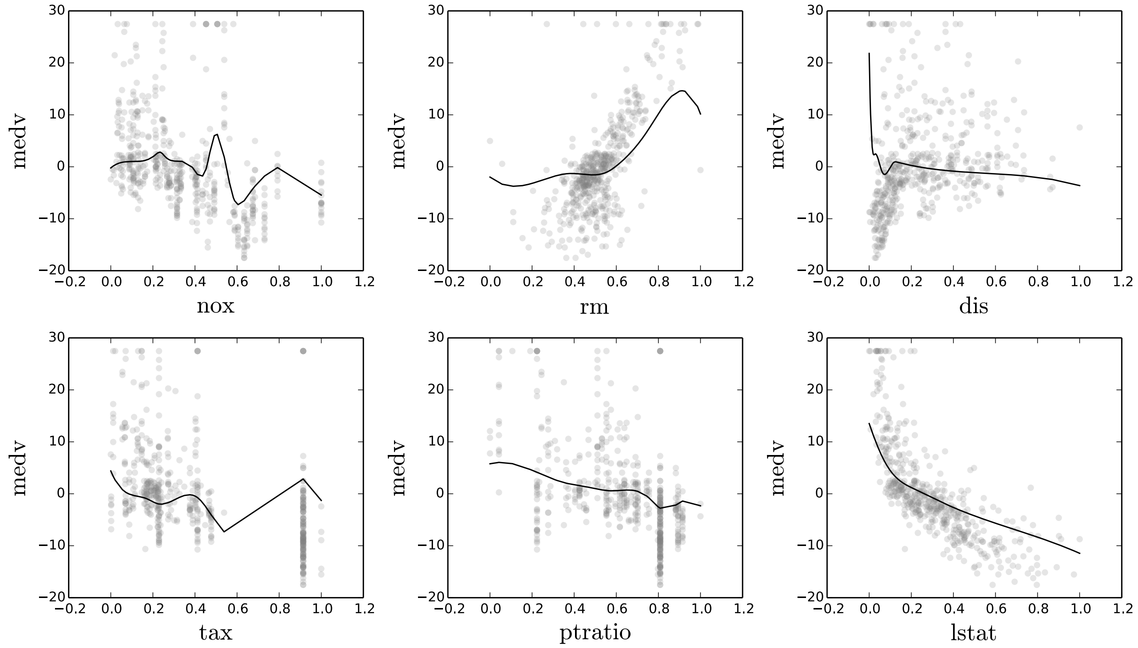

Figure 2.5: The fitted curves of ‘nox’, ‘rm’, ‘dis’, ‘tax’, ‘ptratio’ and ‘lstat’ for Boston housing data. The gray dots are the real observations and the black lines are the fitted curves.

studies show that this approach is robust to the underlying function relationship compared with plain spline fitting. We also discuss how to select λ_0 and the data-driven strategy produces ISE that is quite close to the oracle ISE.

For additive fitting, we suggest two methods to find knots for each predictor, the simple marginal curve fitting and the adaptive marginal curve fitting. When the number of predictors is small and the predictors are independent, the adaptive method has smaller ISE and detects predictors with extremely small marginal G^2’s. When the number of predictors is large, the adaptive method introduces error itself so we prefer the simple marginal fitting method. For variable selection, we try the adaptive group LASSO and the BIC approaches. When the number of predictors is small, we suggest the BIC approach; when the number of predictors is large, we suggest the adaptive group LASSO approach.

In the future, we need more theoretical understandings of the variable selection nature for both methods. Besides, for the adaptive group LASSO approach, we select λ with five-fold cross validation. Ravikumar et al.[39] utilized two heuristic estimates of the risk to choose the regularization parameters. This can also be a future method to tune λ. Simulation studies show that when the sample size is small, the variable selection performance is not as good as other methods. Similarly, the performance of G^2_{m} in independence testing is slightly worse than DDP with $n = 50,\ 100$ (see Section A.4). This is because when n is small, the G^2_{m} estimator sometimes fails to find the change points. A potential solution is to change the penalty term. For example, we already use a penalty that resembles BIC. We can try Akaike information criterion[1] (AIC) as an adjustment because AIC penalizes less than BIC on the log-likelihood.

3

Calibration concordance by multiplicative shrinkage

3.1 Introduction

The calibration of instruments is fundamental if measurements obtained with different instruments are to be compared or combined. In many settings, calibration is based on a data set obtained by using several instruments to measure one or more well-understood sources. The goal is to derive adjustments that can be applied to future observations for reliable absolute measurements. Direct linear adjustments, however, often result in poor calibration of the instruments, no quantification of the calibration error, and hence they do not permit researchers to assess the effect of calibration uncertainty on final estimates. The main difficulty of deriving reliable adjustments for instruments springs from the variations intrinsic to the sources and instruments along with individual measurement errors.

Several complications arise when attempting to properly modeling a calibration data set. First, the physical models, which are derived using various approximations based on scientists' current understandings of the instruments, are not exact. Second, known physical quantities are typically estimates themselves and even when their estimated errors are available, standard plug-in estimators and error propagation techniques may lead to biased or overly optimistic results. Third, data quality varies in ways that cannot be fully quantified, especially across instruments or in the presence of outliers. Finally, the number of unknown model parameters increases with both the number of instruments and the number of sources, leading to well-known model challenges. Together these challenges and subtleties explain that although many researchers have worked on the calibration problem, principled statistical adjustments have yet to be developed.

In this chapter, we resolve these challenges by first introducing a multiplicative observation model and then developing an approximate log-normal approach to model the mean signals for each source measured by each instrument while considering measurement errors. Furthermore, because the number of parameters grows with both the number of instruments and sources, the model fitting requires advanced Bayesian computational algorithms. Lastly, we propose a more general log-t model to handle the outliers that are often present in such data.

3.1.1 Calibration in astronomical instruments

In astrophysics, various instruments are used by different teams of scientists to understand intrinsic properties of astronomical objects. Although it is possible to make relative comparisons of different sources observed with the same instrument, unless the instruments are properly calibrated, we cannot make reliable absolute measurements or make comparisons of sources observed with different instruments. Therefore, calibration of different instruments is an important, and on-going, problem for astrophysicists.

To perform in-flight calibration, a set of well understood sources are observed with multiple instruments to derive adjustments that can be made to future observations and obtain reliable absolute measurements. Deriving adjustments for astronomical instruments based on observing multiple sources with multiple instruments is defined as the calibration concordance problem. This chapter is motivated by the need for a statistically principled solution and is a joint effort between astrophysicists and statisticians, both expertise is needed to appropriately quantify the uncertainties while properly incorporating scientific understanding. This chapter describes the general calibration problem in terms of its manifestation with astronomical instruments.

First, two basic concepts that are essential to describe precisely the scientific question are the flux of each astronomical source and the effective area of each instrument.

- Flux: the absolute flux is the quantity of luminous energy incident upon the aperture of a telescope per unit area per unit time. The absolute flux of an astronomical source depends on the luminosity of the object and its distance from the earth, both of which are intrinsic to the object. For a fixed source spectrum, i.e., the distribution of photon energies, the measured flux is directly proportional to the number of photons detected in each detector on an astronomical instrument. If the spectrum changes, or the detector on the instrument changes, then so will the number of photons and the measured flux.

- Effective Area: the geometric area of an instrument is an upper bound on its capacity to collect photons. Many factors can reduce the efficiency of photon collection, including mirror reflectivity, structural obscuration, filter transmission, detector sensitivity, etc. This reduction in efficiency is also photon-energy dependent. The effective area is the equivalent geometric size of an ideal detector that would have the same collection capability and it is empirically measured or theoretically calculated and tabulated as a function of energy. The effective area of the instrument is used to estimate the absolute flux of an astronomical source given its measured flux. Since the effective area varies with energy, astronomers often consider different energy bands for comparing observations as different instruments. We will adopt the same convention.

Second, the calibration problem arises because the effective areas of the instruments are not known precisely, and thus absolute measurements of the flux of an astronomical source cannot be obtained: different instruments yield different measured fluxes for the same source. In other words, the problem of calibration among different instruments is equivalent to estimating the effective area of each instrument.

Astronomers may use several instruments with different and uncertain effective areas to measure the fluxes of astronomical sources. The measurements are the numbers of photons from each object received on each detector. Since we do not know the effective areas precisely, we aim to improve them using data for common sources. After proper adjustments of the effective areas, instruments measuring a common source should agree within statistical uncertainty on the absolute flux of each astronomical source.

3.1.2 Notations and a multiplicative physical model

Suppose we observe photon counts, c_{ij} $(i = 1, \ldots, N,\ j = 1, \ldots, M)$, where i indexes the instruments and j indexes the sources. We denote the expected photon counts by C_{ij} for each instrument-source pair. Here and elsewhere we use lower case for observed quantities and estimators and upper case for unobserved estimands. Let F_j be the absolute flux of source j. To estimate F_j from the observed photon counts, we need the effective area A_i of each instrument. An estimate of each effective area a_i is obtained through the knowledge of the instrument designers, but we assume that we do not have information for estimating each F_j other than the c_{ij}.

The observed photon counts depend on two factors: the absolute flux of the source and the effective area of the instrument. Because source fluxes have units of photons per second and per square centimeter, they are multiplied by instrument effective areas and exposure times T_{ij} to obtain expected photon counts. Thus, the multiplicative model is

$$C_{ij} = T_{ij} A_i F_j \ (i = 1, \ldots, N,\ j = 1, \ldots, M). \tag{3.1}$$

Although we omit details here, the multiplicative constant T_{ij} contains not only the exposure time, but also other factors that can be calculated approximately by astrophysicists; see Marshall et al.[32] for details. We regard T_{ij} as a fixed constant for now and the uncertainties related to T_{ij} will be considered and discussed for further improvements.

Generally, astronomical effective areas come with estimated systematic uncertainties given by the calibration scientists based on their empirical knowledge about each instrument as well as statistical uncertainties based on the assumed Poisson nature of the detected light. With a typical dataset, astronomers provide their estimated uncertainties for each of the c_{ij} and for the a_i. How to utilize both uncertainties in statistical models and whether these estimated uncertainties suffice to explain the variance in the data are intriguing statistical questions that we also seek to tackle in this chapter.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce our statistical model for calibration concordance, a log-normal hierarchical model, followed by its properties and extensions to a more general log-t model, which handles outliers. In Section 3.3, we illustrate model fitting results with both simulated and real data. We conclude in Section 3.4 with a brief discussion on the frequentist equivalence of the model, a summary, and areas of future works.

3.2 Building and fitting calibration models

3.2.1 Modeling multiplicative means

We can rewrite (3.1) as

$$\log C_{ij} - \log T_{ij} = \log A_i + \log F_j = B_i + G_j, \tag{3.2}$$

where $B_i = \log A_i$ and $G_j = \log F_j$. Although (3.2) holds at the estimand level, the corresponding estimator/observation equation does not. Specifically, if we let $y_{ij} = \log c_{ij} - \log T_{ij}$, $b_i = \log a_i$ and $g_j = \log f_j$, we cannot expect that $y_{ij} = b_i + g_j + e_{ij}$ and that e_{ij} is independent of $(b_i,\ g_j)$ with mean zero. This is because this observation equation incorrectly assumes that the expectation of y_{ij} is determined by b_i and g_j, while they are estimators of B_i and G_j.

Instead, we assume that the measurement error in c_{ij} for C_{ij} is multiplicative and postulate the regression model,

$$y_{ij} = \alpha_{ij} + B_i + G_j + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma_i^2); \tag{3.3}$$

where y_{ij} is obtained from quantities that are either observed (c_{ij}) or supplied (T_{ij}). Nevertheless, when $c_{ij} = 0$ occurs (as in out simulation study), we will use the conventional 0.5. We call this the zero-modified count and denote as $\tilde{c}$. The measurement error e_{ij} is independent Gaussian with mean 0 and variance σ_i^2. Thus, the observed counts c_{ij} follows a log-normal distribution. We define $\alpha_{ij} = -0.5\sigma_i^2$ as the half-variance correction for the multiplicative mean modeling in (3.1) on the log scale to ensure that $Ec_{ij} = C_{ij}$, because

$$Ec_{ij} = T_{ij}Ee^{y_{ij}} = T_{ij}e^{\alpha_{ij}+0.5\sigma_i^2}e^{B_i}e^{G_j} = T_{ij}A_iF_j = C_{ij}.$$

For convenience, when the σ_i^2 are known, we define $y'_{ij} = y_{ij} + \alpha_{ij}$ and use this notation throughout the chapter.

Recall that $b_i = \log a_i$ is estimated with uncertainty based on expert empirical knowledge of instrument i, thus we can view b_i as a noisy observation of B_i with noise level τ_i, i.e.

$$b_i \sim \mathcal{N}(B_i, \tau_i^2). \tag{3.4}$$

Together with (3.3), this gives a frequentist random-effect regression model, which is discussed in Section 3.4.1. We assume the τ_i are known from expert knowledge as they are in our applied examples.

We adopt a Bayesian perspective, i.e. we reverse the roles of b_i and B_i in (3.4) and assume

$$B_i \sim \mathcal{N}(b_i, \tau_i^2). \tag{3.5}$$

This reversal can be justified more formally by using (3.4) along with a flat prior on the B_i, which implies (3.5). There are three advantages of a Bayesian perspective in this setting: 1) by characterizing empirical knowledge in a prior distribution, we can update the prior information with the observed data and give a full posterior distribution of the quantities of interest, which yields automatic uncertainty quantification for the estimators; 2) the maximum-a-posteriori (MAP) estimators obtained from the hierarchical model are shrinkage estimators, which intuitively summarize how information from various instruments and sources are best combined; 3) since the dimension of the parameter space is large and the parameters of interest are highly correlated, Bayesian computational methods, such as the Markov chain Monte Carlo (MCMC) algorithms, are better suited for exploring the parameter space than optimization algorithms.

In setting up our models, we have made several approximations and simplifications. First, the observations collected by astronomical instruments are in fact photon counts, which are usually modeled with a Poisson distribution. Since the observed photon counts in our real examples are typically large, the Gaussian model is a good approximation. Furthermore, we prefer the Gaussian model because both its mean and variance are free parameters which permits the variance term to accommodate imperfections in the mean model. We will provide a detailed discussion of the Gaussian approximation to a Poisson model in our numerical experiments in Section 3.3.1.2.

In (3.3), we assume that the variance for the measurement error depends on the instrument but not on the source. This assumption works reasonably well in our applied examples, but generally, each e_{ij} can have its own variance, σ_{ij}^2, i.e. $e_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$. If the σ_{ij}^2 are unknown, it is generally necessary to constrain them to ensure identifiability. For example, we can assume either that 1) the variance is only object-dependent, i.e. $\sigma_{ij}^2 = \sigma_j^2$ or 2) the variance is additive, i.e. $\sigma_{ij}^2 = \omega_i^2 + \lambda_j^2$. If the σ_{ij}^2 are known, inference is much easier since the B_i's and the G_j's are the only unknown quantities. Hereafter, we refer to the case when the σ_{ij}^2's are known as the known variance model and the case when the σ_i^2's are unknown as in (3.3) as the unknown variance model.

We consider the known variance model because, as noted in the introduction, astronomers provided their estimate of the uncertainties of measurements, i.e., the σ_{ij}^2's. However, as illustrated in subsequent sections through both simulated and real data, the unknown variance model, which is the primary model in this chapter, is more flexible, robust, and is recommended in practice. This is because the inferred adjustment of effective areas could be either overly-optimistic or overly-conservative if the specified σ_{ij}^2's are inaccurate, which is often the case in practice due to an incomplete understanding of the uncertainties in measurements and data processing.

Finally, because not all sources may be observed with all instruments, we define J_i to be the set of indexes of the objects observed by detector i and I_j the set of indexes of the instruments that observe object j. If the set of objects observed with each instrument reflect a biased selection mechanism, our model may produce misleading results. As a first approximation, we assume there is no selection bias or that the selection mechanism is ignorable[45].

3.2.2 Log-normal hierarchical model

In this section, we embed the log-normal regression model given in (3.3) into a Bayesian hierarchical model. To do this, we need prior distributions for B_i's, G_j's, and σ_i^2's when they are unknown. Since we have no prior information for G_j's, we use independent flat priors on the real line whereas we assume (3.5) for B_i's. When σ_i^2's are unknown, we assume independent Inverse-Gamma distributions with degree of freedom df_g and scale β_g. In summary, the log-normal hierarchical model is written as

$$\begin{aligned} y_{ij} \mid B_i,\ G_j,\ \sigma_i^2 &\sim \mathcal{N}\left(-\frac{\sigma_i^2}{2} + B_i + G_j,\ \sigma_i^2\right), \\ \sigma_i^2 &\sim \text{Inv-Gamma}(df_g,\ \beta_g), \\ B_i &\sim N(b_i,\ \tau_i^2),\ G_j \sim \text{ flat prior.} \end{aligned} \tag{3.6}$$

Let $B = (B_1, \ldots, B_N)^t$, $G = (G_1, \ldots, G_M)^t$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_N^2)^t$, and $\tau^2 = (\tau_1^2, \ldots, \tau_N^2)^t$, where x^t denotes the transpose of a vector x. Let $D = \{y_{ij}, b_i\}$ represent the data and $\psi = \{\sigma^2, \tau^2\}$ be the variance parameters.

Under the prior specifications for $\{B_i, G_j, \sigma_i^2 \mid i = 1, \ldots, N, j = 1, \ldots, M\}$ given in (3.6), we can show that: the posterior is proper when all instruments measure all sources, i.e., $|J_i| = M\,(i = 1, \ldots, N)$, and the MAP estimator of each σ_i^2 is bounded away from zero by a finite constant which only depends on the hyper-parameters. Furthermore, this prior specification avoids the problem of unbounded posterior distribution, that would arise with flat priors on the σ_i^2. The proofs of these claims are in Section C.2.

3.2.3 Posterior distributions and their sampling

A special case of (3.6) occurs when the variances σ_i^2 and τ_i^2 are known. The logarithm of the joint posterior distribution of B and G conditioning on ψ is

$$L\,(B, G \mid \psi) = -\sum_{1 \le i \le N, j \in J_i} \frac{(y'_{ij} - B_i - G_j)^2}{2\sigma_i^2} - \sum_{i=1}^{N} \frac{(b_i - B_i)^2}{2\tau_i^2}. \tag{3.7}$$

This is a quadratic function of the B_i and G_j, thus the joint posterior of B and G is multivariate Gaussian. More precisely, the posterior distribution of (B, G) is a multivariate Gaussian with mean $\mu = \Omega^{-1}\gamma$ and variance-covariance matrix Ω^{-1}, where Ω is an $(N+M) \times (N+M)$ matrix with

$$\Omega_{i,i} = M_i\sigma_i^{-2} + \tau_i^{-2}, \quad \Omega_{j+N,j+N} = \sum_{i \in I_j} \sigma_i^{-2}, \quad \Omega_{i,j+N} = \sigma_i^{-2} I_{j \in J_i};$$

and γ is a column vector of length $(N+M)$ with

$$\gamma_i = \left(\sum_{j \in J_i} y'_{ij}\right) \sigma_i^{-2} + b_i\tau_i^{-2}, \quad \gamma_{j+N} = \sum_{i \in I_j} y'_{ij}\sigma_i^{-2};$$

$M_i = |J_i|$ is the number of elements in J_i $(1 \leq i \leq N,\ 1 \leq j \leq M)$.

When the variances are unknown, numerical techniques are required to explore the joint posterior distribution. Since the dimension of the parameter space, $(2N + M)$, is typically large and the parameters are highly correlated, we use a Hamiltonian Monte Carlo (HMC) algorithm[37], which delivers a less correlated sample than more traditional MCMC techniques[34,20,14,31]. We implement HMC using the STAN package[23,49] in *Python*[50].

We have also implemented a blocked Gibbs sampler, which gives satisfactory performance and enables us to crosscheck the results from STAN. In the blocked Gibbs sampler, we sample the B_i and G_j jointly to improve mixing: since as we just noted in Section 3.2.3, they jointly follow a multivariate Gaussian distribution conditioning on ψ. This is much more efficient than one-parameter-at-a-time Gibbs sampling in both our simulated and real data examples. Section C.1 gives details of the computational algorithms we adopted.

We now derive the MAP estimators, which are shrinkage estimators[8,35] of the B_i and the G_j, and thus correspond to power shrinkage on the original scale, i.e., A_i and F_j. The shrinkage estimators enjoy the intuitive interpretation of combining information among all the instruments and sources, which well serves the purpose of calibration concordance across instruments and sources.

To derive the MAP estimators conditioning on ψ, we set the derivative of the log-posterior in (3.7) to be zero. The conditional MAP estimators, denoted by $\widehat{B}_i(\psi)$ and $\widehat{G}_j(\psi)$, satisfy

$$\widehat{B}_i(\psi) = W_i b_i + (1 - W_i)(\bar{y}'_{i\cdot} - \bar{G}_i), \quad \widehat{G}_j(\psi) = \bar{y}'_{\cdot j} - \bar{B}_i, \tag{3.8}$$

where $\bar{G}_i$ is the precision weighted average of the $\widehat{G}_j(\psi)$ over $j \in J_i$ and $\bar{B}_j$ is the precision weighted average of the $\widehat{B}_i(\psi)$ over $j \in J_i$, i.e.

$$\bar{G}_i = \frac{\sum_{j\in J_i} \widehat{G}_j(\psi)\sigma_i^{-2}}{\sum_{j\in J_i} \sigma_i^{-2}}, \quad \bar{B}_j = \frac{\sum_{i\in I_j} \widehat{B}_i(\psi)\sigma_i^{-2}}{\sum_{i\in I_j} \sigma_i^{-2}},$$

$\bar{y}'_{i\cdot}$ is the precision weighted average of the y'_{ij} over $j \in J_i$, and $\bar{y}'_{\cdot j}$ is the precision weighted average of the y'_{ij} over $i \in I_j$, i.e.

$$\bar{y}'_{i\cdot} = \frac{\sum_{j\in J_i} y'_{ij}\sigma_i^{-2}}{\sum_{j\in J_i} \sigma_i^{-2}}, \quad \bar{y}'_{\cdot j} = \frac{\sum_{i\in I_j} y'_{ij}\sigma_i^{-2}}{\sum_{i\in I_j} \sigma_i^{-2}},$$

and the weights,

$$W_i = \frac{\tau_i^{-2}}{\tau_i^{-2} + M_i\sigma_i^{-2}},$$

are the precisions of the direct information in the b_i relative to the indirect information for estimating the B_i. Thus W_i can be regarded as the proportion of information from the prior. From studying this, we can make more informative choices for the prior variances τ_i^2 when we do not have precise values of τ_i^2 from experts. In our real applications, we can choose reasonable values of the τ_i^2 by examining this quantity, to ensure our results are largely data-driven rather than prior-driven. We elaborate on this in the data analysis in Section 3.3.2.1, with detailed results given in Section C.5.

When the σ_i^2 are unknown, we use independent conjugate priors for the σ_i^2 as in Section 3.2.2. In this case, the MAP estimators satisfy both (3.8) and

$$\widehat{\sigma}_i^2 = 2\left(\sqrt{1 + S_{y,i}^2} - 1\right), \quad S_{y,i}^2 = \frac{1}{M_i + df_g}\left(\sum_{j\in J_i}(y_{ij} - \widehat{B}_i - \widehat{G}_j)^2 + \beta_g\right). \tag{3.9}$$

We simultaneously solve (3.8) and (3.9) numerically and denote the MAP estimators that solve these equations by $\{\widehat{B}_i, \widehat{\sigma}_i^2, \widehat{G}_j\}$. Because of the log transformation in the regression in (3.3), the estimate for the variance in (3.9) is on the same scale as the data and the mean. This seeming contradiction can be explained by observing that the data are already on the log-scale.

Intriguingly, we notice that the MAP estimator for the variance is also of a shrinkage form. This is most clearly seen by re-expressing (3.9) as

$$\widehat{\sigma}_i^2 = 2\left(\sqrt{1+S_{y,i}^2}-1\right) = \frac{2}{1+\sqrt{1+S_{y,i}^2}}\, S_{y,i}^2 \equiv R_i S_{y,i}^2, \tag{3.10}$$

where $S_{y,i}^2$ as defined in (3.9) is similar to the natural residual variance estimator for σ_i^2, except for the prior distribution. The half-variance correction leads to a shrinkage of $S_{y,i}^2$ because R_i is bounded above by 1. The degree of shrinkage depends on $S_{y,i}^2$ itself. The larger $S_{y,i}^2$ is, the smaller R_i is, and the more shrinkage there is in the estimator of σ_i^2.

3.2.4 Extensions to handle outliers: log-t model

The framework of Section 3.2.2 assumes Gaussian noise on the log scale and the half-variance correction depends on this Gaussian assumption. However, taking logs is not enough to get rid of some extreme outliers, which are not rare in astronomical observations since some sources can be quite dim, causing the photon collection highly imprecise. Therefore, the log-normal hierarchical model may not be robust to outliers. Here we propose a generalization of the log-normal model by introducing a unique variance σ_{ij}^2 for each observation y_{ij}. In this way, we can down weight outliers for more robust inference.

For any observation y_{ij}, we assume

$$y_{ij} \mid B_i,\ G_j,\ \sigma_{ij}^2 \quad = \quad -0.5\sigma_{ij}^2 + B_i + G_j + e_{ij}, \quad e_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2). \tag{3.11}$$

Under (3.11), $E(e^{y_{ij}} \mid B_i,\ G_j) = E\left\{E(e^{y_{ij}} \mid B_i,\ G_j,\ \sigma_{ij})|B_i, G_j\right\} = A_iF_j$, so the multiplicative model (3.1) still holds. Depending on the assumptions we place on σ_{ij}'s, (3.11) includes the following cases:

> *Case 1*: *log-normal model with known variances*. If the σ_{ij} are known constants, the noise terms are independent Gaussians with mean 0 and variance σ_{ij}^2. Thus the model in (3.11) is equivalent to (3.6) with known variances.
>
> *Case 2*: *log-normal model with unknown variances*. If $\sigma_{ij}^2 = \sigma_i^2 \sim \text{Inv-Gamma}(df_g,\ \beta_g)$ for all j, the model in (3.11) is equivalent to (3.6) when the variances ψ are unknown.
>
> *Case 3*: *log-t model*. If $\sigma_{ij}^2 \sim \text{Inv-Gamma}(df_g,\ \beta_g)$, i.e. independent Inv-Gamma distribution for all i, j. The error term e_{ij} follows independent student-t distributions, with degree of freedom $2df_g$ and scale $\sqrt{\beta_g/df_g}$.

The fitting of the log-t model in Case 3 is also achieved with the HMC algorithm using the STAN package.

It is worth emphasizing that besides down weighting outliers, Case 3 also permits a unique variance for each instrument-source combination, which is impossible for the log-normal regression model (3.6) where the observational noise is only instrument dependent. Therefore, the log-t model is more flexible than the log-normal model, but with a price of more computational cost: the dimension of the parameter space for the log-t model is higher than for the log-normal hierarchical model. Thus, the convergence of the HMC algorithm is harder to achieve and the sampling takes longer. A potential limitation of the log-t model is when σ_{ij} is large, it is very likely that the half variance correction $0.5\sigma_{ij}^2$ gives a larger value than the absolute value of the error e_{ij}, considering the relative order in terms of σ_{ij}^2. This results in y_{ij} being very small, i.e., the model is more likely to generate small outliers as opposed to large outliers. We verify this numerically by simulating data from the model and examining the left and right tails of the observations.

We demonstrate the effectiveness of the log-t model compared with the log-normal hierarchical model for simulated and real data in Sections 3.3.1.3 and 3.3.2.4. In general, the log-normal model is robust enough for real applications. Thus, we recommend users to fit the log-normal model before resorting to the more complicated log-t model. However, the log-t model can provide more precise results in the presence of outliers, especially when computational cost is not a concern.

3.3 Examples: simulated and applied results

In Section 3.3.1, we validate the methodology proposed in Section 3.2 through a series of simulation studies. We show that 1) the log-normal hierarchical model works well when the model is correctly specified; 2) under realistic model misspecification, the log-normal hierarchical model can still give valid results; 3) plugging-in known variances given by astronomers can give overly optimistic results which leads to misleading adjustments, especially when there exist unknown uncertainties; 4) the log-t model performs better than the log-normal model in fitting data with outliers. We illustrate model fitting using real data compiled by IACHEC researchers in Section 3.3.2.

3.3.1 Numerical simulations

3.3.1.1 Simulations with correctly specified model

In Simulation I, we simulate from the log-normal model with $N = 10$ instruments and $M = 40$ sources. We set each $B_i = 5$ and each $G_j = 3$, and independently sample $b_i = \log a_i$ from $\mathcal{N}(B_i, 0.05^2)$. The variances are specified as $\sigma_i^2 = 0.1^2$ and $\tau_i^2 = 0.05^2$ for each i. When the σ_i^2 are assumed known, Section 3.2.3 gives the posterior distributions of the B_i and G_j. If, on the other hand, the σ_i^2 are unknown, we specify the priors with $df_g = 2$, $\beta_g = 0.1^2$ and use the HMC algorithm to obtain a Monte Carlo sample from the joint posterior distribution. The results of Simulation I in Fig. 3.1 show that the posterior distributions of the effective areas and variances match the true values and that the posterior distributions of the B_i's are similar regardless of whether the variances are known.

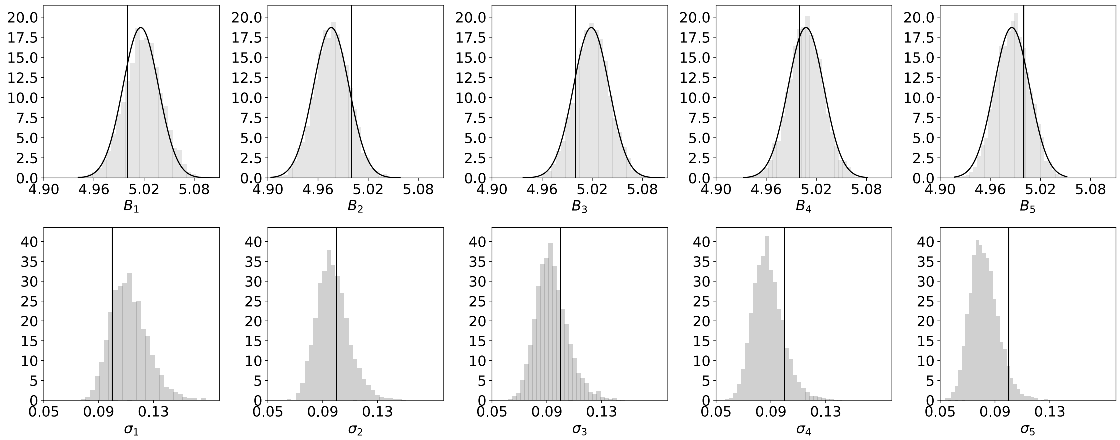

Figure 3.1: Simulation I. Posterior distributions of the B_i (row 1) and the σ_i (row 2) under known and unknown variance scenarios. The gray histograms represent the posterior samples of the B_i and the σ_i with unknown variances. The solid vertical lines are the true values. The solid black density curves on top of the histograms in the first row denote the closed-form posterior densities of the B_i when the σ_i^2 are known (0.1^2).

We also find through simulations in Section C.4.1 that for the same number of instruments, the larger the number of sources, the better the estimated effective areas are but the estimated fluxes may not be better. Whereas for the same number of sources, the larger the number of instruments, the better the estimated fluxes are but the estimated effective areas may not be better.

3.3.1.2 Simulations with misspecified model

There are several approximations we make in the log-normal hierarchical model. Specifically, we model Poisson photon counts using a log-normal distribution and we assume the T_{ij} are known in (3.1). These approximations are justifiable theoretically. Here we study the influences of the former approximation with simulation studies and leave more results of the latter approximation and a combination of both in Section C.4.2.

Suppose more appropriately that $c_{ij} \sim$ Poisson(C_{ij}) with $C_{ij} = A_i F_j$. As mentioned in Section 3.2, the Gaussian assumption is a good approximation to the Poisson model when the counts are large. We can directly verify this using numerical experiments.

In Simulation II and III, we set $N = 10$, $M = 40$, and choose each $\tau_i = 0.05$. We generate the prior mean b_i from $\mathcal{N}(B_i, 0.05^2)$. We use independent inverse gamma priors with shape $df_g = 2$ and rate $\beta_g = 0.1^2$. Besides, in Simulation II, each $B_i = 1$ and $G_j = 1$ while in Simulation III, each $B_i = 5$ and $G_j = 3$. Thus Simulation II represents a low count scenario where the normal approximation may not be appropriate. Applying the delta method to the zero-modified Poisson model $\tilde{c}_{ij}$ gives that $\sigma_{i1}^2 \approx 0.367^2$ and $\sigma_{ij}^2 \approx 0.018^2$ $(i = 1, \ldots, N,\, j = 2, \ldots, M)$. As expected, the fitted values of the B_i and G_j are much better in Simulation III; Figures 3.2 and 3.3 give detailed comparisons under Simulations II and III.

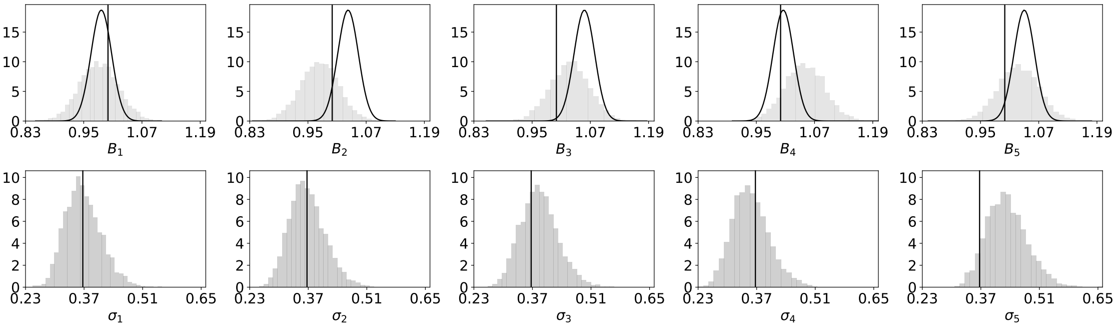

Figure 3.2: Simulations II. The legend is the same as in Fig. 3.1.

Suppose a user plugged in $\sigma_i^2 = 0.1^2$, a hypothetical value of the σ_i^2, as the known variances. Comparing the histograms and the overlying curves from Fig.s 3.2 and 3.3, we see that our model fitting with known variances: 1) is overly optimistic if the specified variances are smaller than the variances estimated under the unknown variance model; 2) is conservative if the specified variances are larger than the variances estimated under the

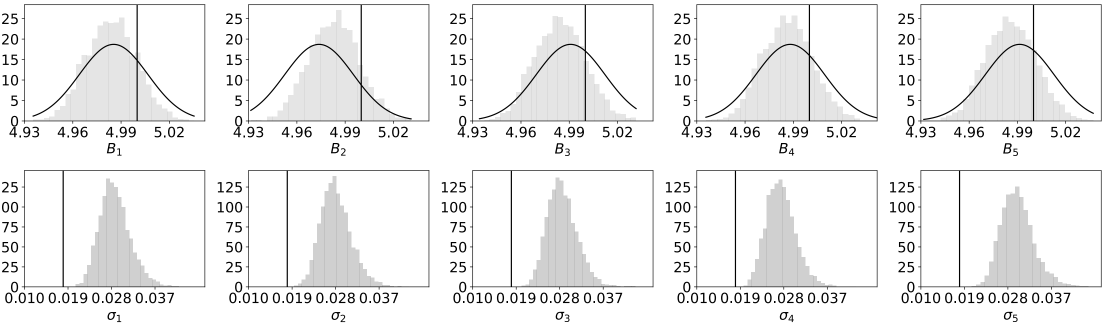

Figure 3.3: Simulations III. The legend is the same as in Fig. 3.1.

unknown variance model.

In summary, when the data is generated from a Poisson model, assuming a hypothetical known variance can possibly be detrimental. If the assumed known variances are not large enough to account for the model misspecification, the estimated effective areas and fluxes can be biased and the posterior coverage can be bad; furthermore, the model gives overly optimistic results – possibly false positive signals. On the contrary, assuming large hypothetical known variances is a much safer choice since the inflated variances consider of the model misspecification. However, larger variances mean less precision, which leads to less informative results.

3.3.1.3 Simulation studies with outliers

We demonstrate the effectiveness of the log-t model in dealing with outliers through Simulation IV. We simulate from a Poisson model with $N = 10$ and $M = 40$. We set each $B_i = 5$ and $G_1 = -2$, $G_j = 3$ $(j = 2, \ldots, M)$. The b_i are independently sampled from $\mathcal{N}(B_i,\, 0.05^2)$. The prior for σ_i^2/σ_{ij}^2 is Inv-Gamma with $df_g = 4$ and $\beta_g = 0.1^2$. Applying the delta method to the zero-modified Poisson model $\tilde{c}_{ij}$ gives that $\sigma_{i1}^2 \approx 0.232^2$ and $\sigma_{ij}^2 \approx 0.018^2$ $(i = 1, \ldots, N,\ j = 2, \ldots, M)$. As discussed in Section 3.2.4, when the true σ_{ij}^2 for some sources are much larger than the others, the corresponding observations are likely, but not necessarily, to be outliers. Thus, in this example, the observations from the first source are likely to be outliers.

Figure 3.4 compares the fitted results of the log-normal model and the log-t model through the standardized residuals,

$$\widehat{\mathcal{R}}_{ij} = \frac{y_{ij} - \widehat{B}_i - \widehat{G}_j + 0.5\widehat{\sigma}^2_{ij}}{\widehat{\sigma}_{ij}} = \frac{y_{ij} - \widehat{B}_i - \widehat{G}_j}{\widehat{\sigma}_{ij}} + 0.5\widehat{\sigma}_{ij}, \tag{3.12}$$

for the observations from the first three sources. Some observations from the first source (blue circles) appear to be outliers, with standardized residuals lying outside the $[-2, 2]$ interval, in the log-normal model but not in the log-t model. In the log-normal model, setting $\sigma^2_{ij} = \sigma^2_i$ causes failure due to some source-dependent large variances: $\sigma^2_{i1} >> \sigma^2_{ij}$ $(j = 2, \ldots, M)$. Because we model σ^2_{ij} separately for the log-t model, the log-t model is more capable of handling outliers than the log-normal model. Figure 3.5 shows the posterior distributions of B_i by the log-normal and log-t model and both the models capture the true value. It is reasonable that the results by the log-t model have slightly larger variance since the log-t model is more flexible.

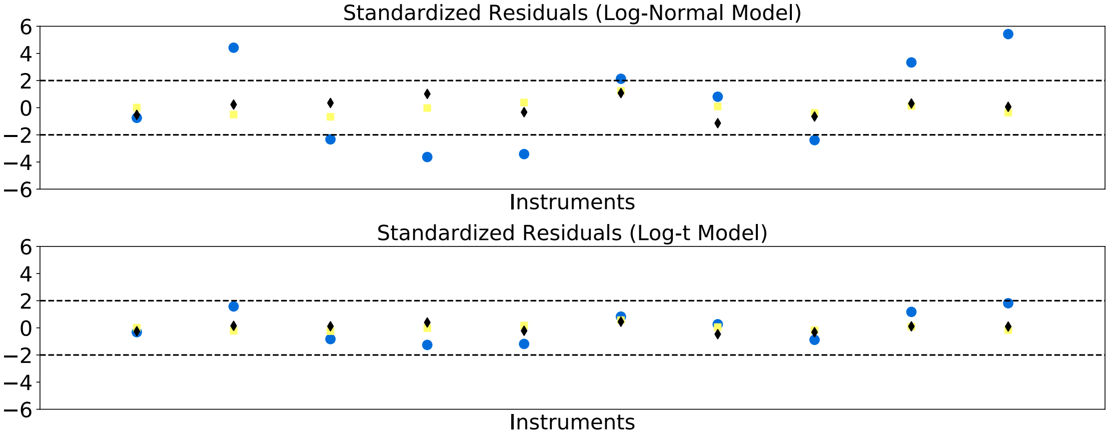

Figure 3.4: Simulation IV. Standardized residuals of the log-normal hierarchical model (row 1) and the log-t model (row 2). The blue circles, yellow squares and black rhombuses represent the first three sources respectively. The instruments are plotted on the x-axes. The dashed horizontal lines denote the $[-2, 2]$ intervals.

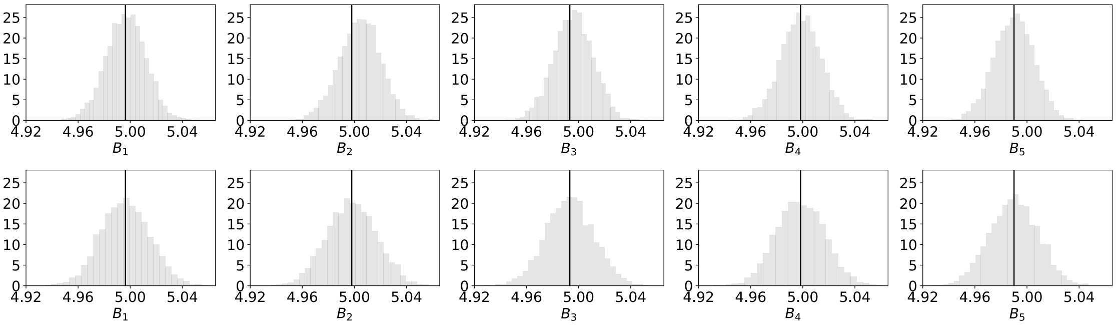

Figure 3.5: Simulation IX. Posterior distributions (gray histograms) of $\{B_i\}_{i=1}^5$ of the log-normal hierarchical model (upper panel) and the log-t model (lower panel).

3.3.2 Data analysis

In this section, we fit the log-normal hierarchical model to three data sets compiled by IACHEC[26] researchers, with the aim of understanding calibration properties of various X-ray telescopes (instruments) such as Chandra, XMM-Newton, Suzaku, Swift, etc. See Marshall et al.[32] for detailed descriptions of the data collection and preprocessing.

3.3.2.1 E0102 data

E0102 is the remnant of a supernova that exploded in a neighboring galaxy known as the Small Magellanic Cloud[4] and is a calibration target for a variety of X-ray missions. We consider four photon sources associated with E0102. Each of the sources is a local peak or line that appears in the E0102 spectrum. (A spectrum can be thought of as a high-resolution histogram of the energies of photons originating from E0102. We consider the photon counts in four bins of this histogram.) Two of the lines are associated with highly ionized Oxygen (Hydrogen Lyman-α like O VIII at 18.969Å and the resonance line of O VII from the He-like triplet at 21.805Å) and the other two are associated with Neon (H-like Ne X at 12.135Å and He-like resonance line Ne IX at 13.447Å). We consider replicate data obtained with 13 different detectors configurations respectively over 4 separate telescopes, *Chandra* (HETG and ACIS-S), XMM-*Newton* (RGS, EPIC-MOS, EPIC-pn), *Suzaku* (XIS), and *Swift* (XRT). Details of how the spectra are preprocessed to obtain relevant line counts can be found in Plucinsky et al.[38]. Because the energies of the two O lines are similar, it is reasonable to assume that the associate effective areas are also similar, likewise for the Neon lines are their effective areas. Therefore, we consider two separate data sets, one with O VII and O VIII treated as its two sources, and the other with Ne IX and Ne X treated as its two sources. The measured fluxes (counts) have been normalized relative to those measured in one of the detectors (RGS1). This is an arbitrary choice: we do not expect that RGS1 represents the ground truth, see Plucinsky et al.[38].

We apply the log-normal hierarchical model in Section 3.2.2 to the two data sets. The hyper-parameters are $df_g = 1.5,\ \beta_g = 0.014^2$ for O VII, O VIII and $df_g = 1.5,\ \beta_g = 0.009^2$ for Ne IX, Ne X. These values are chosen based on empirical knowledge about the measurement uncertainties. We set each $b_i = 0$, i.e., with an expectation that no adjustment is needed across detectors, with confidence τ_i. We use two possible values $\tau_i = 0.025$ and $\tau = 0.05$ according to the empirical knowledge of astronomers to study the influence of the τ_i on the analysis.

Figure 3.6 shows the adjustments of the log-scale effective area for O (row 1) and Ne (row 2) in the E0102 data sets. We can find that the estimated values of the B_i are not sensitive to the choices of the τ_i except for detector XRT-PC. We compute the estimated prior influence as defined in Section 3.2.3, i.e., $\widehat{W}_i = \frac{\tau_i^{-2}}{\tau_i^{-2}+|J_i|\widehat{\sigma}_i^{-2}}$, for XRT-PC in the Ne data; the value is 0.91 when $\tau_i = 0.025$ and 0.02 when $\tau_i = 0.05$. When the prior variance of B_i is too small ($\tau_i = 0.025$), the model treats the observations as being less accurate (by fitting a large σ_i) instead of adjusting the effective area of the corresponding instrument more (a larger deviation from b_i). Figure 3.6 suggests that the effective areas of MOS1, MOS2, XIS1, XIS2, XIS3 needs to be adjusted downward and those of pn, XRT-WT, XRT-PC needs to be adjusted upward.

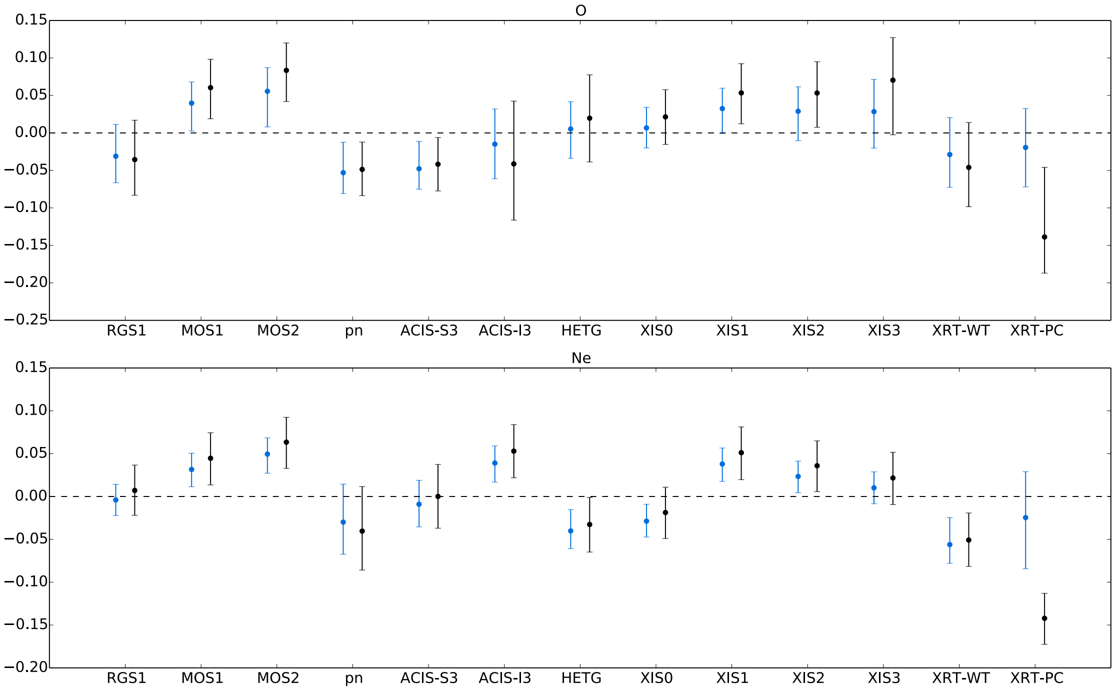

Figure 3.6: Adjustments of the log-scale of detector effective area for O (row 1) and Ne (row 2) in E0102 data set. The x-axis labels the detectors and the y-axis is B_i. The horizontal dashed lines represent zero, which is the baseline. The vertical bars denote 95% posterior interval for each B_i, whereas the dots denote the posterior means. The blue bars correspond to $\tau_i = 0.025$ and the black bars correspond to $\tau_i = 0.05$.

3.3.2.2 2XMM data

The 2XMM catalogue can be used to generate large, well-defined samples of various types of astrophysical objects, notably active galaxies (AGN), clusters of galaxies, interacting compact binaries, and active stellar coronae, using the power of X-ray selection[59]. The 2XMM data are from the XMM-Newton European Photon Imaging Cameras (EPIC). Briefly, there are three EPIC instruments: the EPIC-pn (pn) and the two EPIC-MOS detectors (MOS1 and MOS2). These detectors have separate X-ray focusing optics but are co-aligned so that the sources in our samples are observed simultaneously in the pn, MOS1, and MOS2 detectors.

The 2XMM data contains three data sets, corresponding to the hard (2.5 - 10.0 keV), medium (1.5 - 2.5 keV) and soft (0.5 - 1.5 keV) energy bands. The three detectors (pn, MOS1 and MOS2) are used to measure 41 sources in the hard band, 41 in the medium band, and 42 in the soft band. The sources are from the 2XMM EPIC Serendipitous Source Catalog[58], selected to be sufficiently faint that pileup, which occurs when several photons hit the detector at the same time and causes extra uncertainty in observations, is not a problem. With sufficient exposure, on average 1,500 counts are collected from the faint sources in each band for each detector.

We fit the log-normal hierarchical model in Section 3.2.2 to the three data sets individually. We set $df_g = 1.5$ for all energy bands and set $\beta_g = 0.116^2$ for hard band, $\beta_g = 0.288^2$ for medium band, and $\beta_g = 0.148^2$ for soft band. We again use $b_i = 0$ and try $\tau_i = 0.025$ and $\tau_i = 0.05$. Figure 3.7 shows the adjustments of the log-scale effective area for hard band (left), medium band (middle) and soft band (right) for 2XMM data, with both values of the τ_i. The results confirm the astronomers' intuition that no adjustment of the effective areas of the different detectors are needed for 2XMM data, regardless of the choice of the τ_i. We tabulate the proportion of prior information for each detector-source pair in Section C.5.

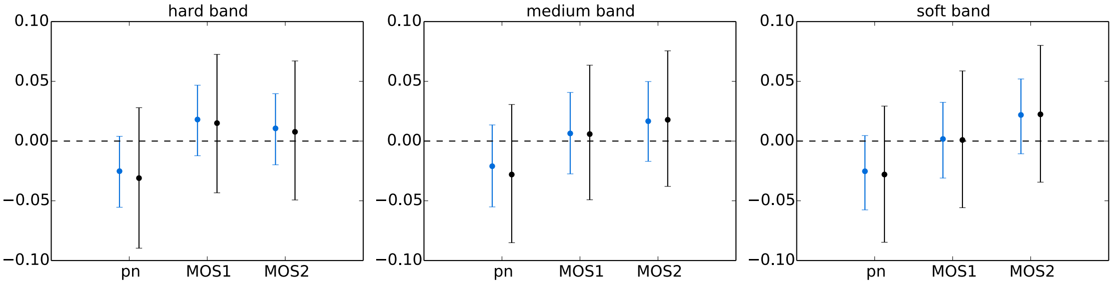

Figure 3.7: Adjustments of the log-scale effective areas for hard band (left), medium band (middle) and soft band (right) of the 2XMM data set. The legend is the same as in Figure 3.6.

3.3.2.3 XCAL data

Another XMM data consists of bright active galactic nuclei from the XMM-Newton cross-calibration sample, denoted as the XCAL data set. The pileup is very important for XCAL data, so the image data are clipped to eliminate the regions affected by pileup and the estimated effective area is adjusted according to lookup tables (from other in-flight data) that account for the unused regions. The region that is clipped out is determined using a standard XMM software task (called epatplot) and depends on the observed source intensity: unused regions are larger for brighter sources. This process is described in more detail in our companion paper[32].

Like the 2XMM data, XCAL data are composed of three data sets: the hard, medium, and soft energy bands. For each energy band, three detectors, MOS1, MOS2 and pn, are used to measure 94 (hard band), 103 (medium band), and 108 (soft band) sources. The model fitting follows the same procedure as detailed in Section 3.3.2.2. We again use $b_i = 0$ and try $\tau_i = 0.025$ and $\tau_i = 0.05$. For the prior on each σ_i^2, we set $df_g = 1.5$ for all three energy bands and $\beta_g = 0.028^2$ for the hard band, $\beta_g = 0.093^2$ for the medium band, and $\beta_g = 0.026^2$ for the soft band.

Figure 3.8 demonstrates that adjustment of the effective areas is needed to make the measured fluxes consistent across different detectors. We take four sources from the medium band data and use black vertical bars to denote the 95% interval (mean ± 2 given standard deviations) for the log-fluxes obtained with a standard astronomical method for each of the three detectors. The intervals match in some cases but are quite distinct in others. For each source, we also plot the 95% posterior intervals of the estimated log-fluxes after adjustment using the log-normal hierarchical model. The fitting results with different τ_i are consistent, regardless of the length of the posterior interval. This simple visualization gives us evidence that calibration of the effective areas is necessary to obtain consistent flux estimates.

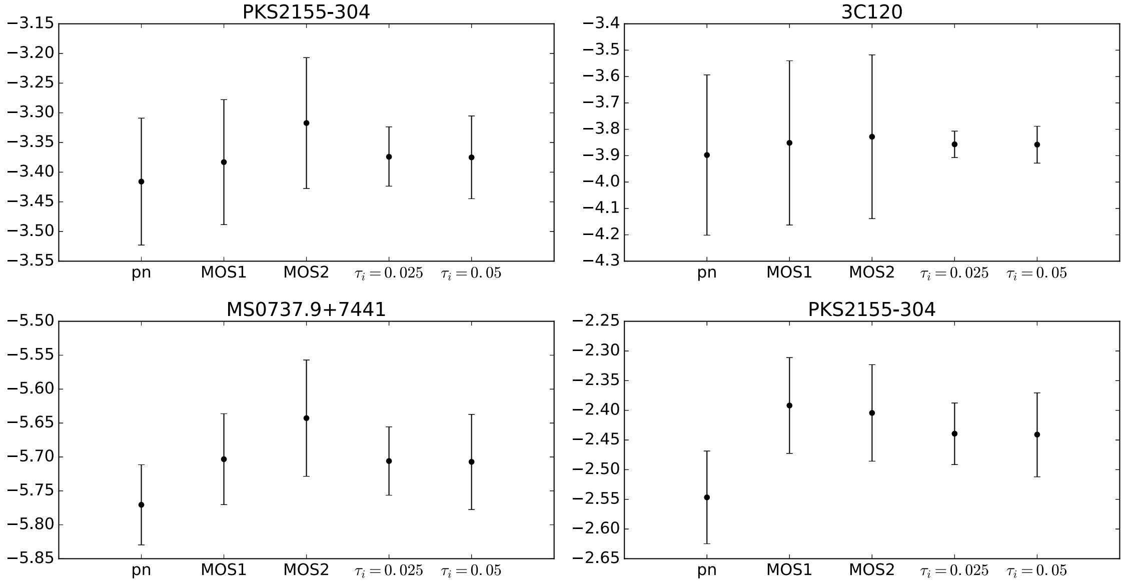

Figure 3.8: Comparison of estimated 95% intervals for log-fluxes using standard astronomical method (left three bars) and those by fitting the log-normal hierarchical model (right two bars) for four representative sources from medium band measurements. The titles of each panel give the names of the sources.

Finally, we show how to adjust the effective areas of each instrument to obtain the results illustrated in the rightmost interval in each panel of Fig. 3.8. Figure 3.9 shows the necessary adjustment of the B_i for hard band (left), medium band (middle) and soft band (right). For all these bands, we must adjust pn upward and MOS2 downward.

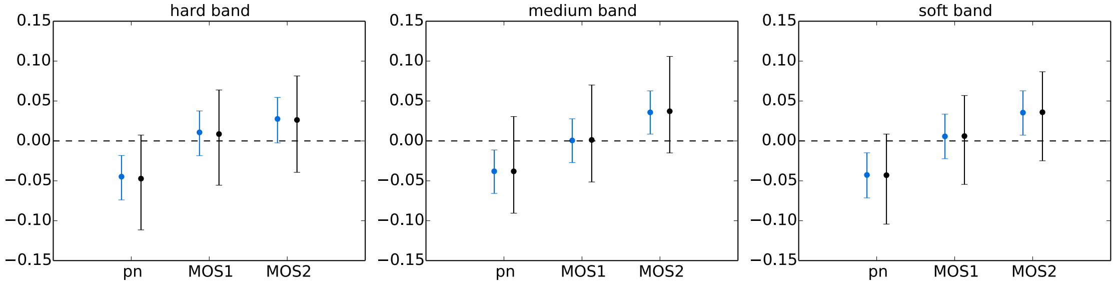

Figure 3.9: Adjustments of the log-scale effective areas for hard band (left), medium band (middle) and soft band (right) for XCAL data. The legend is the same as in Figure 3.6.

3.3.2.4 Model checking

In this section, we study how well the log-normal hierarchical model captures the observed variability in the data. We visualize the residuals of the fitted log-normal hierarchical model and deploy a posterior predictive check.

We propose using residual plots to visualize the goodness-of-fit. Specifically, Figure 3.10 plots the standardized residuals, as shown in (3.12), for data analyzed in Section 3.3.2.3 with $\tau_i = 0.05$, with the left panel denoting the results from the log-normal hierarchical model and the right panel denoting the results from the log-t model. Nearly all of the standardized residuals fall in the interval $[-3, 3]$ for the log-normal hierarchical model and $[-2, 2]$ for the log-t model. The observations of 3C111 in all three energy bands are the only outliers (with large standardized residuals) in the log-normal hierarchical model, but not for the log-t model which down weights the outliers. The adjusted effective areas and the estimated fluxes are not too sensitive to whether or not we include the outliers in the analysis. Thus the log-normal hierarchical model is good enough for the data in Section 3.3.2.3.

We use a posterior predictive check[33,13] to detect if there is any serious error with the log-normal hierarchical model. In a posterior predictive check, one chooses a test statistic and computes the corresponding posterior predictive p-value. The test statistic we choose

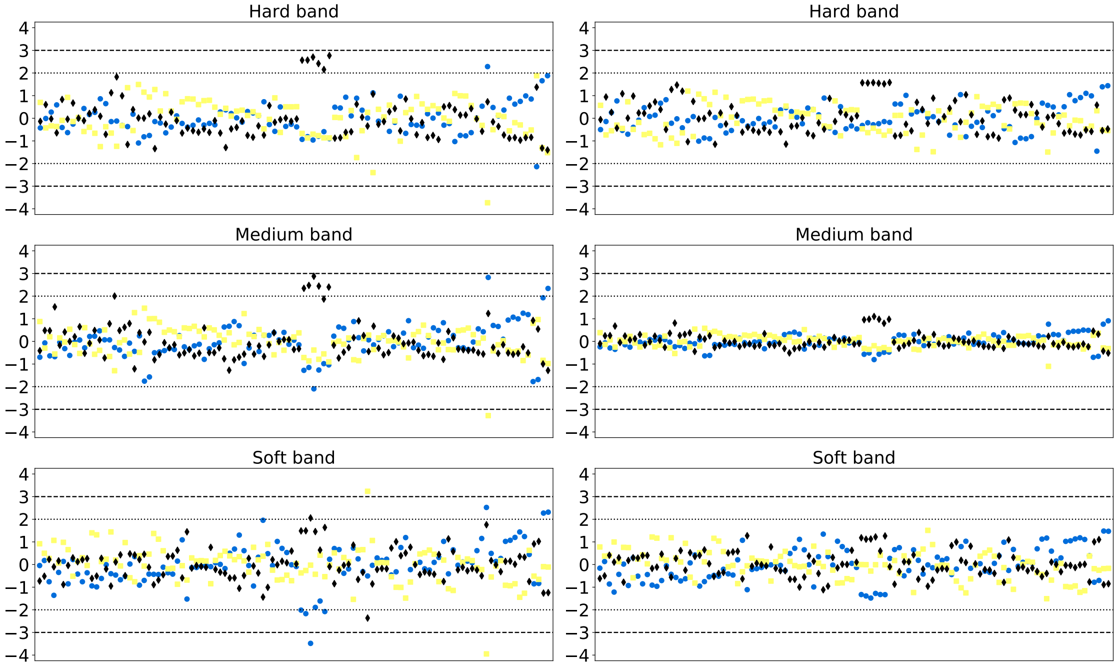

Figure 3.10: Standardized residuals of hard (row 1), medium (row 2) and soft (row 3) band data in Section 3.3.2.3 with $\tau_i = 0.05$, the log-normal hierarchical model on the left panel and log-t model on the right panel. The blue circles, yellow squares and black rhombuses denote the instruments pn, MOS1 and MOS2 respectively. The dashed horizontal lines denote the $[-3, 3]$ intervals and the horizontal dots denote $[-2, 2]$ intervals.

is

$$\left\{ T_i = \bar{y}_{i\cdot} - \bar{y} = \frac{\sum_{j=1}^{M} y_{ij}}{M} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{NM} \right\}_{i=1}^{N},$$

which reflects the relative magnitudes of the log scale effective areas. None of the posterior predictive p-values for any of our datasets is significant, i.e., we never fail the posterior predictive check. Therefore, regarding potential serious defects of our model, the results are encouraging so far because the tests do not show any serious discrepancy.

3.4 Discussions and conclusions

3.4.1 Discussion of frequentist method

In Section 3.2, we adopt a Bayesian perspective, which leads to the log-normal hierarchical model elaborated in Sections 3.2 and 3.3. Here we discuss the alternative frequentist method of tackling the calibration concordance problem.

3.4.1.1 MLEs and asymptotic properties

The regression model in (3.3) together with (3.4) yields a special case of a multivariate linear regression and can be fitted as such. Nonetheless, it is more straightforward and instructive to fit the model via maximum likelihood.

When the variances ψ are known, the regression model we consider is in fact a Gaussian model. Thus, the variances of the MLEs can be obtained through inverting the Fisher information matrix. Theorem C.2 in Section C.3.1 gives the MLEs of the B_i and the G_j. Proposition C.1 in Section C.3.1 gives the closed-form solution of the variance-covariance matrix for the MLEs of the B_i and the G_j when all the instruments measure all sources. Furthermore, the standardized residual sum of squares follow a chi-squared distribution, which enables easy testing of the goodness-of-fit; see Section C.3.2 for details.

When the variances ψ are unknown, in principle, we can also obtain the (asymptotic) variance of the MLEs by calculating the observed/expected Fisher information. However, the number of unknown parameters $(2N + M)$, grows infinitely as the number of observations $(NM + N)$ goes to infinity. Whether the classical MLE asymptotic theory can be easily adapted to this situation is a problem for future study.

These estimators are approximately valid even if the Gaussian assumptions made in (3.3) and for the b_i are not valid. In this case, the variance of the estimator requires a more complicated sandwich formula, which involves both the Fisher information and the variance of the score function. Here we say approximately valid because the half-variance correction of Section 3.2 would still depend on the normal assumption. Consequently, when the variance is large, our bias correction may be off if the normal assumption is severely violated.

3.4.1.2 Comparison to Bayesian method

It is easy to check that when the variances ψ are known, the MLEs of the B_i and the G_j corresponds to the MAP estimation defined in (3.8), which also have the intuitive interpretation as shrinkage estimators. When the variances are unknown, the likelihood function is unbounded on the boundary ($\sigma_i^2 = 0$) and the maximization algorithm converges to the boundary of the parameter space. The conjugate priors for the variance parameters in the Bayesian model regularizes the likelihood and gives a proper posterior distribution. This is another reason why we adopt the Bayesian model when the variances are unknown.

3.4.2 Conclusions and future work

In this chapter, we propose a log-normal approach to tackle the calibration concordance problem which consists of measurements of intrinsic properties of multiple astronomical objects with multiple instruments. This approach well represents the physical multiplicative model on the mean captured by the residuals on the log scale and is shown reasonably robust to misspecification of the physical model, which is typically the case in practice. In addition, we generalize the log-normal hierarchical model to a more flexible log-t model which is more robust to outliers, which are prevalent in astrophysical observations. We resolve the identifiability problem of the measurement model by incorporating the imprecise, empirical knowledge of scientists accordingly. Intuitively, the different pieces of information coming from experts' knowledge, as vague or as precise as it is, and the observations are combined using the shrinkage estimators. We adopt the Hamiltonian Monte Carlo algorithm to obtain the posterior distribution, which resides in a high-dimensional space with highly correlated parameters. We give detailed descriptions of the model fitting and illustrate our method via a variety of simulation studies and real data results.

The log-normal hierarchical model proposed in this chapter works well for real data and yields important astronomical findings – concrete guidance about systematic adjustments of the effective areas for each instrument are given thus concordance of an intrinsic property for each astronomical object across different instruments is achieved. Calibration scientists are thus able to make absolute measurements of properties of astronomical objects using different instruments. Furthermore, we detect the danger of wrongly fixing the observation noise, which scientists are tempted to do, through various simulation experiments that mimic possible realistic uncertainties.

There are several future works that can improve the current model. First, we assume that the effective areas are independent as a priori, which is not always true in practice. Sometimes the effective areas across different energy bands are correlated. We plan to take this correlation structure into account in future modeling. Second, the log-normal hierarchical model gives conservative results under realistic model misspecification according to our simulation studies. Theoretical properties of the log-normal approach under model misspecification need to be further investigated. Third, the statistical properties of the log-t model need more study. Last, the asymptotic properties of the models proposed in this chapter are intriguing issues to be addressed, under the bigger umbrella of the asymptotic behaviors of models with both the number of parameters and the number of observations approaching infinity.

A

Supplementary materials for Chapter 1

A.1 Computation time

We study the computing time for different methods with sample sizes $n = 50,\ 100,\ 225$ and 500. For each n we simulate 1,000 observations and record the computing time for every method; the average time is shown in Fig A.1. The computing time for G_{t}^2 is twice as much as the computing time for G_{m}^2 due to the normalizing constant. This time can be further reduced by tabulating the normalizing constant for pairs of (n, λ_0). G_{m}^2 and G_{t}^2 are more time efficient compared with DCOR, DDP and MIC$_e$.

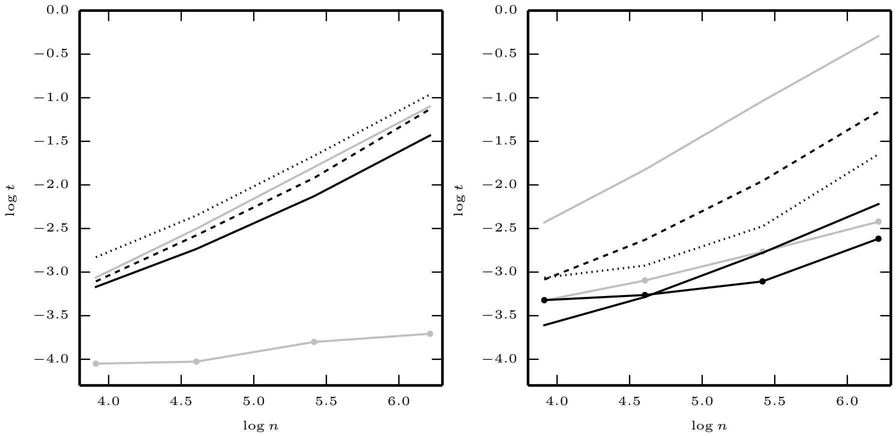

Figure A.1: The left figure shows the average computing time of G_{m}^2 (black solid), G_{t}^2 (grey solid), COR (grey markers), DCOR (black dashes) and DDP (black dots) for 1,000 simulations with sample sizes $n = 50,\ 100,\ 225$ and 500; the right figure shows the average computing time of mutual information (black solid), MIC$_e$ (grey solid), ACE (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers). The x-axis is the logarithm of n with base e and the y-axis is the logarithm of the computing time in seconds with base 10.

A.2 Segmented regression

The R-squared for segmented regression with predictor X and response Y is

$$R^2 = 1 - \frac{\sum_{h=1}^{K} n_h \widehat{\sigma}_h^2}{n\widehat{v}^2},$$

where $\widehat{v}^2$ is the sample variance of Y, n_h and $\widehat{\sigma}_h^2$ are sample size and residual variance of Y after regressing on X in segment h $(h = 1, \ldots, K)$. R^2 can be viewed as an estimator of

$$R_{Y|X}^2 = 1 - \frac{E\{\mathrm{var}(Y \mid X)\}}{\mathrm{var}(Y)};$$

it is zero if and only if $E(Y \mid X)$ is a constant. $G_{Y|X}^2$ is zero if and only if both $E(Y \mid X)$ and $\mathrm{var}(Y \mid X)$ are constant. $G_{Y|X}^2$ equals $R_{Y|X}^2$ when $\mathrm{var}(Y \mid X)$ is a constant, but $G_{Y|X}^2$ is more general than $R_{Y|X}^2$ since it can capture heteroscedastic effects.

Given a fixed number of segments K, computing $R_{Y|X}^2$ with the optimal segmentation is more computationally intensive than computing G^2_{m} and G^2_{t}, especially when K is large. When K is unknown, we can apply the same dynamic programming algorithm for G^2_{m} or G^2_{t} and fit a penalized version of the segmented regression to avoid over-fitting. If we also require that the fitted curve be continuous, no exact numerical solution is available; we can potentially design a Markov chain Monte Carlo algorithm under a Bayesian framework.

A.3 Proofs of consistency and relationship with R-squared

A.3.1 Proof of Theorem 1.1 - consistency

The following lemma is needed for the main theorem.

Lemma A.1. *Suppose X and Y are univariate continuous random variables with $|X|,\ |Y| < B$ and $\mathrm{var}(Y) > b^{-2}$. Given n observations (x_i, y_i) $(i = 1, \ldots, n)$ and let $\widehat{\sigma}^2$ be the residual variance after regressing Y on X. Then,*

$$\mathrm{pr}\left[\left|\widehat{\sigma}^2 - \left\{\mathrm{var}(Y) - \frac{\mathrm{cov}^2(X, Y)}{\mathrm{var}(X)}\right\}\right| > \varepsilon\right] \ \leq\ 10e^{-C_1(B,b)n\varepsilon^2}$$

with $C_1(B, b) = (288b^2B^4)^{-1} \min\{1,\ (4b^2B^2)^{-1}\}$ and $\varepsilon > 0$ small enough.

Proof of Lemma A.1. Without loss of generality, we assume $E(X) = E(Y) = 0$, $\mathrm{var}(X) = \mathrm{var}(Y) = 1$ and $E(XY) = \rho$. By definition

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 - \frac{\left\{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} y_i)\right\}^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}.$$

Then $x_i^2, y_i^2 \in [0,\ B^2], x_iy_i \in [-B^2,\ B^2]$. According to Hoeffding's inequality,

$$\begin{aligned}&\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| > \varepsilon/6\right), \quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i\right| > \varepsilon/6\right), \quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1\right| > \varepsilon/6\right),\\ &\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 1\right| > \varepsilon/6\right), \quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_iy_i - \rho\right| > \varepsilon/6\right) \leq 2\exp\{-c(B)n\varepsilon^2\}\end{aligned}$$

with $c(B) = (72B^2)^{-1}\min(1,\ B^{-2})$. If $\varepsilon < 1$ and

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right|, \quad \left|\frac{1}{n}\sum_{i=1}^{n} y_i\right|, \quad \left|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1\right|, \quad \left|\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 1\right|, \quad \left|\frac{1}{n}\sum_{i=1}^{n} x_iy_i - \rho\right| \leq \varepsilon/6,$$

we have

$$\begin{aligned}\left|\widehat{\sigma}^2 - 1 + \rho^2\right| &\leq \left|1 - \frac{1}{n}\sum_{i=1}^{n} y_i^2\right| + \left|\frac{1}{n}\sum_{i=1}^{n} y_i\right|^2 + \frac{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 - 1|\rho^2}{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2|}\\ &\quad \frac{|\left\{\frac{1}{n}\sum_{i=1}^{n} x_iy_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} y_i)\right\}^2 - \rho^2|}{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2|}\\ &\leq \frac{4(\varepsilon/6 + \varepsilon^2/36)}{1 - \varepsilon/6 - \varepsilon^2/36} < \varepsilon.\end{aligned}$$

So $\mathrm{pr}\left(\left|\widehat{\sigma}^2 - 1 - \rho^2\right| > \varepsilon\right) \leq 10\exp\{-c(B)n\varepsilon^2\}$. For general cases, define

$$X' = \frac{X - E(X)}{\mathrm{sd}(X)}, \quad Y' = \frac{Y - E(Y)}{\mathrm{sd}(Y)}.$$

Then $E(X') = E(Y') = 0$, $\mathrm{var}(X') = \mathrm{var}(Y') = 1$ and $|X'|, |Y'| < 2bB$. Thus,

$$\begin{aligned}&\mathrm{pr}\left[\left|\widehat{\sigma}^2 - \left\{\mathrm{var}(Y) - \frac{\mathrm{cov}^2(X, Y)}{\mathrm{var}(X)}\right\}\right| > \varepsilon\right]\\ = \ &\mathrm{pr}\left[\left|\widehat{\sigma}'^2 - \{1 - \mathrm{cov}^2(X', Y')\}\right| > \frac{\varepsilon}{\mathrm{var}(Y)}\right]\\ \leq \ &10\exp\{-\frac{c(2bB)}{\mathrm{var}(Y)^2}n\varepsilon^2\} = 10\exp\{-C_1(B, b)n\varepsilon^2\}\end{aligned}$$

with $C_1(B,b) = (288b^2B^4)^{-1}\min\{1,\ (4b^2B^2)^{-1}\}$. □

Proof of Theorem 1.1. We only need to prove that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$ are consistent estimators of $G^2_{Y|X}$. If so, by switching X and Y, we must have that $G^2_{\mathrm{m}}(X \mid Y, \lambda_0)$ and $G^2_{\mathrm{t}}(X \mid Y, \lambda_0)$ are consistent estimators of $G^2_{X|Y}$ which guarantees the consistency of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$.

We first introduce some notations that will appear later. Suppose $|X|,\ |Y| < B$. Condition 1 shows that $v_X(y) > b^{-2}$ almost surely. Let $m = \lceil n^{1/2} \rceil$ be the minimum size of slices, and let $s \in S$ denote a slice and p_s be the probability that an observation falls in s. Let E_s, var_s, and cov_s denote the mean, variance and covariance conditional on slice s. Finally, define

$$\sigma_s^2 = \mathrm{var}_s(Y) - \frac{\mathrm{cov}_s^2(X,Y)}{\mathrm{var}_s(X)}.$$

Then by definition

$$\sigma_s^2 \geq \mathrm{var}_s(Y) - \mathrm{var}_s\{E(Y \mid X)\} = E_s\{\mathrm{var}(Y \mid X)\} \geq \exp[E_s\{\log \mathrm{var}(Y \mid X)\}] \geq b^{-2}.$$

For observations (x_i, y_i) $(i = 1, \ldots, n)$, let $\hat{v}^2$ be the estimated variance of Y and $\widehat{\sigma}_s^2$ be the residual variance after regressing Y on X in slice s. Besides, we use the following inequality

$$1 - x^{-1} < \log x < x - 1, \quad x > 0$$

throughout the proof.

Now we prove that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ is a consistent estimator for $G^2_{Y|X}$. Define

$$d_{Y|X} = \log \mathrm{var}(Y) - E\left\{\log \mathrm{var}(Y \mid X)\right\},$$

so $G^2_{Y|X} = 1 - \exp(-d_{Y|X})$. Because

$$G^2_{\mathrm{m}}(Y \mid X) = 1 - \exp\{-\max_{S:\ m_S \geq m} D(Y \mid S, \lambda_0)\},$$

we only need to show the consistency of $\max_{S:\ m_S \geq m} D(Y \mid S, \lambda_0)$, denoted as $D(Y \mid X, \lambda_0)$. We prove this in two steps:

Step 1: We show that there exists $\eta_1(n) > 0$ and $\eta_1(n) \to 0$ as $n \to \infty$, such that

$$\mathrm{pr}\left\{\limsup_{n\to\infty} D(Y \mid X, \lambda_0) < d_{Y|X} + \eta_1(n)\right\} = 1,$$

which means that $D(Y \mid X, \lambda_0)$ is almost surely smaller than $d_{Y|X}$. Because for any slicing scheme S, $\log \mathrm{var}(Y) - \sum_{s\in S} p_s \log \sigma_s^2 \leq d_{Y|X}$, it is enough to show that there is $\eta_1(n)$ such that

$$\mathrm{pr}\left\{\limsup_{n\to\infty} D(Y \mid S, \lambda_0) - \log \mathrm{var}(Y) + \sum_{s\in S} p_s \log(\sigma_s^2) < \eta_1(n)\right\} = 1.$$

Let $\delta(n) = \log(n)n^{-1/4}$. By definition of $D(Y \mid S, \lambda_0)$, we have

$$\begin{aligned} & D(Y \mid S, \lambda_0) - \log \mathrm{var}(Y) + \sum_{s\in S} p_s \log(\sigma_s^2) \\ \leq\ & \left\{\log \hat{v}^2 - \log \mathrm{var}(Y)\right\} + \sum_{s\in S}\left(p_s - \frac{n_s}{n}\right)\log \sigma_s^2 + \sum_{s\in S}\frac{n_s}{n}\left(\log \sigma_s^2 - \log \widehat{\sigma}_s^2\right). \end{aligned}$$

First, we consider $\log \hat{v}^2 - \log \mathrm{var}(Y)$. By Hoeffding's inequality, for $0 < \varepsilon < 2$,

$$\begin{aligned} & \mathrm{pr}\left\{|\hat{v}^2 - \mathrm{var}(Y)| > \varepsilon\right\} \\ \leq\ & \mathrm{pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\{y_i - E(Y)\}^2 - \mathrm{var}(Y)\right| > \varepsilon/2\right] + \mathrm{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} y_i - E(Y)\right| > \varepsilon/2\right\} \\ \leq\ & 4\exp\left[-n\varepsilon^2 \min\{1,\ (4B^2)^{-1}\}(8B^2)^{-1}\right], \end{aligned}$$

we have

$$\begin{aligned} & \operatorname{pr}\left\{\log \hat{v}^2 - \log \operatorname{var}(Y) > \delta(n)\right\} \\ \leq\ & \operatorname{pr}\left\{\hat{v}^2 - \operatorname{var}(Y) > \operatorname{var}(Y)\delta(n)\right\} \leq 4n^{-c_1 n^{1/2} \log n} \end{aligned} \qquad \text{(A.1)}$$

with $c_1 = \min\{1,\ (4B^2)^{-1}\}(8b^4B^2)^{-1}$.

Second, we consider $\sum_{s\in S}(p_s - n_s/n)\log\sigma_s^2$. Let us define a new random variable Z and $Z = \log\sigma_s^2$ if X is in slice s. Let z_i $(i = 1, \ldots n)$ be n independent observations of Z, then,

$$E(Z) = \sum_{s\in S} p_s \log\sigma_s^2, \quad \frac{1}{n}\sum_{i=1}^{n} z_i = \sum_{s\in S}\frac{n_s}{n}\log\sigma_s^2.$$

By Hoeffding's inequality and the fact that $\sigma_s^2 \in [b^{-2},\ B^2]$,

$$\operatorname{pr}\left\{\left|\sum_{s\in S}(p_s - \frac{n_s}{n})\log\sigma_s^2\right| > \delta(n)\right\} \ \leq\ 2n^{-c_2 n^{1/2}\log n} \qquad \text{(A.2)}$$

with $c_2 = \min(1/|\log B|^2,\ 1/|\log b|^2)/2$.

Third, we focus on the difference between $\log\widehat{\sigma}_s^2$ and $\log\sigma_s^2$. Consider a slicing scheme Q_n of n^4 slices such that an observation falls in each slice equally. Given n observations, the probability for any of the n^4 slices containing more than one observations is smaller than

$$n^4\left\{1 - \left(1 + n^{-3}\right)\left(1 - n^{-4}\right)^n\right\} \leq n^{-2}.$$

Then event

$$E_{1,n} = \{\text{each slice of } Q_n \text{ has at most one observation}\}$$

satisfies $\operatorname{pr}(\liminf_{n\to\infty} E_{1,n}) = 1$. Thus, we only need to consider slicing schemes that are more refined than Q_n, denoted as $S \preceq Q_n$. Define the set of slices as

$$\Xi = \{s \mid \text{there exists } S \preceq Q_n \text{ such that } s \in S\}.$$

The set Ξ contains at most $n^4(n^4+1)/2 = O(n^8)$ slices. Each slice $s \in \Xi$ contains at least m observations. By Lemma A.1, if $\delta(n) < 0.5b^{-2}$,

$$\begin{aligned} & \operatorname{pr}\left\{\log \sigma_s^2 - \log \widehat{\sigma}_s^2 > \delta(n)\right\} & \text{(A.3)} \\ \leq\ & P\{\sigma_s^2/\widehat{\sigma}_s^2 - 1 > \delta(n)\} \\ \leq\ & \operatorname{pr}\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| > \delta(n)\right\} + P\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| > \delta(n)\widehat{\sigma}_s^2,\ |\widehat{\sigma}_s^2 - \sigma_s^2| \leq \delta(n)\right\} \\ \leq\ & 20n^{-c_3 \log(n)}. \end{aligned}$$

with $c_3 = C_1(B,b)\min\{1,\ (4b^4)^{-1}\}$. Let $\eta_1(n) = 3\delta(n)$ and event

$$E_{2,n} = \{\max_{S \preceq Q_n} D(Y \mid S, \lambda_0) < d_{Y|X} + \eta_1(n)\}.$$

Combine the results of (A.1)–(A.3), we have $\operatorname{pr}\left(\liminf_{n\to\infty} E_{1,n} \cap E_{2,n}\right) = 1$, which means that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ is almost surely smaller than $G^2_{Y|X}$.

Step 2: Next, we show that there exists $\eta_2(n) > 0$ and $\eta_2(n) \to 0$ as $n \to \infty$, such that

$$\operatorname{pr}\left\{\liminf_{n\to\infty} D(Y \mid X, \lambda_0) > d_{Y|X} - \eta_2(n)\right\} = 1,$$

which means that $D(Y \mid X, \lambda_0)$ is almost surely larger than $d_{Y|X}$. We just need to prove that for any sample size n, there exists a slicing scheme T_n such that

$$\operatorname{pr}\left(\liminf_{n\to\infty} E_{3,n} \cap E_{4,n}\right) = 1,$$

where

$$E_{3,n} = \{\text{each slice of } T_n \text{ contains at least } m \text{ samples}\}$$

and

$$E_{4,n} = \{D(Y \mid T_n, \lambda_0) > d_{Y|X} - \eta_2(n)\}.$$

Consider a slicing scheme T_n of $\lfloor n^{1/4} \rfloor$ slices such that an observation falls in one slice equally. Then, we further divide each slice into $\lfloor n^{1/2} \rfloor$ bins such that an observation falls in each bin equally. Given n observations, the probability that each bin contains at least one observation is greater than

$$1 - \lfloor n^{1/4} \rfloor \lfloor n^{1/2} \rfloor (1 - n^{-3/4})^n > 1 - \lfloor n^{1/4} \rfloor \lfloor n^{1/2} \rfloor e^{-n^{1/4}},$$

so each slice of T_n contains at least m observations. Then, $\mathrm{pr}\left(\liminf_{n\to\infty} E_{3,n}\right) = 1$. Define

$$\Delta_n(T_n) = \log \mathrm{var}(Y) - \sum_{s \in T_n} p_s \log \mathrm{var}_s(Y).$$

We first consider the difference between $D(Y \mid T_n, \lambda_0) - \Delta_n(T_n)$:

$$\begin{aligned}
& D(Y \mid T_n, \lambda_0) - \Delta_n(T_n) \\
\geq\ & \left\{\log \hat{v}^2 - \log \mathrm{var}(Y)\right\} + \sum_{s \in T_n} \left(p_s - \frac{n_s}{n}\right) \log \mathrm{var}_s(Y) + \sum_{s \in T_n} \frac{n_s}{n} \{\log \mathrm{var}_s(Y) - \log \widehat{\sigma}_s^2\} \\
& - \lambda_0 n^{-3/4} \log n.
\end{aligned}$$

Similar as (A.1), if $\delta(n) < 0.5b^{-2}$,

$$\begin{aligned} & \operatorname{pr}\left\{\log \hat{v}^2 - \log \operatorname{var}(Y) < -\delta(n)\right\} && \text{(A.4)} \\ \leq\ & \operatorname{pr}\left\{1 - \operatorname{var}(Y)/\hat{v}^2 < -\delta(n)\right\} \\ \leq\ & \operatorname{pr}\left\{|\hat{v}^2 - \operatorname{var}(Y)| > \delta(n)\right\} + P\left\{|\hat{v}^2 - \operatorname{var}(Y)| > \delta(n)\hat{v}^2,\ |\hat{v}^2 - \operatorname{var}(Y)| \leq \delta(n)\right\} \\ \leq\ & 4n^{-c_4 n^{1/2}\log n} \end{aligned}$$

with $c_4 = (8B^2)^{-1}\min\{1,\ (4B^2)^{-1}\}\min\{1,(4b^4)^{-1}\}$. Similar as (A.2), we have

$$\operatorname{pr}\left\{\left|\sum_{s\in S}(p_s - \frac{n_s}{n})\log \operatorname{var}_s(Y)\right| > \delta(n)\right\} \ \leq\ 2n^{-c_2 n^{1/2}\log n}. \qquad \text{(A.5)}$$

Besides, $\operatorname{var}_s(Y) \geq \sigma_s^2$ and

$$\begin{aligned} & \operatorname{pr}\left\{\log \operatorname{var}_s(Y) - \log \widehat{\sigma}_s^2 < -\delta(n)\right\} && \text{(A.6)} \\ \leq\ & \operatorname{pr}\left\{\log \sigma_s^2 - \log \widehat{\sigma}_s^2 < -\delta(n)\right\} \\ \leq\ & \operatorname{pr}\left\{1 - \widehat{\sigma}_s^2/\sigma_s^2 < -\delta(n)\right\} \\ \leq\ & \operatorname{pr}\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| \geq b^{-2}\delta(n)\right\} \leq 10n^{-C(B,b)b^{-4}\log(n)}. \end{aligned}$$

Now, define $\delta_1(n) = 3\delta(n) + \lambda_0 \log(n) n^{-3/4}$ and event

$$E_{5,n} = \{D(Y \mid T_n, \lambda_0) > \Delta_n(T_n) - \delta_1(n)\}.$$

By (A.4)–(A.6), $\operatorname{pr}(\liminf_{n\to\infty} E_{3,n} \cap E_{5,n}) = 1$.

The only problem left is how to control the difference between $\Delta_n(T_n)$ and $d_{Y|X}$, which is

$$\Delta_n(T_n) - d_{Y|X} = \sum_{s\in T_n} p_s \left\{ \frac{1}{p_s} \int_s \log v_Y^2(x) f_X(x) dx - \log \mathrm{var}_s(Y) \right\}.$$

Denote the probability density function of X as $f_X(x)$. For one slice s, because X is a continuous random variable, set

$$\frac{1}{p_s} \int_s \mu_Y(x) f_X(x) dx = \mu_Y(x_s'), \quad \frac{1}{p_s} \int_s \log v_Y^2(x) f_X(x) dx = \log v_Y^2(x_s''),$$

where x_s' and x_s'' lie in the slice almost surely. Then

$$\begin{aligned}
& \log v_Y^2(x_s'') - \log \mathrm{var}_s(Y) \\
= \;& \log v_Y^2(x_s'') - \log \left[\frac{1}{p_s} \int_s v_Y^2(x) f_X(x) dx + \frac{1}{p_s} \int_s \{\mu_Y(x) - \mu_Y(x_s')\}^2 f_X(x) dx \right] \\
= \;& \log v_Y^2(x_s'') - \log \left[v_Y^2(x_s'') + \frac{1}{p_s} \int_s \int_{x_s''}^{x} 2 v_Y(z) v_Y'(z) dz f_X(x) dx \right. \\
& \qquad\qquad \left. + \frac{1}{p_s} \int_s \left\{ \int_{x_s'}^{x} \mu_Y'(z) dz \right\}^2 f_X(x) dx \right] \\
\geq \;& \log v_Y^2(x_s'') - \log \left[v_Y^2(x_s'') + \int_s 2 v_Y(x) |v_Y'(x)| dx + \left\{ \int_s |\mu_Y'(x)| dx \right\}^2 \right].
\end{aligned}$$

According to Condition 3, we have

$$\begin{aligned}
& \log v_Y^2(x_s'') - \log \mathrm{var}_s(Y) \\
\geq \;& \log v_Y^2(x'') - \log \left\{ v_Y^2(x'') + 2C \int_s v_Y^2(x) dx + C^2 \int_s 1 dx \int_s v_Y^2(x) dx \right\} \\
\geq \;& -\frac{\int_s v_Y^2(x) dx \left(2C + C^2 \int_s 1 dx\right)}{v_Y^2(x'')} \\
\geq \;& -2b^2 B^2 C(1 + BC) \int_s 1 dx.
\end{aligned}$$

Then, we can conclude

$$\begin{aligned} \Delta_n(T_n) - d_{Y|X} &\geq -2p_s b^2 B^2 C(1+BC) \sum_{s \in T_n} \int_s 1 dx \\ &\geq -4\lfloor n^{1/4} \rfloor^{-1}(1+BC)Cb^2B^3 = -\delta_2(n). \end{aligned}$$

Therefore, let $\eta_2(n) = \delta_1(n) + \delta_2(n)$, we have $\mathrm{pr}\,(\liminf_{n\to\infty} E_{3,n} \cap E_{4,n}) = 1$, which means $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ is almost surely larger than $G^2_{Y|X}$. By Steps 1 and 2, we can conclude that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ is a consistent estimator of $G^2_{Y|X}$.

To prove the consistency of $G^2_{\mathrm{t}}(Y \mid X, \lambda)$, we introduce a new quantity

$$Z(\lambda_0) = \sum_{m_s \geq m} n^{-\lambda_0(|S|-1)/2};$$

$Z(\lambda_0)$ is bounded by 1 and $(1 + n^{-\lambda_0/2})^n$. By definition of $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$, we have

$$\begin{aligned} \{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_0)\}^{-n/2} &= Z(\lambda_0)^{-1} \sum_{S:\, m_S \geq m} \exp\{\frac{n}{2} D(Y \mid S, \lambda_0)\} \\ &\geq Z(\lambda_0)^{-1} \exp\{\frac{n}{2} D(Y \mid X, \lambda_0)\}, \\ \{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_0)\}^{-n/2} &\leq Z(\lambda_0)^{-1} \sum_{S:\, m_S \geq m} \exp\{\frac{n}{2} D(Y \mid S, \frac{\lambda_0}{2}) - \frac{\lambda_0}{4}(|S|-1)\log(n)\} \\ &\leq Z(\lambda_0)^{-1} Z(\frac{\lambda_0}{2}) \exp\{\frac{n}{2} D(Y \mid X, \frac{\lambda_0}{2})\}. \end{aligned}$$

By the consistency of $D(Y \mid X, \lambda_0)$ and $D(Y \mid X, \lambda_0/2)$, we prove that $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$ is an consistent estimator of $G^2_{Y|X}$. □

A.3.2 Consistency of G^2_{m} and G^2_{t} with empirical Bayes selection of λ_0

Suppose λ_0^* is the optimal λ_0 that maximizes $\mathrm{bf}(\lambda_0)$ from a range $[\lambda_1,\ \lambda_2]$ with $\lambda_1\ >\ 0$. Then $Z(\lambda_2) \leq Z(\lambda_0^*) \leq Z(\lambda_1)$ and

$$
\begin{aligned}
G^2_{\mathrm{m}}(Y \mid X, \lambda_0^*) &\leq G^2_{\mathrm{m}}(Y \mid X, \lambda_1), \\
\left\{1 - G^2_{\mathrm{m}}(Y \mid X, \lambda_0^*)\right\}^{-n/2} &= \exp\{\frac{n}{2} D(Y \mid X, \lambda_0^*)\} \\
&\geq Z(\lambda_2)^{-1} \sum_{S:\ m_S \geq m} \exp\{\frac{n}{2} D(Y \mid S, \lambda_0^* + \lambda_2)\} \\
&\geq Z(\lambda_2)^{-1} \left\{1 - G^2_{\mathrm{m}}(Y \mid X, 2\lambda_2)\right\}^{-n/2}, \\
\left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_0^*)\right\}^{-n/2} &= Z(\lambda_0^*)^{-1} \sum_{S:\ m_S \geq m} \exp\{\frac{n}{2} D(Y \mid S, \lambda_0^*)\} \\
&\geq Z(\lambda_1)^{-1} \left\{1 - G^2_{\mathrm{m}}(Y \mid X, \lambda_2)\right\}^{-n/2}, \\
\left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_0^*)\right\}^{-n/2} &\leq Z(\lambda_0^*)^{-1} \sum_{S:\ m_S \geq m} \exp\{\frac{n}{2} D(Y \mid S, \lambda_1)\} \\
&\leq Z(\lambda_2)^{-1} Z(\lambda_1) \left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_1)\right\}^{-n/2}.
\end{aligned}
$$

By the consistency of $G^2_{\mathrm{m}}(Y \mid X, \lambda_1)$, $G^2_{\mathrm{m}}(Y \mid X, 2\lambda_2)$, $G^2_{\mathrm{m}}(Y \mid X, \lambda_2)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_1)$, we conclude that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0^*)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0^*)$ are consistent estimators. Then the estimators with data-driven λ_0 are consistent.

A.3.3 Proof of Theorem 1.2 - Equivalence between G^2_{m} and R^2

The following lemma is needed for the main theorem.

Lemma A.2. *Let* $(p_1, p_2, p_3) \sim \mathrm{Dir}(k_1, k_2, 2)$ *and*

$$\Lambda(q, p) = (k_1 - 1)\log\frac{q_1}{p_1} + (k_2 - 1)\log\frac{q_2}{p_2}.$$

Then for any $k_1,\ k_2 \geq 3$, $q_1,\ q_2 > 0,\ q_1 + q_2 = 1$ *and function* $\delta(p) > 0$,

$$\mathrm{pr}\left\{\Lambda(q,p) \geq \delta(p)\right\} \leq (k_1 + k_2)^3 \int_0^1 e^{-\delta(p)} dp.$$

Proof of Lemma A.2. By definition, we have

$$p_1^{k_1-1} p_2^{k_2-1}(1 - p_1 - p_2) \leq q_1^{k_1-1} q_2^{k_2-1} e^{-\Lambda(q,p)},$$

so that

$$\begin{aligned}
& \mathrm{pr}\left\{\Lambda(q,p) \geq \delta(p)\right\} \\
= \ & \frac{(k_1 + k_2 + 1)!}{(k_1 - 1)!(k_2 - 1)!} \int_{\Lambda(q,p)\geq\delta(p)} p_1^{k_1-1} p_2^{k_2-1}(1 - p_1 - p_2) dp_1 dp_2 \\
\leq \ & \frac{(k_1 + k_2 + 1)!}{(k_1 - 1)!(k_2 - 1)!} q_1^{k_1-1} q_2^{k_2-1} \int_{\Lambda(q,p)\geq\delta(p)} e^{-\Lambda(q,p)} dp_1 dp_2 \\
\leq \ & (k_1 + k_2)^3 \frac{(k_1 + k_2 - 2)!}{(k_1 - 1)!(k_2 - 1)!} q_1^{k_1-1} q_2^{k_2-1} \int_{\Lambda(q,p)\geq\delta(p)} e^{-\Lambda(q,p)} dp_1 dp_2 \\
\leq \ & (k_1 + k_2)^3 \int_0^1 e^{-\delta(p)} dp.
\end{aligned}$$

□

Proof of Theorem 1.2. If the slice scheme on X has only one slice, we have

$$D(Y \mid S, \lambda_0) = \log \hat{v}^2 - \log \widehat{\sigma}^2 = -\log(1 - R^2),$$

where $\widehat{\sigma}^2$ is the residual variance after regressing Y on X. Intuitively, if Y and X follow a bivariate normal, the optimal slice scheme is only one slice in each direction. Now, we show that

$$\mathrm{pr}\left\{D(Y \mid X, \lambda_0) + \log(1 - R^2) > 0\right\} < 1.5 n^{-\lambda_0/3+5}.$$

For any slice scheme S,

$$D(Y \mid S, \lambda_0) + \log(1 - R^2) = \log \widehat{\sigma}^2 - \sum_{s \in S} \frac{n_s}{n} \log(\widehat{\sigma}_s^2) - \frac{\lambda_0}{n}(|S| - 1)\log n.$$

Without loss of generality, we assume that $\mathrm{var}(Y) = 1$ and $x_1 < \ldots < x_n$. Suppose the connected slices each has n_i $(i = 1, \ldots |S|)$ observations. For $1 \leq j < k \leq n$, define

$$\Delta(j, k, \lambda_0) = \frac{k}{n} \log\{\widehat{\sigma}^{(k)}\}^2 - \frac{j}{n} \log\{\widehat{\sigma}^{(j)}\}^2 - \frac{k-j}{n} \log\{\widehat{\sigma}^{(k,j)}\}^2 - \frac{\lambda_0}{n} \log n.$$

Here, $\{\widehat{\sigma}^{(j)}\}^2$ is the residual variance of regressing y_i on x_i $(i = 1, \ldots, j)$, $\{\widehat{\sigma}^{(k)}\}^2$ is the residual variance of regressing y_i on x_i $(i = 1, \ldots, k)$ and $\{\widehat{\sigma}^{(k,j)}\}^2$ is the residual variance of regressing y_i on x_i $(i = j+1, \ldots, k)$. For given j, k, let

$$p_1 = \frac{j\{\widehat{\sigma}^{(j)}\}^2}{k\{\widehat{\sigma}^{(k)}\}^2}, \quad p_2 = \frac{(k-j)\{\widehat{\sigma}^{(k,j)}\}^2}{k\{\widehat{\sigma}^{(k)}\}^2}, \quad q_1 = \frac{j}{k}, \quad q_2 = 1 - q_1.$$

Then according to Cochran's theorem, we have

$$\begin{aligned} &(p_1, p_2, 1 - p_1 - p_2) \sim \mathrm{Dir}(j - 2, k - j - 2, 2), \\ &n\Delta(j, k, \lambda_0) = \Lambda(q, p) - \lambda_0 \log(n) + 3\log(q_1/p_1) + 3\log(q_2/p_2). \end{aligned}$$

By Lemma A.2 we have

$$\mathrm{pr}\{\Lambda(q, p) > \lambda_0 \log(n)/3\} \quad \leq \quad k^3 n^{-\lambda_0/3} \leq n^{-\lambda_0/3+3}.$$

At the same time,

$$\begin{aligned}
& \operatorname{pr}\left\{3 \log \left(q_1/p_1\right) > \lambda_0 \log(n)/3\right\} \\
= \; & \frac{(k-3)!}{(j-3)!(k-j-1)!} \int_0^{q_1 n^{-\lambda_0/9}} p^{j-3}(1-p)^{k-j-1} dp \\
\leq \; & \frac{(k-3)!}{(j-3)!(k-j-1)!} \frac{1}{j-2} (q_1 n^{-\lambda_0/9})^{j-2} \\
= \; & (j/k)^{j-2} \frac{(k-3)!}{(j-2)!(k-j-1)!} \frac{1}{n^{\lambda_0(j-2)/9}} \leq \frac{1}{n^{(j-2)(\lambda_0/9-1)}}
\end{aligned}$$

If $n \geq 25$, we have $\operatorname{pr}\{\Delta(j,k,\lambda_0) > 0\} \leq 3n^{-\lambda_0/3+3}$. On the other hand, for any slicing scheme with $|S| \geq 2$, $D(Y \mid S, \lambda_0) + \log(1-R^2)$ equals

$$\sum_{h=1}^{|S|-1} \Delta(\sum_{l=1}^{h} n_l, \sum_{l=1}^{h+1} n_l, \lambda_0)$$

So

$$\begin{aligned}
& \operatorname{pr}\left\{D(Y \mid X, \lambda_0) + \log(1-R^2) > 0\right\} \\
\leq \; & \operatorname{pr}\left\{\max_{m \leq j < k \leq n-m} \Delta(j,k,\lambda_0) > 0\right\} \\
\leq \; & \sum_{m \leq j < k \leq n-m} \operatorname{pr}\{\Delta(j,k,\lambda_0) > 0\} < 1.5n^{-\lambda_0/3+5}.
\end{aligned}$$

Since X and Y are symmetric, the result tells us that $P\left\{G^2_{\mathrm{m}}(\lambda_0) = R^2\right\} > 1 - 3n^{-\lambda_0/3+5}$. When $\lambda_0 > 18$, we have $G^2_{\mathrm{m}}(\lambda_0) = R^2$ almost surely. □

A.4 More simulations

A.4.1 Power analysis

Table A.1 lists twenty relationships for power analysis. For all relationships, we normalize them so that var$\{f(X)\} = 1$ with $X \sim U(0,1)$. As an intuitive presentation, Figure A.2

shows the twenty simulated relationships with $G^2_{Y|X} = 0.8$. The power analysis results with six methods for the first eight relationships are in Chapter 1. Figure A.3 presents the power for the eight relationships with the remaining six methods. The power analysis of the remaining twelve relationships with the entire twelve methods are in Fig.s A.4–A.6. Figures A.5 and A.6 have the same legend as Fig. A.4. We find G^2_{m} and G^2_{t} are among the most powerful test statistics and G^2_{t} shows a higher power than G^2_{m} in most examples.

Table A.1: Relationships for power analysis.

relationship name	function
linear	x
quadratic	$(x-1/2)^2$
cubic	$32(x-1/3)^3 - 12(x-1/3)^2 - 3(x-1/3)$
radical	$x^{0.25}$
low freq sine	$\sin(2\pi x)$
triangle	$(1-x)I_{x<0.5} + xI_{x\geq 0.5}$
high freq sine	$\sin(8\pi x)$
piecewise constant	$0.287I_{x\leq 0.2} + 0.796I_{0.2<x\leq 0.4} + 0.290I_{0.4<x\leq 0.6}$
	$+0.924I_{0.6<x\leq 0.8} + 0.717I_{x>0.8}$
unimodal cubic	$32(x-2/3)^3 - 12(x-2/3)^2 - 3(x-2/3)$
low order polynomial	$x^4(1-x)$
high order polynomial	$x(1-x)^9$
reciprocal	$1/(x+0.5)$
L-shaped	$(x/90)I_{x\leq 0.9} + (90x-81)I_{x>0.9}$
lopsided L-shaped	$200xI_{x\leq 0.005} + (-198x+19.9)I_{0.005<x\leq 0.01} + (-x/99+1/99)I_{x>0.1}$
spike	$20xI_{x\leq 0.05} + (-18x+1.9)I_{0.05<x\leq 0.1} + (-x/9+1/9)I_{x>0.1}$
sigmoid	$\{50(x-0.5)+0.5\}I_{0.4<x\leq 0.6} + I_{x>0.6}$
medium freq sine	$\sin(4\pi x)$
very high freq sine	$\sin(16\pi x)$
sine with drift	$\sin\{2\pi(2x-1)\} + (2x-1)/2$
vary freq sine	$\sin\{4\pi x(1+x)\}$

A.4.2 Influence of sample size

We run simulations with the same setup with $n = 50,\ 100,\ 225$ and 500. Figure A.7 shows the average power of G^2_{m}, G^2_{t}, COR, DCOR, DDP and TIC$_e$ against different sample sizes. We find that G^2_{m} and G^2_{t} are among the most powerful methods when n is larger

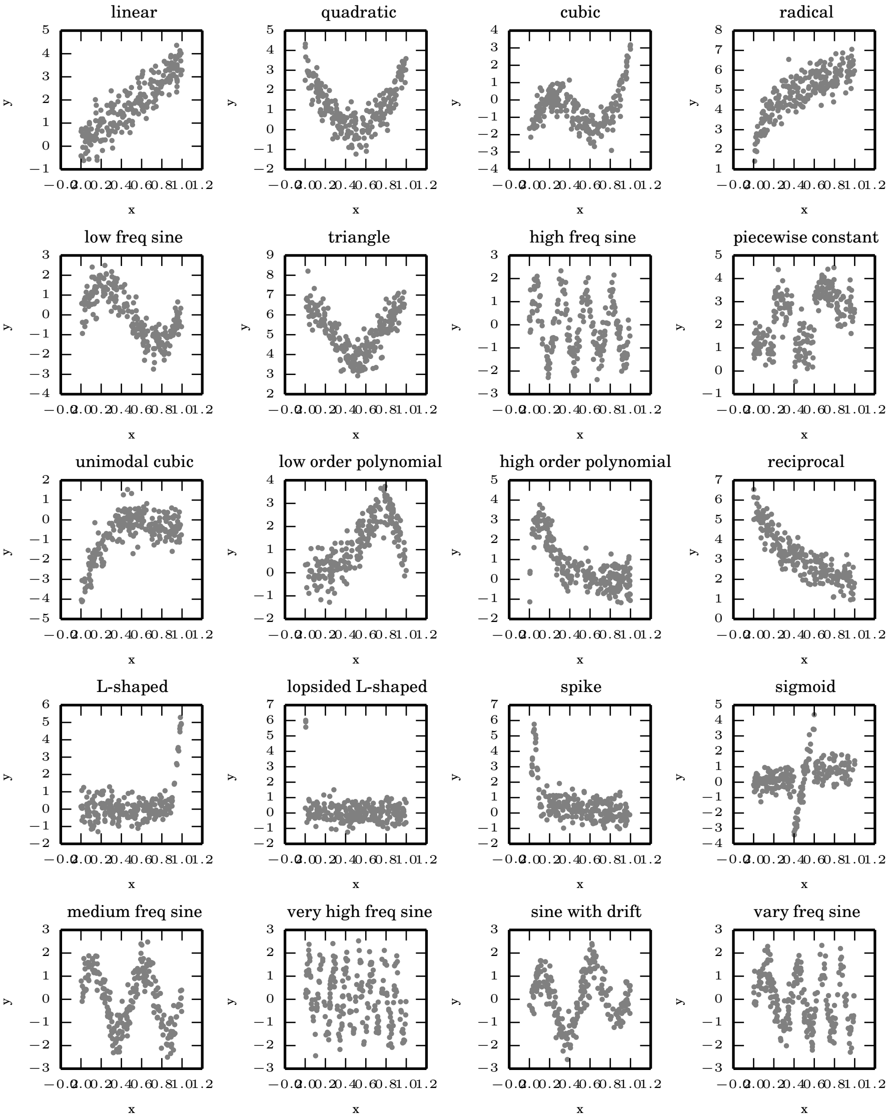

Figure A.2: Scatter plots for the twenty relationships in Table A.1 with $n = 225$. We choose $\sigma = 0.5$ for each relationship so $G^2_{Y|X} = 0.8$.

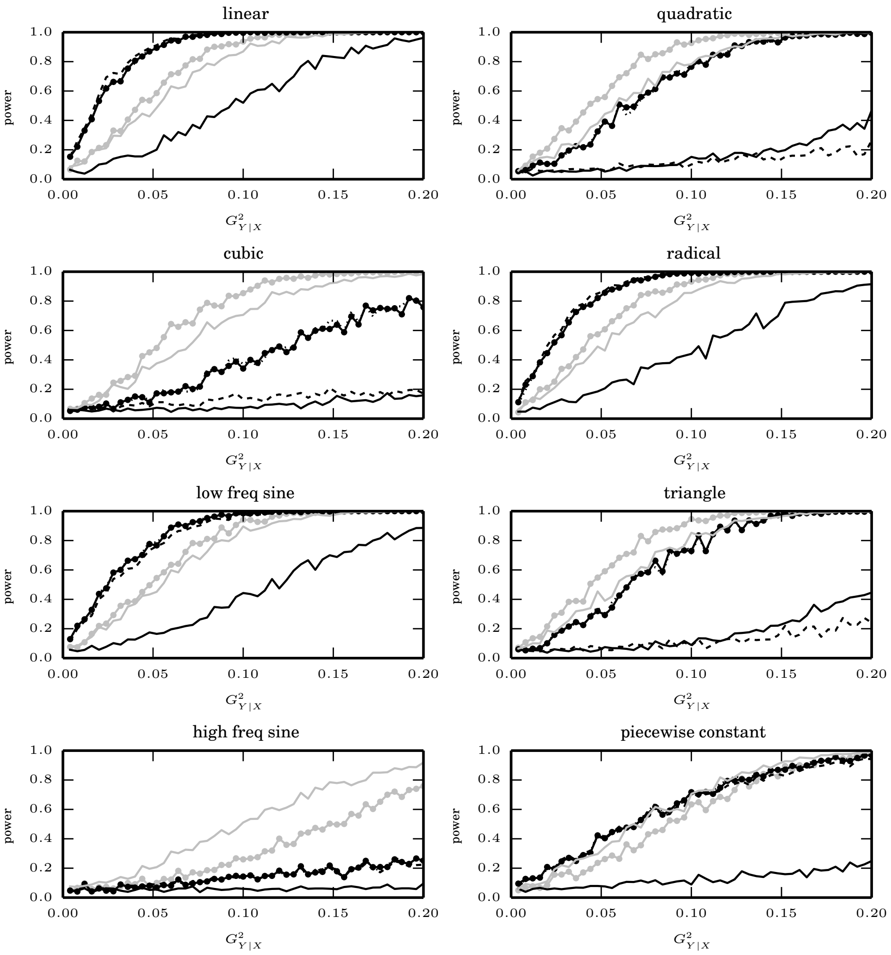

Figure A.3: The powers of mutual information (black solid), MIC$_e$ (grey solid), ACE (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers) for independence test between X and Y when the relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant. The x-axis is $G^2_{Y|X}$ and the y-axis is the power.

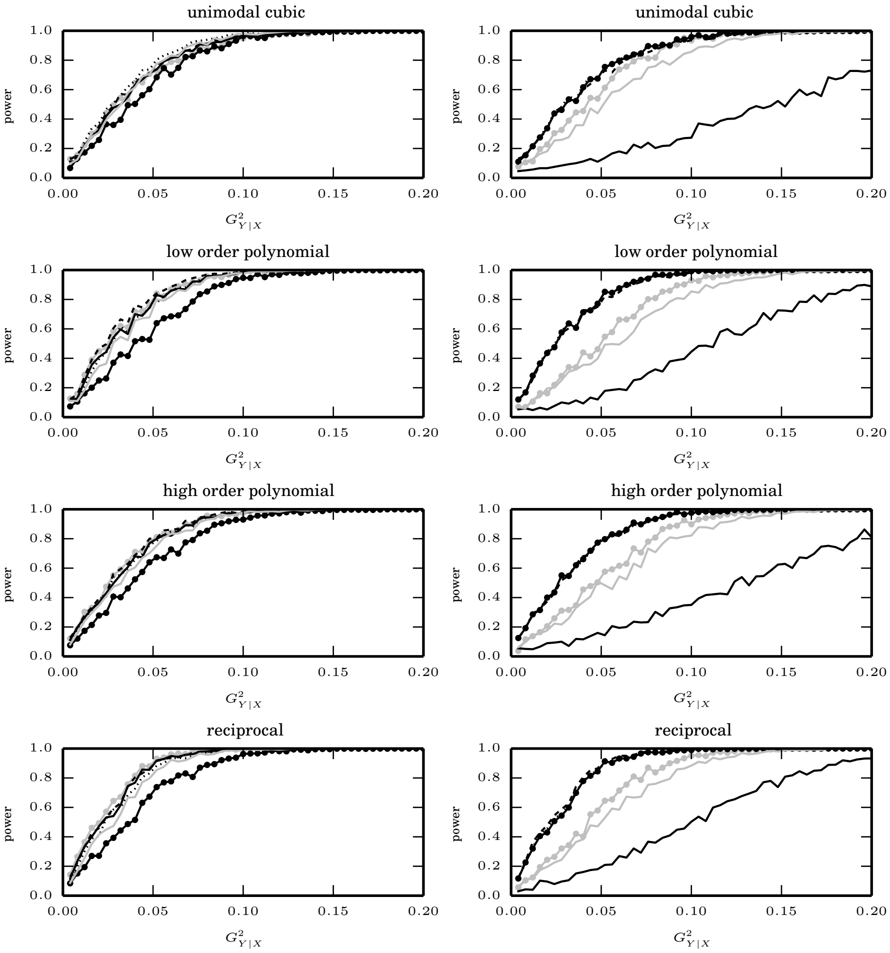

Figure A.4: The left column presents the powers of G^2_{m} (black solid), G^2_{t} (grey solid), COR (grey markers), DCOR (black dashes), DDP (black dots) and TIC$_e$ (black markers) for independence test between X and Y when the relationships are power functions; the right column presents the powers of mutual information (black solid), MIC$_e$ (grey solid), ACE (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers). The x-axis is $G^2_{Y|X}$ and the y-axis is the power.

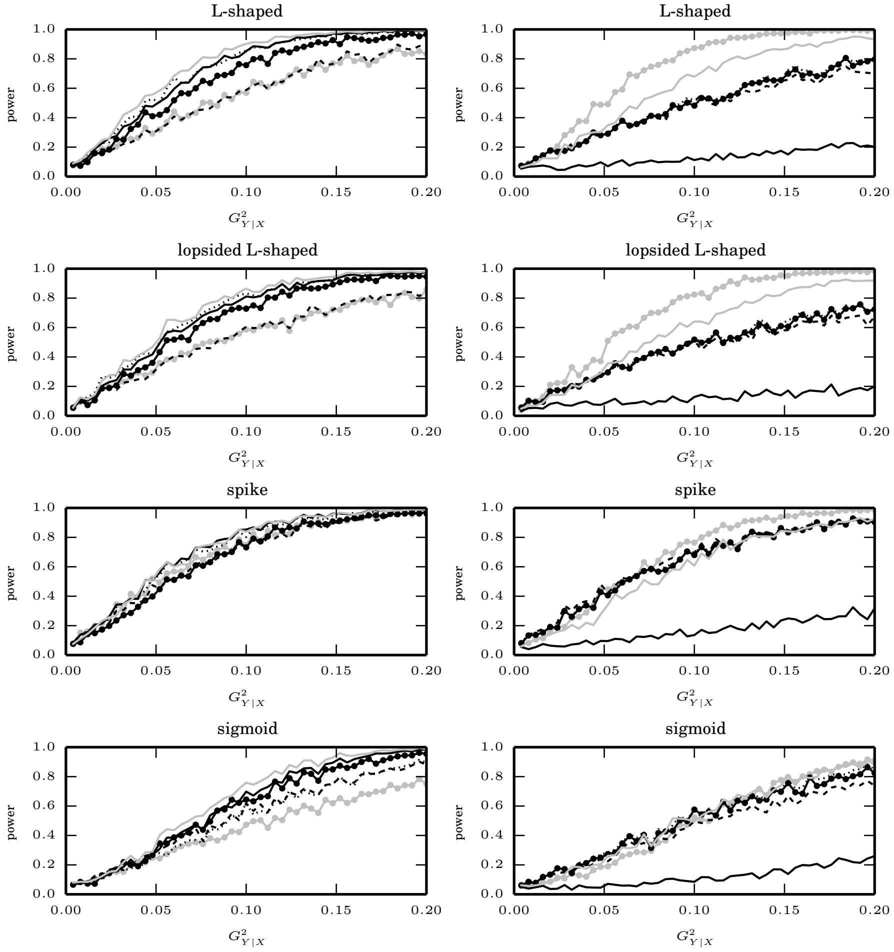

Figure A.5: The powers for independence test between *X* and *Y* when the relationships are piecewise linear functions. The legends is the same as in Fig. A.4.

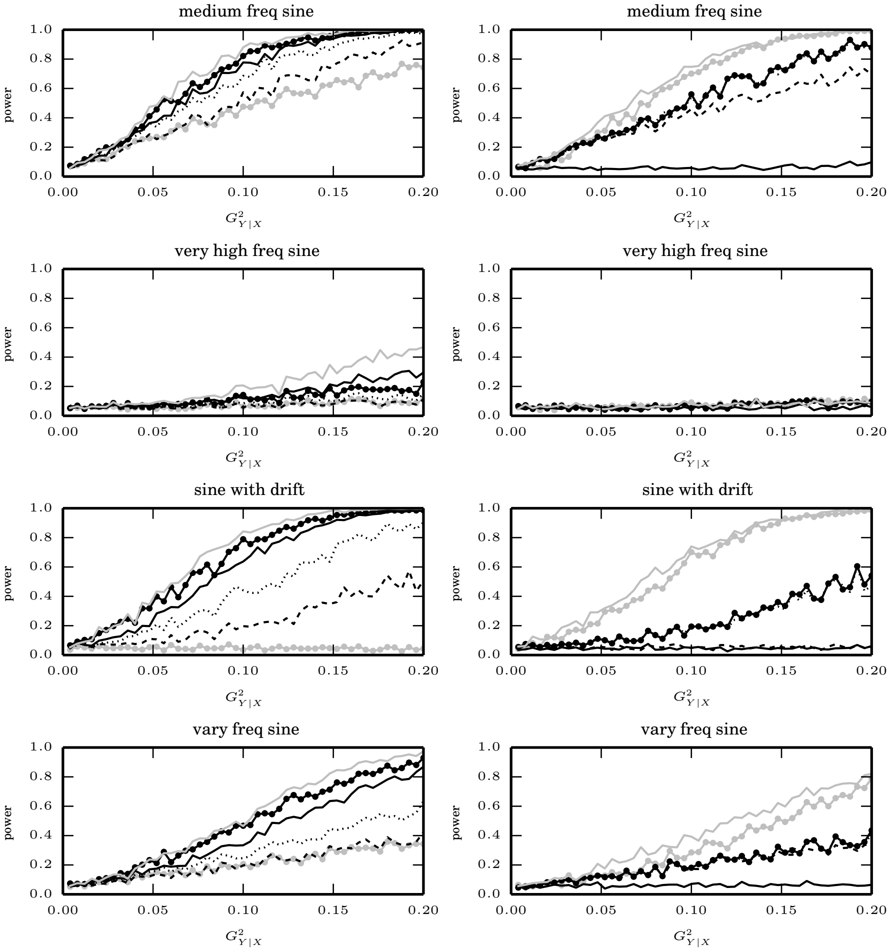

Figure A.6: The powers for independence test between X and Y when the relationships are trigonometric functions. The legends is the same as in Fig. A.4.

than 100. When the sample size is small, the powers of G^2_m and G^2_t are slightly lower than DDP in some cases but are still among the most powerful methods. Power analysis for more relationships are in Fig.s A.8–A.10.

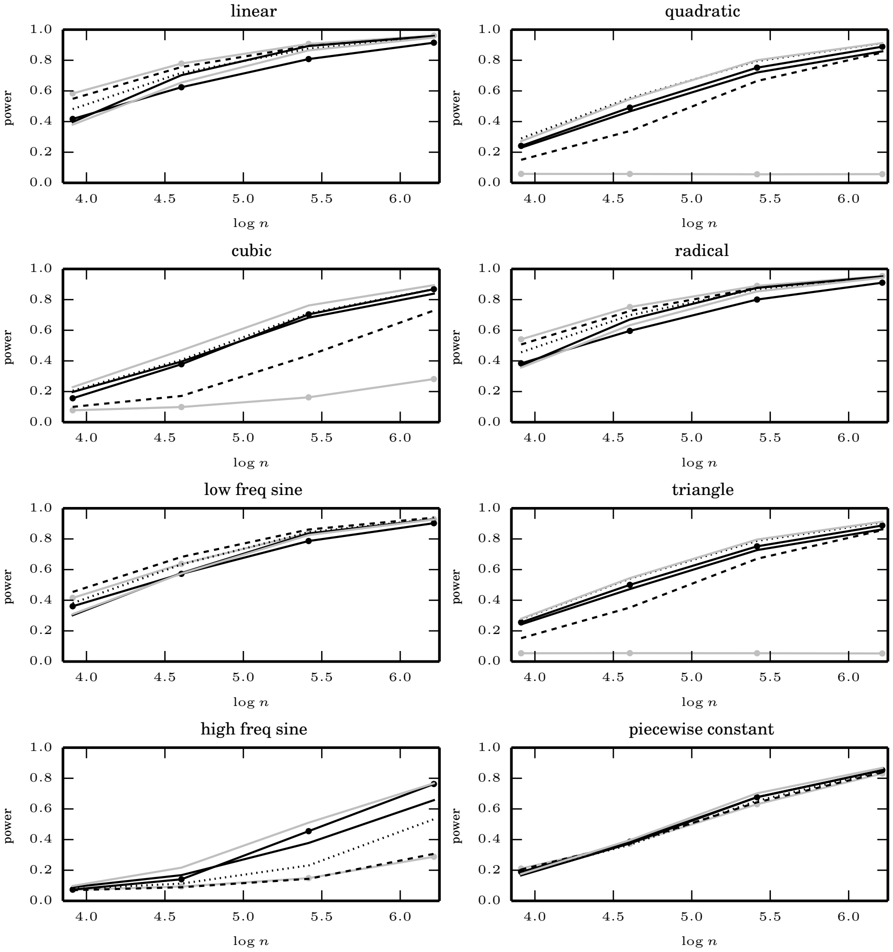

Figure A.7: The average powers of G^2_m (black solid), G^2_t (grey solid), COR (grey markers), DCOR (black dashes), DDP and TIC$_e$ (black markers) for testing independence between X and Y with $n = 50,\ 100,\ 225$ and 500. The underlying true relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant. The x-axis is logarithm of n with base 10 and the y-axis is the average power.

A.5 Relationships for equitability study

The relationships for equitability study are in Table A.2.

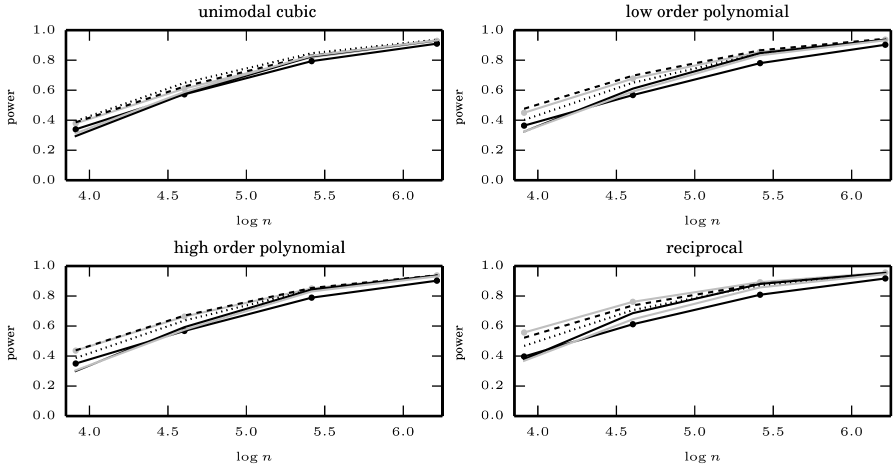

Figure A.8: The average powers for independence test between *X* and *Y* when the relationships are power functions. The legends is the same as in Fig. A.7.

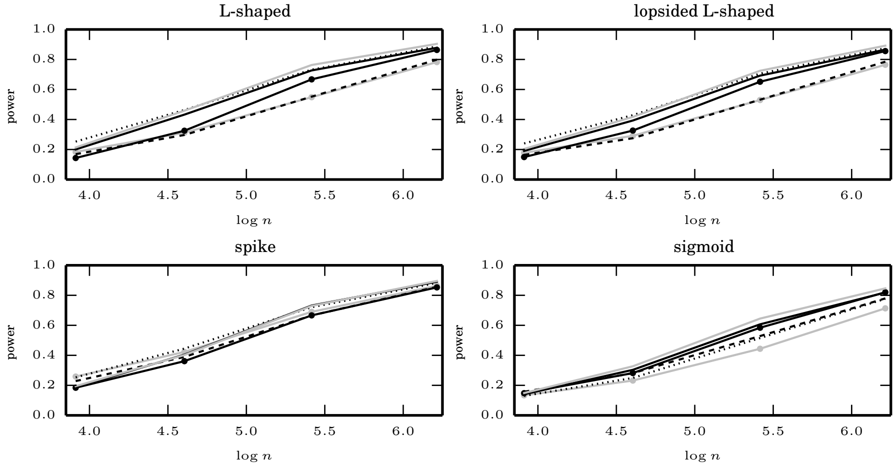

Figure A.9: The average powers for independence test between *X* and *Y* when the relationships are piecewise linear functions. The legends is the same as in Fig. A.7.

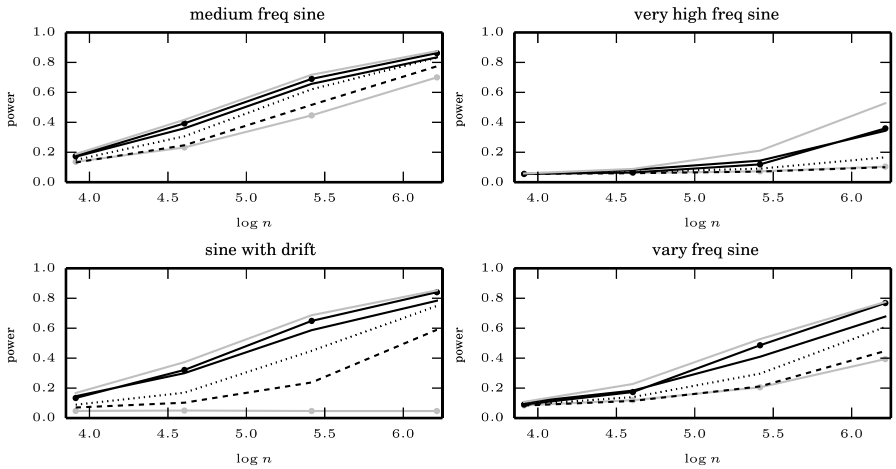

Figure A.10: The average powers for independence test between X and Y when the relationships are trigonometric functions. The legends is the same as in Fig. A.7.

Table A.2: Relationships for equitability study.

relationship name	function
line	x
quadratic	$(x-1/2)^2$
cubic	$4(2.4x-1.3)^3+(2.4x-1.3)^2-4(2.4x-1.3)$
exponential (10^x)	10^{10x}
exponential (2^x)	2^{2x}
L-shaped	$(x/99)I_{x\leq 0/99}+1I_{x>0.99}$
lopsided L-shaped	$200xI_{x\leq 0.005}+(-198x+19.9)I_{0.005<x\leq 0.01}+(-x/99+1/99)I_{x>0.1}$
spike	$20I_{x\leq 0.05}+(-18x+1.9)I_{0.05<x\leq 0.1}+(-x/9+1/9)I_{x>0.1}$
sigmoid	$\{50(x-0.5)+0.5\}I_{0.49<x\leq 0.51}+1I_{x>0.51}$
linear + high freq periodic	$0.1\sin\{10.6(2x-1)\}+1.1(2x-1)$
linear + high freq periodic 2	$0.2\sin\{10.6(2x-1)\}+1.1(2x-1)$
linear + low freq periodic	$0.2\sin\{4(2x-1)\}+1.1(2x-1)$
linear + medium freq periodic	$\sin(10\pi x)+x$
high freq sine	$\sin(8\pi x)$
non-Fourier freq sine	$\sin(9\pi x)$
very high freq sine	$\sin(16\pi x)$
varying freq sine	$\sin\{6\pi x(1+x)\}$
high freq cosine	$\cos(14\pi x)$
non-Fourier freq cosine	$\cos(7\pi x)$
varying freq cosine	$\sin\{5\pi x(1+x)\}$

B

Supplementary materials for Chapter 2

B.1 Proof of consistency

Lemma B.1. *Suppose X and Y are univariate continuous random variables with $|X|, |Y| < B$ and $\mathrm{var}(X),\ \mathrm{var}(Y) > b^{-2}$. Given n observations as (x_i, y_i) $(i = 1, \ldots, n)$, let $\hat{\beta}$ be the slope coefficient after regressing Y on X. Then*

$$\mathrm{pr}\left\{\left|\hat{\beta} - \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}\right| > \varepsilon\right\} \leq 8\exp\{-C_2(B, b)n\varepsilon^2\},$$

with $C_2(B, b) = (72b^4B^4)^{-1}\min\{1, (4b^2B^2)^{-1}\}$.

Proof of Lemma B.1. Without loss of generality, we assume $EX = EY = 0$, $\mathrm{var}(X) = \mathrm{var}(Y) = 1$ and $E(XY) = \rho$. By definition, we have

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)(\frac{1}{n}\sum_{i=1}^{n} y_i)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}$$

Then $x_i^2, y_i^2 \in [0, B^2]$, $x_i y_i \in [-B^2, B^2]$. According to Hoeffding's inequality, we have

$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| > \varepsilon/6\right), \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i\right| > \varepsilon/6\right), \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1\right| > \varepsilon/6\right),$$
$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \rho\right| > \varepsilon/6\right) \leq 2\exp\{-c(B)n\varepsilon^2\}$$

with $c(B) = (72B^2)^{-1}\min\{1, B^{-2}\}$. If $\varepsilon < 1/2$ and

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right|, \left|\frac{1}{n}\sum_{i=1}^{n} y_i\right|, \left|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1\right|, \left|\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \rho\right| \leq \varepsilon/6,$$

we can derive that

$$\begin{aligned}\left|\hat{\beta} - \rho\right| &\leq \frac{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 - 1||\rho|}{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2|} \\ &\quad \frac{+|\frac{1}{n}\sum_{i=1}^{n} x_i||\frac{1}{n}\sum_{i=1}^{n} y_i| + |\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \rho|}{|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2|} \\ &\leq \frac{2(\varepsilon/6 + \varepsilon^2/36)}{1 - \varepsilon/6 - \varepsilon^2/36} < \varepsilon.\end{aligned}$$

So we can conclude that

$$\mathrm{pr}\left(\left|\hat{\beta} - \rho\right| > \varepsilon\right) \leq 8\exp\{-c(B)n\varepsilon^2\}.$$

For general cases, define $X' = (X - EX)/\mathrm{sd}(X)$, $Y' = (Y - EY)/\mathrm{sd}(Y)$. Then $EX' = EY' = 0$,

$\text{var}(X') = \text{var}(Y') = 1$ and $|X'|, |Y'| < 2bB$. Thus,

$$\begin{aligned} \text{pr}\left\{\left|\hat{\beta} - \frac{\text{cov}(X,Y)}{\text{var}(X)}\right| > \varepsilon\right\} &= \text{pr}\left\{\left|\hat{\beta}' - \text{cov}(X',Y')\right| > \frac{\varepsilon sd(X)}{sd(Y)}\right\} \\ \leq \ 8\exp\left\{-\frac{c(2bB)\text{var}(X)}{\text{var}(Y)}n\varepsilon^2\right\} &= 8\exp\{-C_2(B,b)n\varepsilon^2\}, \end{aligned}$$

with $C_2(B,b) = (288b^4B^4)^{-1}\min\{1,(4b^2B^2)^{-1}\}$. □

Proof of Theorem 2.1. First, we prove the first part of the theorem. Let $\lambda(n) = \lambda_0 \log(n)/n$. By definition, we know that

$$S_n = \text{argmin}_{m_S \geq m} \sum_{s \in S} \frac{n_s}{n} \log \widehat{\sigma}_s^2 + (|S| - 1)\lambda(n)$$

For each slice s, define $\sigma_s^2 = \text{var}_s(Y) - \text{cov}_s^2(X,Y)/\text{var}_s(X)$ and $p_s = P(X \in s)$. Then

$$\sigma_s^2 \geq \text{var}_s(Y) - \text{var}_s\{E(Y \mid X)\} = \sigma^2.$$

Step 1: We show that there is $\eta_3(n) > 0$, and $\eta_3(n) \to 0$ as $n \to \infty$ such that

$$\text{pr}\left[\liminf_{n\to\infty}\sum_{s\in S_n}\frac{n_s}{n}\widehat{\sigma}_s^2 > \sigma^2 - \eta_3(n)\right] = 1.$$

Because for any slicing scheme S, $\sum_{s\in S} p_s\sigma_s^2 \geq \sigma^2$, it is enough to show that there is $\eta_3(n)$ such that

$$\text{pr}\left\{\liminf_{n\to\infty}\sum_{s\in S_n}\frac{n_s}{n}\widehat{\sigma}_s^2 > \sum_{s\in S_n} p_s\sigma_s^2 - \eta_3(n)\right\} = 1.$$

Let $\delta(n) = \log(n)n^{-1/4}$. We have

$$\begin{aligned} & \sum_{s\in S_n} \frac{n_s}{n}\widehat{\sigma}_s^2 - \sum_{s\in S_n} p_s\sigma_s^2 \\ = & \sum_{s\in S_n} (\frac{n_s}{n} - p_s)\sigma_s^2 + \sum_{s\in S_n} \frac{n_s}{n}\left(\widehat{\sigma}_s^2 - \sigma_s^2\right). \end{aligned}$$

First, we consider $\sum_{s\in S_n}(\frac{n_s}{n} - p_s)\sigma_s^2$. Let us define a new random variable: Z, and $Z = \sigma_s^2$ if X is in slice s, and let z_i $(i = 1, \dots n)$ be n independent observations of Z. Then,

$$E(Z) = \sum_{s\in S_n} p_s\sigma_s^2, \quad \frac{1}{n}\sum_{i=1}^{n} z_i = \sum_{s\in S_n} \frac{n_s}{n}\sigma_s^2.$$

By Hoeffding's inequality and the fact that $\sigma_s^2 \in [b^{-2}, B^2]$,

$$\text{pr}\left\{\sum_{s\in S_n}(\frac{n_s}{n} - p_s)\sigma_s^2 < -\delta(n)\right\} \leq \exp\{-2n\delta(n)^2/B^4\}. \qquad \text{(B.1)}$$

Second, we focus on the difference between $\widehat{\sigma}_s^2$ and σ_s^2. Consider a slicing scheme Q_n of n^4 slices such that an observation falls in each slice equally. Given n observations, the probability for any of the n^4 slices containing more than one observations is smaller than $n^4\left\{1 - \left(1 + n^{-3}\right)\left(1 - n^{-4}\right)^n\right\} \leq n^{-2}$. Then event

$$E_{1,n} = \{\text{each slice of } Q_n \text{ has at most one observation}\}$$

satisfies $\text{pr}\left(\liminf_{n\to\infty} E_{1,n}\right) = 1$. Thus, we only need to consider slicing schemes that are more refined than Q_n, denoted as $S \preceq Q_n$. Define the set of slices

$$\Xi = \{s | \text{there exists } S \preceq Q_n \text{ such that } s \in S\}.$$

The set Ξ contains at most $n^4(n^4+1)/2 = O(n^8)$ slices. Each slice $s \in \Xi$ contains at least m observations.

By Lemma A.1, if $\delta(n) < 0.5b^{-2}$,

$$\text{pr}\left\{\widehat{\sigma}_s^2 - \sigma_s^2 < -\delta(n)\right\} \leq 10e^{-C_1(B,2\sigma^{-1})\delta(n)^2 m} \leq 10n^{-C_1(B,2\sigma^{-1})\log(n)}. \qquad \text{(B.2)}$$

Let $\eta_3(n) = 2\delta(n)$ and event

$$E_{2,n} = \left\{\min_{S \preceq Q_n} \sum_{s \in S} \frac{n_s}{n} \widehat{\sigma}_s^2 > \sigma^2 - \eta_3(n)\right\}.$$

Combine the result of (B.1) and (B.2), we have $\text{pr}\left(\liminf_{n\to\infty} E_{1,n} \cap E_{2,n}\right) = 1$, which means that $n^{-1}\sum_{s\in S_n} n_s \widehat{\sigma}_s^2$ is almost surely larger than σ^2.

Step 2: Next, we show that there exists $\eta_4(n) > 0$, and $\eta_4(n) \to 0$ as $n \to \infty$ such that

$$\text{pr}\left[\limsup_{n\to\infty} \sum_{s\in S_n} \frac{n_s}{n} \widehat{\sigma}_s^2 < \sigma^2 + \eta_4(n)\right] = 1.$$

We already know that there $\eta_2(n) > 0$ and $\eta_2(n) \to \infty$ as $n \to \infty$, such that

$$\text{pr}\left[\limsup_{n\to\infty} \sum_{s\in S_n} \frac{n_s}{n} \log \widehat{\sigma}_s^2 < \log \sigma^2 + \eta_2(n)\right] = 1.$$

We know

$$\begin{aligned}
& \sum_{s\in S_n} \frac{n_s}{n} \log \widehat{\sigma}_s^2 - \log \sigma^2 \\
= & \sum_{s\in S_n} \frac{n_s}{n} (\log \widehat{\sigma}_s^2 - \log \sigma_s^2) - \sum_{s\in S_n} \frac{n_s}{n} (\log \sigma^2 - \log \sigma_s^2) \\
\geq & \sum_{s\in S_n} \frac{n_s}{n} \{ \frac{\widehat{\sigma}_s^2}{\sigma_s^2} - 1 - \delta(n)^2 B^{-4} \} - \sum_{s\in S_n} \frac{n_s}{n} (\frac{\sigma^2}{\sigma_s^2} - 1) \\
\geq & \sum_{s\in S_n} \frac{n_s \widehat{\sigma}_s^2}{n\sigma_s^2} - \sum_{s\in S_n} \frac{n_s \sigma^2}{n\sigma_s^2} - \delta(n)^2 B^{-4},
\end{aligned}$$

if

$$\frac{\widehat{\sigma}_s^2}{\sigma_s^2} \in [1 - \delta(n)B^{-2}, 1 + \delta(n)B^{-2}]$$

and if $\delta(n)B^{-2} < 0.5$. Besides,

$$\text{pr}\left\{ |\frac{\widehat{\sigma}_s^2}{\sigma_s^2} - 1| > \delta(n)B^{-2} \right\} \leq \text{pr}\left\{ |\widehat{\sigma}_s^2 - \sigma_s^2| > \delta(n)B^{-2}\sigma^2 \right\} \leq 10 e^{-C_1(B, 2\sigma^{-1})\sqrt{n}\delta(n)^2 B^{-4}\sigma^4}.$$

Because $\sigma_s^2 \geq \sigma^2$, under (B.1), we have

$$\begin{aligned}
\sum_{s\in S_n} \frac{n_s}{n} (\widehat{\sigma}_s^2 - \sigma^2) = & \sum_{s\in S_n} \frac{n_s}{n} (\widehat{\sigma}_s^2 - \sigma_s^2) + \sum_{s\in S_n} \frac{n_s(\sigma_s^2 - \sigma^2)}{n\sigma_s^2} \sigma_s^2 \\
\leq & \delta(n) + B^2 \sum_{s\in S_n} \frac{n_s(\sigma_s^2 - \sigma^2)}{n\sigma_s^2} \\
\leq & \delta(n) + B^2 \sum_{s\in S_n} \frac{n_s(\sigma_s^2 - \widehat{\sigma}_s^2)}{n\sigma_s^2} + B^2 \sum_{s\in S_n} \frac{n_s(\widehat{\sigma}_s - \sigma^2)}{n\sigma_s^2} \\
\leq & 2\delta(n) + B^2 \{ \eta_2(n) + \delta(n)^2 B^{-4} \}
\end{aligned}$$

Let $\eta_4(n) = 2\delta(n) + \delta(n)^2 B^{-2} + B^2 \eta_2(n)$, by (B.2) and (B.2), we have

$$\text{pr}\left\{ \limsup_{n\to\infty} \sum_{s\in S_n} \frac{n_s}{n} \widehat{\sigma}_s^2 < \sigma^2 + \eta_4(n) \right\} = 1.$$

For the second part of the theorem, denote that $f_i = f(X_i), \hat{f}_i = \hat{f}(X_i)$, then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(f_i + e_i - \hat{f}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(f_i - \hat{f}_i)^2 + \frac{1}{n}\sum_{i=1} e_i^2 + \frac{2}{n}\sum_{i=1}^{n} e_i(f_i - \hat{f}_i)\end{aligned}$$

where $e_i = y_i - f_i$ and $\frac{1}{n}\sum_{i=1}^{n} e_i^2 \to \sigma^2$. We show that there is $\eta_5(n) > 0$ axnd $\eta_5(n) \to 0$ as $n \to \infty$ such that

$$\text{pr}\left\{\limsup_{n\to\infty} \frac{1}{n}|\sum_{i=1}^{n} e_i(\hat{f}_i - f_i)| < \eta_5(n)\right\} = 1.$$

Similar as before, we only need to consider slicing schemes that are more refined than Q_n. For slice $s \in \Xi$, we have

$$\begin{aligned}\sum_{i\in s} e_i(\hat{f}_i - f_i) &= \sum_{i\in s} e_i\left[\bar{f}_s - f_i + \bar{e}_s + \hat{\beta}_s(x_i - \bar{x}_s)\right] \\ &= n_s\left(\bar{e}_s^2 + \bar{e}_s\bar{f}_s - \hat{\beta}_s\bar{e}_s\bar{x}_s\right) - \sum_{i\in s} e_i f_i + \hat{\beta}_s \sum_{i\in_s} e_i x_i\end{aligned}$$

Because $f_i \in [-B, B]$, $e_i \in [-2B, 2B]$, $e_i f_i \in [-2B^2, 2B^2]$ and $e_i x_i \in [-2B^2, 2B^2]$. By Hoeffding's inequality,

$$\begin{aligned}&\text{pr}\left\{|\bar{e}_s| > \delta(n)\right\}, \quad \text{pr}\left\{\frac{1}{n_s}|\sum_{i\in s} e_i f_i| > \delta(n)\right\}, \quad \text{pr}\left\{\frac{1}{n_s}|\sum_{i\in s} e_i x_i| > \delta(n)\right\} \\ \leq\; & 2\exp\{-C_3(B) n_s \delta(n)^2\} \leq 2n^{-C_3(B)\log(n)}\end{aligned}$$

with $C_3(B) = (2B^2)^{-1}\min\{1,\ B^{-2}\}$. Then if

$$|\bar{e}_s| \leq \delta(n), \quad \frac{1}{n_s}|\sum_{i\in s} e_i f_i| \leq \delta(n), \quad \frac{1}{n_s}|\sum_{i\in s} e_i x_i| \leq \delta(n),$$

we have

$$\frac{1}{n_s}|\sum_{i\in s} e_i(\hat{f}_i - f_i)| \quad \leq \quad |\bar{e}_s|(4B + B|\hat{\beta}_s|) + \frac{1}{n_s}|\sum_{i\in s} e_i f_i| + \frac{1}{n_s}|\hat{\beta}_s||\sum_{i\in S} e_i x_i|.$$

By Lemma B.1, if $\delta(n) < B^2\sigma^{-2}$,

$$\mathrm{pr}\left(|\hat{\beta}| > 2B^2\sigma^{-2}\right) \leq 8\exp\left\{-C_2(B, 2\sigma^{-1})\delta(n)^2\sqrt{n}\right\}.$$

Let $\eta_5(n) = (1 + 4B + 4B^2b^2 + 4B^3b^2)\delta(n)$, then

$$\mathrm{pr}\left\{\limsup_{n\to\infty}\frac{1}{n}|\sum_{i=1}^{n} e_i(\hat{f}_i - f_i)| < \eta_5(n)\right\} = 1,$$

which means $\frac{1}{n}\sum_{i=1}^{n}(f_i - \hat{f}_i)^2 \to 0$.

□

B.2 More simulations

We present the average ISEs for $t = 1$ and 2 in Tables B.1–B.5. As the sample size increases, our method can outperform the other methods. We also present the false negative and false positive selections in Tables B.6–B.10. Our methods perform better with larger sample size ($n \geq 225$). When the number of predictors is not large, we suggest the AMGS-BIC for a smaller prediction error. When the number of predictors is large, we suggest MGS-AGL for a more consistent variable selection result.

Table B.1: The average ISEs with $p = 10$ for k_j $(j = 1, 2, 3, 4)$

		$t = 1$	
n	100	225	400
MGS-AGL	1.50 (1.55)	0.60 (0.34)	0.32 (0.21)
MGS-BIC	1.48 (2.98)	**0.45 (0.30)**	**0.24 (0.18)**
SPAM	**1.05 (0.30)**	0.50 (0.12)	0.34 (0.07)
AGL	1.65 (1.25)	0.56 (0.24)	0.33 (0.12)
		$t = 2$	
n	100	225	400
MGS-AGL	1.93 (12.68)	0.39 (0.31)	0.21 (0.11)
MGS-BIC	1.89 (8.84)	**0.26 (0.23)**	**0.13 (0.08)**
SPAM	**1.10 (0.33)**	0.49 (0.14)	0.31 (0.08)
AGL	1.66 (1.52)	0.54 (0.24)	0.30 (0.10)

Table B.2: The average ISEs with $p = 20$ for k_j $(j = 1, 2, 3, 4)$

		$t = 1$	
n	100	225	400
MGS-AGL	1.61 (1.64)	0.59 (0.31)	0.33 (0.21)
MGS-BIC	1.74 (3.81)	**0.46 (0.28)**	**0.25 (0.18)**
SPAM	**1.21 (0.34)**	0.56 (0.13)	0.37 (0.08)
AGL	1.72 (1.30)	0.56 (0.26)	0.33 (0.12)
		$t = 2$	
n	100	225	400
MGS-AGL	1.63 (2.09)	0.41 (0.31)	0.20 (0.11)
MGS-BIC	1.82 (5.08)	**0.28 (0.34)**	**0.13 (0.08)**
SPAM	**1.24 (0.35)**	0.57 (0.15)	0.35 (0.08)
AGL	1.78 (2.55)	0.56 (0.32)	0.29 (0.10)

Table B.3: The average ISEs with $p = 50$ for k_j $(j = 1, 2, 3, 4)$

		$t = 1$	
n	100	225	400
MGS-AGL	1.51 (1.56)	0.58 (0.35)	0.33 (0.21)
SPAM	**1.45 (0.38)**	0.63 (0.15)	0.40 (0.07)
AGL	1.79 (1.21)	**0.54 (0.26)**	**0.32 (0.12)**
		$t = 2$	
n	100	225	400
MGS-AGL	1.96 (3.93)	**0.42 (0.28)**	**0.20 (0.12)**
SPAM	**1.42 (0.37)**	0.67 (0.16)	0.41 (0.08)
AGL	1.98 (2.18)	0.56 (0.29)	0.29 (0.11)

Table B.4: The average ISEs with $p = 100$ for k_j $(j = 1, 2, 3, 4)$

		$t = 1$	
n	100	225	400
MGS-AGL	1.70 (2.51)	0.59 (0.42)	**0.32 (0.21)**
SPAM	**1.61 (0.41)**	0.70 (0.16)	0.43 (0.08)
AGL	1.99 (1.41)	**0.57 (0.30)**	0.32 (0.11)
		$t = 2$	
n	100	225	400
MGS-AGL	2.00 (4.15)	**0.43 (0.31)**	**0.20 (0.11)**
SPAM	**1.55 (0.37)**	0.73 (0.16)	0.45 (0.09)
AGL	1.98 (1.32)	0.57 (0.28)	0.28 (0.10)

Table B.5: The average ISEs with $p = 200$ for k_j $(j = 1, 2, 3, 4)$

		$t = 1$	
n	100	225	400
MGS-AGL	1.97 (4.67)	0.58 (0.50)	**0.30 (0.20)**
SPAM	**1.77 (0.44)**	0.75 (0.16)	0.46 (0.08)
AGL	2.31 (1.64)	**0.55 (0.27)**	0.32 (0.12)
		$t = 2$	
n	100	225	400
MGS-AGL	2.52 (7.95)	**0.44 (0.29)**	**0.19 (0.13)**
SPAM	**1.67 (0.40)**	0.81 (0.18)	0.48 (0.09)
AGL	2.04 (1.17)	0.59 (0.31)	0.28 (0.11)

Table B.6: The average of false negative and positive selections with $p = 10$ for k_j $(j = 1, 2, 3, 4)$

FN		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.49	0.12	0.01	0.77	0.05	0
MGS-BIC	0.17	0	0	0.61	0.01	0
SPAM	0	0	0	0.06	0	0
AGL	0.30	0.02	0	1.00	0.13	0
FP		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.04	0	0	0.18	0.04	0.01
MGS-BIC	0.31	0.15	0.10	0.16	0.01	0
SPAM	4.01	4.59	4.88	3.61	4.86	5.28
AGL	0.12	0.01	0	0.16	0.05	0

Table B.7: The average of false negative and positive selections with $p = 20$ for k_j $(j = 1, 2, 3, 4)$

FN		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.50	0.15	0.02	0.98	0.07	0
MGS-BIC	0.19	0	0	0.64	0.02	0
SPAM	0	0	0	0.10	0	0
AGL	0.37	0.02	0	1.19	0.18	0
FP		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.11	0.01	0	0.33	0.10	0
MGS-BIC	0.82	0.41	0.24	0.39	0.02	0
SPAM	7.81	9.21	9.73	6.77	9.77	11.19
AGL	0.21	0.03	0.01	0.30	0.13	0.02

Table B.8: The average of false negative and positive selections with $p = 50$ for k_j $(j = 1, 2, 3, 4)$

FN		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.51	0.12	0.01	1.22	0.10	0
SPAM	0.01	0	0	0.25	0	0
AGL	0.59	0.01	0	1.46	0.23	0.02
FP		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.29	0.03	0	0.64	0.26	0.06
SPAM	13.47	16.35	17.43	11.52	16.95	20.56
AGL	0.44	0.12	0.04	0.58	0.31	0.10

Table B.9: The average of false negative and positive selections with $p = 100$ for k_j $(j = 1, 2, 3, 4)$

FN		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.56	0.14	0.01	1.42	0.11	0
SPAM	0.05	0	0	0.47	0	0
AGL	0.73	0.02	0	1.62	0.25	0.01
FP		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.59	0.09	0.01	0.97	0.46	0.16
SPAM	18.20	22.78	23.68	15.18	22.13	27.22
AGL	0.55	0.20	0.07	0.76	0.56	0.21

Table B.10: The average of false negative and positive selections with $p = 200$ for k_j $(j = 1, 2, 3, 4)$

FN		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	0.70	0.12	0.02	1.65	0.17	0
SPAM	0.11	0	0	0.66	0.02	0
AGL	1.04	0.01	0	1.81	0.35	0.02
FP		$t = 1$			$t = 2$	
n	100	225	400	100	225	400
MGS-AGL	1.00	0.19	0.04	1.13	0.87	0.29
SPAM	21.85	28.14	31.88	18.62	27.45	34.84
AGL	0.71	0.27	0.17	0.99	0.81	0.37

C

Supplementary materials for Chapter 3

C.1 Details of fitting the log-normal hierarchical model

We fit the hierarchical regression model by sampling from its posterior distributions using Markov chain Monte Carlo. We introduce three different algorithms: the Gibbs sampling algorithm which updates parameters one at a time sequentially, the block Gibbs sampling algorithm which jointly updates vectors of correlated parameters, and the Hamiltonian Monte Carlo algorithm which uses the Hamiltonian dynamics to propose efficient moves around the parameter space.

1. **Gibbs Sampling Algorithm**.

   The Gibbs sampler iterates the following steps until convergence.

   (a) For $i = 1, \ldots, N$, sample B_i from

   $$\mathcal{N}\left\{ \frac{b_i/\tau_i^2 + \sum_{j \in J_i} (y'_{ij} - G_j)/\sigma_i^2}{1/\tau_i^2 + \sum_{j \in J_i} 1/\sigma_i^2}, \frac{1}{1/\tau_i^2 + \sum_{j \in J_i} 1/\sigma_i^2} \right\}.$$

   (b) For $j = 1, \ldots, M$, sample G_j from

   $$\mathcal{N}\left\{ \frac{\sum_{i \in I_j} (y'_{ij} - B_i)/\sigma_i^2}{\sum_{i \in I_j} 1/\sigma_i^2}, \frac{1}{\sum_{i \in I_j} 1/\sigma_i^2} \right\}.$$

   (c) Update each σ_i^2 one-at-a-time using Metropolis-Hastings.

2. **Block Gibbs Sampling Algorithm**.

   The block Gibbs sampler iterates the following step and step (1c) until convergence.

   (a) Sample the vector $(B^t, G^t)^t$ using Section 3.2.3, i.e., sample $(B^t, G^t)^t$ from a multivariate Gaussian distribution with mean $\Omega^{-1}\gamma$ and variance Ω^{-1}.

3. **Hamiltonian Monte Carlo Algorithm**.

   In the Hamiltonian Monte Carlo algorithm, we sample the whole vector of unknown parameters, i.e., $\{B_i, G_j, \sigma_i^2\}$ through the non-U-turn HMC sampler[23]. The algorithm is implemented with the STAN package.

We compare the performance of the afore mentioned algorithms using auto-correlation plots of the posterior samples and the effective sample size, in both the simulated and real data examples. We find that the Gibbs sampler converges very slowly relative to the other two algorithms. We can cross check our results by comparing the samples obtained with the block Gibbs sampler and HMC – they give the same posterior distributions.

C.2 Proprieties of the posterior distribution

Theorem C.1. *Under the prior specifications for $\{B_i, G_j, \sigma_i^2 : i = 1, \dots, N,\ j = 1, \dots, M\}$ given in (3.6), we have (i) the posterior is proper if all instruments measure all sources, i.e. $|J_i| = M$ for all $1 \leq i \leq N$, (ii) the MAP estimator of each σ_i^2 is bounded away from zero by a finite constant which only depends on the hyper-parameters. Furthermore, flat priors on the σ_i^2 would result in an unbounded posterior distribution.*

Proof. **Part 1.** Under the prior specifications in (3.6), the joint posterior distribution is

$$p(B, G, \sigma^2 \mid D, \tau^2) \propto \prod_{i=1}^{N} \sigma_i^{-M-2-2df_g} \exp\left\{ -\sum_{i=1}^{N} \frac{\sum_{j=1}^{M}(y'_{ij} - B_i - G_j)^2 + 2\beta_g}{2\sigma_i^2} - \sum_{i=1}^{N} \frac{(b_i - B_i)^2}{2\tau_i^2} \right\}.$$

Integrating out (B, G) gives

$$p(\sigma^2|D, \tau^2) \propto \prod_{i=1}^{N} \sigma_i^{-M-2-2df_g} \mid \det(\Omega)|^{-1/2} \exp\left\{ \frac{1}{2}\mu^t\Omega\mu - \sum_{i=1}^{N} \left\{ \beta_g + \sum_{j\in J_i} \frac{(y'_{ij})^2}{2} \right\} \sigma_i^{-2} \right\},$$

where μ and Ω, both of which depends on the σ_i^2, are defined in Section 3.2.3.

Claim 1: $\mu^t\Omega\mu < \sum_{i=1}^{N}\sum_{j=1}^{M}(y'_{ij})^2\sigma_i^{-2}$.

Claim 2: $|\det(\Omega)|^{-1/2} \leq D\prod_{i=1}^{N}\sigma_i$ for some constant D.

From Claims 1 and 2, whose proofs are given after the current proof, we conclude that $p(\sigma^2 \mid D, \tau)$ is integrable on the positive real line when all $|J_i| = M$, thus the posterior is proper.

Part 2. For fixed B and G, the σ_i^2 which maximizes the posterior probability satisfies

$$\sigma_i^2 \;=\; 2\sqrt{u^2 + 2\beta_g/M + v_i} - 2u \; (i = 1, \dots, N),$$

where $u = 1 + (2df_g + 2)/M, v_i = \sum_{j=1}^{M}(y_{ij} - B_i - G_j)^2/M$. Then,

$$\sigma_i^2 \geq \frac{2\beta_g/M}{\sqrt{u^2 + 2\beta_g/M} + u},$$

thus the MAP estimator of σ_i^2 is bounded away from 0 by a finite constant independent of B and G.

Part 3. If we assign flat priors on σ_i^2, the posterior distribution may be unbounded near the boundary. For example, if $J_i \neq \emptyset$, let $B_i = 0$ and $G_j = y_{ij}, j \in J_i$, then $\sum_{j \in J_i} \frac{(y'_{ij} - B_i - G_j)^2}{2\sigma_i^2} = \sigma_i^2/8$. Then $p(B, G, \sigma^2 \mid D, \tau^2) \to \infty$ as $\sigma_i \to 0$.

□

First, let us study the properties of the Ω matrix. We use $p_i = \sigma_i^{-2}$ for simplicity of notations and assume $b_i = 0$ without loss of generality. Let A be an $(N + M) \times (N + M)$ diagonal matrix and the diagonal values are the same as Ω. Let U be an $(N{+}M){\times}2$ matrix such that

$$U_{i,1} = p_i,\ U_{i,2} = 0\ (i = 1, \ldots, N), \quad U_{j+N,1} = 0,\ U_{j+N,2} = 1\ (j = 1, \ldots, M).$$

Let C be a 2×2 matrix such that $C_{i,j} = I_{i \neq j}\ (i,j = 1, 2)$. Then $\Omega = A + UCU^t$. By the Woodbury matrix identity, we have

$$\Omega^{-1} = A^{-1} - A^{-1}U\left(C + U^tA^{-1}U\right)^{-1}U^tA^{-1}. \tag{C.1}$$

For simplicity, let $\alpha_i = Mp_i + \tau^{-2}$, $\beta = \sum_{i=1}^{N} p_i$ and $\omega_i = \tau_i^{-2}\alpha_i^{-1}$. Then $Mp_i\alpha_i^{-1} = 1 - \omega_i$.

Let

$$\begin{aligned} \delta &= \frac{1}{1 - M(\sum_{i=1}^{N} \alpha_i^{-1} p_i^2)\beta^{-1}} \\ &= \frac{\sum_{i=1}^{N} p_i}{\sum_{i=1}^{N} \omega_i p_i}. \end{aligned}$$

Then we have

$$\begin{aligned} (A)^{-1}_{i,i} &= \alpha_i^{-1}\ (i = 1, \ldots, N), \quad (A)^{-1}_{j+N,j+N} = \beta^{-1}\ (j = 1, \ldots, M); \\ \left(A^{-1}U\right)_{i,1} &= (1-\omega_i)/M\ (i = 1, \ldots, N), \quad \left(A^{-1}U\right)_{j+N,2} = \beta^{-1}\ (j = 1, \ldots, M); \\ \left(C + U^t A^{-1} U\right)^{-1}_{1,1} &= -\delta M \beta^{-1}, \quad \left(C + U^t A^{-1} U\right)^{-1}_{1,2} = \delta \\ \left(C + U^t A^{-1} U\right)^{-1}_{2,1} &= \delta, \left(C + U^t A^{-1} U\right)^{-1}_{2,2} = -\delta(\sum_{i=1}^{N}(1-\omega_i)p_i/M). \end{aligned}$$

<u>Proof of Claim 1</u>. First, we define some new variables:

$$\bar{y}_{i\cdot} = M^{-1}\sum_{j=1}^{M} y'_{ij} p_i, \quad \bar{y}_{\cdot j} = N^{-1}\sum_{i=1}^{N} y'_{ij} p_i, \quad \bar{y}_{\cdot\cdot} = (MN)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} y'_{ij} p_i.$$

then $\gamma_i = M\bar{y}_{i\cdot}$ $(i = 1, \ldots, N)$ and $\gamma_{j+N} = N\bar{y}_{\cdot j}$ $(j = 1, \ldots, M)$. So

$$\begin{aligned} \left(U^t A^{-1} \gamma\right)_1 &= \sum_{i=1}^{N}(1-\omega_i)\bar{y}_{i\cdot} = N\bar{y}_{\cdot\cdot} - \sum_{i=1}^{N}\omega_i \bar{y}_{i\cdot} \\ \left(U^t A^{-1} \gamma\right)_2 &= N\beta^{-1}\sum_{j=1}^{M}\bar{y}_{\cdot j} = \beta^{-1} MN\bar{y}_{\cdot\cdot} \end{aligned}$$

which will give us

$$\begin{aligned}
\mu^t\Omega\mu = \gamma^t\Omega^{-1}\gamma &= \gamma^t A^{-1}\gamma - \gamma^t A^{-1} U \left(C + U^t A^{-1} U\right)^{-1} U^t A^{-1}\gamma \\
&= M^2 \sum_{i=1}^{N} \overline{y}_{i\cdot}^2 \alpha_i^{-1} + N^2\beta^{-1} \sum_{j=1}^{M} \overline{y}_{\cdot j}^2 + \delta M \beta^{-1} \left(N\overline{y}_{\cdot\cdot} - \sum_{i=1}^{N} \omega_i \overline{y}_{i\cdot} \right)^2 \\
&\quad + MN^2\beta^{-2}\delta\overline{y}_{\cdot\cdot}^2 \sum_{i=1}^{N} (1-\omega_i) p_i - 2\delta \left(N\overline{y}_{\cdot\cdot} - \sum_{i=1}^{N} \omega_i \overline{y}_{i\cdot} \right) \beta^{-1} MN\overline{y}_{\cdot\cdot} \\
&= M^2 \sum_{i=1}^{N} \overline{y}_{i\cdot}^2 \alpha_i^{-1} + N^2\beta^{-1} \sum_{j=1}^{M} \overline{y}_{\cdot j}^2 + \delta M \beta^{-1} \left(\sum_{i=1}^{N} \omega_i \overline{y}_{i\cdot} \right)^2 - MN^2\beta^{-1}\overline{y}_{\cdot\cdot}^2 \\
&= M \sum_{i=1}^{N} p_i^{-1} \overline{y}_{i\cdot}^2 + N^2\beta^{-1} \sum_{j=1}^{M} \overline{y}_{\cdot j}^2 - MN^2\beta^{-1}\overline{y}_{\cdot\cdot}^2 \\
&\quad - \left\{ M \sum_{i=1}^{N} \omega_i p_i^{-1} \overline{y}_{\cdot\cdot}^2 - M \left(\sum_{i=1}^{N} \omega_i p_i \right)^{-1} \left(\sum_{i=1}^{N} \omega_i \overline{y}_{i\cdot} \right)^2 \right\}
\end{aligned}$$

Applying the the Cauchy-Schwarz inequality gives

$$\left(\sum_{i=1}^{N} \omega_i p_i^{-1} \overline{y}_{\cdot\cdot}^2 \right) \left(\sum_{i=1}^{N} \omega_i p_i \right) \geq \left(\sum_{i=1}^{N} \omega_i \overline{y}_{i\cdot} \right)^2 . \tag{C.2}$$

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{j=1}^{M} (y'_{ij})^2 p_i - M \sum_{i=1}^{N} p_i^{-1} \overline{y}_{i\cdot}^2 && \text{(C.3)} \\
=\ & \sum_{j=1}^{M} \sum_{i=1}^{N} p_i^{-1} \left(y'_{ij} p_i - \overline{y}_{i\cdot} \right)^2 \\
\geq\ & \sum_{j=1}^{M} \left(\sum_{i=1}^{M} p_i \right)^{-1} \left(\sum_{i=1}^{N} y'_{ij} p_i - \overline{y}_{i\cdot} \right)^2 \\
=\ & \beta^{-1} N^2 \sum_{j=1}^{M} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \right)^2 = \beta^{-1} N^2 \left(\sum_{j=1}^{M} \overline{y}_{\cdot j}^2 - M\overline{y}_{\cdot\cdot}^2 \right)
\end{aligned}$$

Combing (C.2) and (C.3), we can get $\mu^t\Omega\mu \leq \sum_{i=1}^{N} \sum_{j=1}^{M} (y'_{ij})^2 p_i$.

Proof of Claim 2. From the matrix determinant lemma,

$$\begin{aligned}
\det(\Omega) &= \det\left(C^{-1} + U^t A^{-1} U\right) \det(A) \det(C) \\
&= \delta^{-1} \prod_{i=1}^{M} \alpha_i \beta^M \\
&= \left\{1 - \sum_{i=1}^{N} p_i(1-\omega_i)\beta^{-1}\right\} \prod_{i=1}^{M} \alpha_i \beta^M \\
&= \left\{\sum_{i=1}^{N} p_i \omega_i\right\} \prod_{i=1}^{M} \alpha_i \beta^{M-1} \\
&\geq \beta^{M-1} \sum_{i=1}^{N} p_i \tau_i^{-2} \alpha_i^{-1} \prod_{l=1}^{N} \alpha_l \\
&\geq \beta^{M-1} \sum_{i=1}^{N} \sigma_i^{-2} \prod_{l=1}^{N} \tau_l^{-2} \\
&\geq \prod_{i=1}^{N} \tau_i^{-2} \prod_{i=1}^{N} \sigma_i^{-2}.
\end{aligned}$$

C.3 Frequentist method

C.3.1 MLEs and asymptotic variances

The MLEs of the random-effect regression model given by (3.3) and (3.4) can be obtained by setting the derivative of the log-likelihood equal to zero. Let $\widehat{B}_{1:N}$ and $\widehat{G}_{1:M}$ be the MLEs of B and G.

Theorem C.2. *The MLEs of the B_i and the G_j can be written as*

$$\begin{pmatrix} \widehat{B}_{1:N} \\ \widehat{G}_{1:M} \end{pmatrix} = \begin{pmatrix} B \\ G \end{pmatrix} + \Omega^{-1} \begin{pmatrix} R_B \\ R_G \end{pmatrix}; \tag{C.4}$$

where

$$\begin{aligned}(R_B)_i &= M^{-1}\sum_{j=1}^{M}(y'_{ij} - B_i - G_j)\sigma_i^{-2} + M^{-1}(b_i - B_i)\tau_i^{-2},\\ (R_G)_j &= N^{-1}\sum_{i=1}^{N}\sigma_i^{-2}(y'_{ij} - B_i - G_j).\end{aligned}$$

Assume that (i) the σ_i^2 and the τ_i^2 are uniformly bounded from below and from above by finite positive constants and (ii) $N^{-1}\sum_{i=1}^{N}\sigma_i^{-2}$ converges to a finite positive constant as $N \to \infty$. Then as M and N goes to infinity, $\widehat{B}_{1:N}$ and $\widehat{G}_{1:M}$ converge to the corresponding true values almost surely and the asymptotic variances are $O(M^{-1})$ and $O(N^{-1})$ respectively.

Proof. The almost sure convergence of the MLEs follows from the strong law of large numbers. The rate of the asymptotic variances follows by letting $M, N \to \infty$ in the variance-covariance matrix of the MLEs, which is the inverse of the Fisher information matrix under this Gaussian model. □

Here we give the closed-form solutions of the variances of the MLEs when ψ is known.

Proposition C.1. *When all detectors measure all objects, i.e. $J_i = \{1, \ldots, M\}$, $I_j = \{1, \ldots, N\}$ and $\{\sigma_i^2, \tau_i^2\}$ are known constants; the variances of $\{\widehat{B}_i\}_{i=1}^{N}$, $\{\widehat{G}_j\}_{j=1}^{M}$ are given by*

$$\begin{aligned}\mathrm{var}(\hat{B}_j) &= \left(M\sigma_i^{-2} + \tau_i^{-2}\right)^{-1}\left\{1 + \frac{(1-\omega_i)\sigma_i^{-2}}{\sum_{i=1}^{N}\omega_i\sigma_i^{-2}}\right\},\\ \mathrm{var}(\hat{G}_i) &= \left(\sum_{i=1}^{N}\sigma^{-2}\right)^{-1}\left\{1 + \frac{\sum_{i=1}^{N}(1-\omega_i)\sigma_i^{-2}}{M\sum_{i=1}^{N}\omega_i\sigma^{-2}}\right\}.\end{aligned}$$

Moreover, we have

$$Cov(B_i, G_j) = -\frac{1-\omega_i}{M\sum_{i=1}^{N}\omega_i\sigma_i^{-2}}.$$

Under the additive model, B_i and G_j are negatively correlated for all i, j. The variances are just a direct result of (C.1).

C.3.2 Goodness-of-fit

We now give a goodness-of-fit test statistics for the random-effect regression model. Since the errors e_{ij} are independent normal distributions with mean 0 and variances σ_i^2, and b_i also follows normal distributions with variance τ_i^2, we define the following normalized residual sum of squares:

$$T(B, G) := \sum_{i=1}^{N} \frac{(b_i - B_i)^2}{\tau_i^2} + \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\left(y'_{ij} - B_i - G_j\right)^2}{\sigma_i^2}. \tag{C.5}$$

Theorem C.3. *When the variances σ_i^2, τ_i^2 are known and we plug in the MLEs of B_i and the G_j for the random-effect regression model in Equation (C.5), then statistic T follows a Chi-squared distribution with degree of freedom $MN - M$, i.e.*

$$T(\widehat{B}_{1:N}, \widehat{G}_{1:M}) \sim \chi^2_{NM-M}. \tag{C.6}$$

Proof. We can write our model as a weighted linear regression model, with $(NM + N)$ independent Gaussian noise: $\{e_{ij}, \varepsilon_i : i = 1, \ldots, N, j = 1, \ldots, M\}$. Plugging in the estimators of the B_i and the G_j, which are linear in the observed values $\{y_{ij},\ b_i\}$, costs $N + M$ degrees of freedom. Therefore, the degrees of freedom left is $NM - M$. □

With unknown variances we do not have a closed-form distribution of T as defined in Equation (C.5). Instead, we use the following approximation: plug in the estimated variances and adjust the degrees of freedom as $(MN - M - N)$ to take the estimations the variances σ_i^2 into account. The resulting p-values of the fitted data in Sections 3.3.2.2 and 3.3.2.3 are not significant.

C.4 More simulations

C.4.1 Simulations of correctly specified model

We consider two simulation studies that vary in terms of the relative number of instruments and sources. The parameters are specified the same as Simulation I except that in Simulation V, $N = M = 10$ and in Simulation VI, $N = 10$ and $M = 100$. In Simulation VI, non-surprisingly, the estimates of the effective areas are more precise, whereas the estimates of the fluxes are less precise, as compared with Simulation V. Figure C.1 contrasts the results of Simulation V and VI by comparing the posterior distributions of B_1, B_2, B_3, and B_4 in four columns for Simulation V (row 1) and Simulation VI (row 2). Similarly, the third and fourth rows compare the posterior distributions of G_1, G_2, G_3, and G_4 for Simulation V (row 3) and Simulation VI (row 4).

C.4.2 Simulations of misspecified model

C.4.2.1 Noisy known constants

In Simulation VII, we generates data under

$$y_{ij} = -\sigma_i^2/2 + B_i + G_j + \lambda_{ij} + e_{ij}, \tag{C.7}$$

with $e_{ij} \sim N(0, \sigma_i^2)$ and $\lambda_{ij} \sim N(0, \zeta^2)$. The model in Section 3.2 assumes that $\zeta = 0$, or equivalently that each λ_{ij} is zero. We use model (C.7) to mimic a realistic case where the multiplicative model is not perfectly satisfied. This is equivalent to (3.3) when each σ_i^2 is replaced by $\sigma_i^2 + \zeta^2$, since $e_{ij} + \lambda_{ij} \sim N(0, \sigma_i^2 + \zeta^2)$. This is confirmed numerically in Simulation VII, which has the same setup as Simulation V except that $\sigma_i = \zeta = 0.1$. Figure C.2 compares the posterior distributions of $\{B_i\}_{i=1}^5$ and $\{G_j\}_{j=1}^5$ for model fittings with known $\sigma_i^2 = 0.1^2$ and unknown σ_i^2.

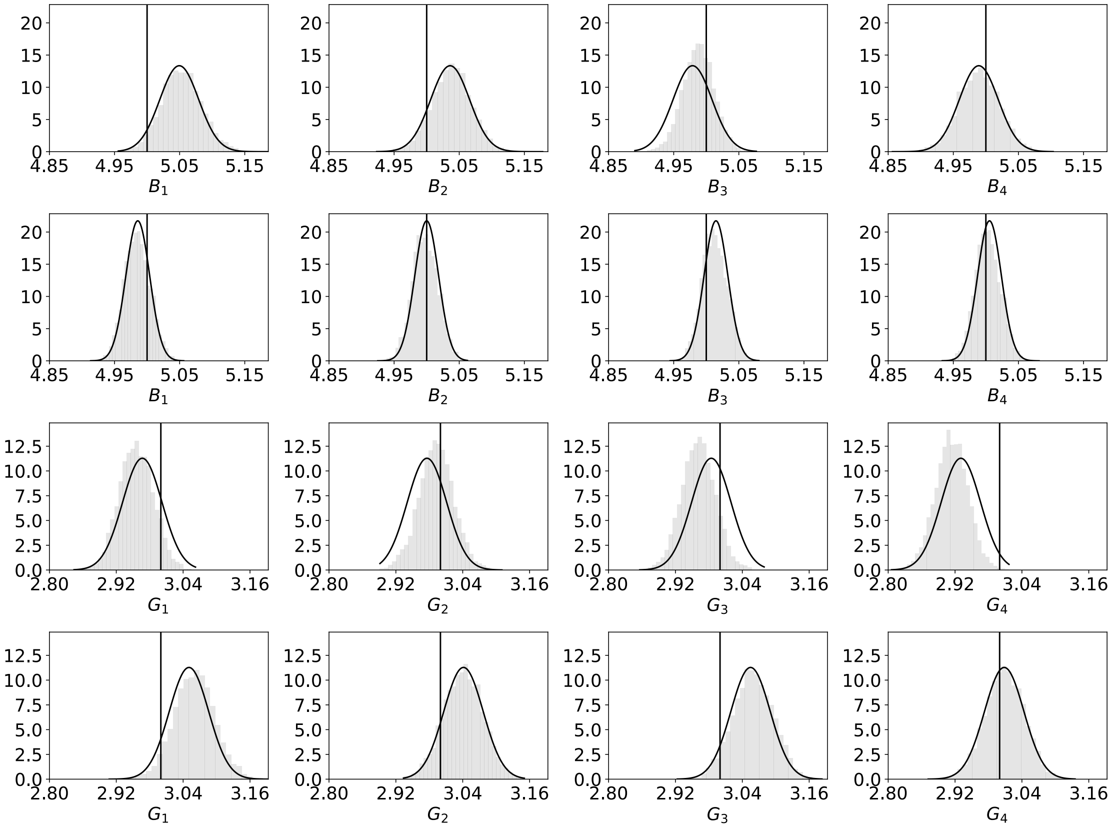

Figure C.1: Simulations V (rows 1 & 3) and VI (rows 2 & 4). The gray histograms are the posterior distributions of the B_i (rows 1 & 2), and the G_j (rows 3 & 4), when the σ_i^2 are unknown. The black vertical lines are the true values. The solid density curves on top of the histograms denote the closed-form posterior densities when the σ_i^2 are known.

Based on Simulation VII, if the true model is the log-normal regression model but there exists uncertainties on the multiplicative constant, the estimated variances are inflated – to account for the extra variability brought in by the uncertainties on the multiplicative constant. As can be seen in Fig. C.2, the estimated σ_i is approximately $\sqrt{\zeta^2 + \sigma_i^2} = \sqrt{0.01 + 0.01} \approx 0.14$ – which means the extra uncertainty comes out as an additive error which is not distinguishable from the measurement error. Again, if the practitioner plugs-in known values for the σ_i^2 which miss other possible uncertainties (represented by the λ_{ij} here), the results could be overly optimistic or even misleading in terms of the suggested adjustments of the effective areas.

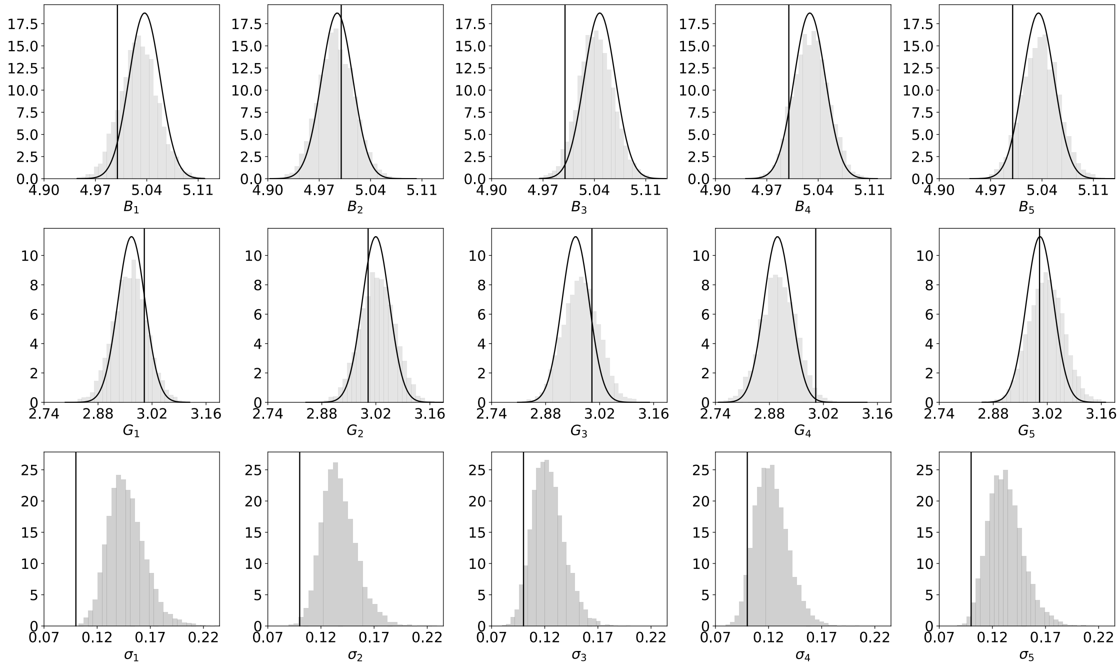

Figure C.2: Simulation VII. The legend is the same as in Fig. C.1.

C.4.2.2 Misspecification in both ways

Next, we combine the model misspecification discussed in Sections 3.3.1.2 and C.4.2.1. Specifically, the data generating model for Simulations VIII and IX was

$$c_{ij} \sim \text{Poisson}(\lambda_{ij} A_i F_j), \tag{C.8}$$

where the λ_{ij} were randomly generated from the uniform distribution on $[0.8, 1.2]$. This resembles the case where the true model is Poisson and the estimation of T_{ij} is volatile. The other parameters are set to be the same as in Simulations IV and V. Figures C.3 and C.4 give the results of Simulation VIII with smaller counts ($B_i = 1$ and $G_j = 3$) and IX with larger counts ($B_i = 5$ and $G_j = 3$) under this scenario. It shows with large Poisson counts, controlling the uncertainty in the multiplicative constant can possibly lead to reasonably good results. Thus even with compounded model misspecification, the log-normal hierarchical model is able to provide reasonable, though not as precise, results, as compared with the correctly-specified case.

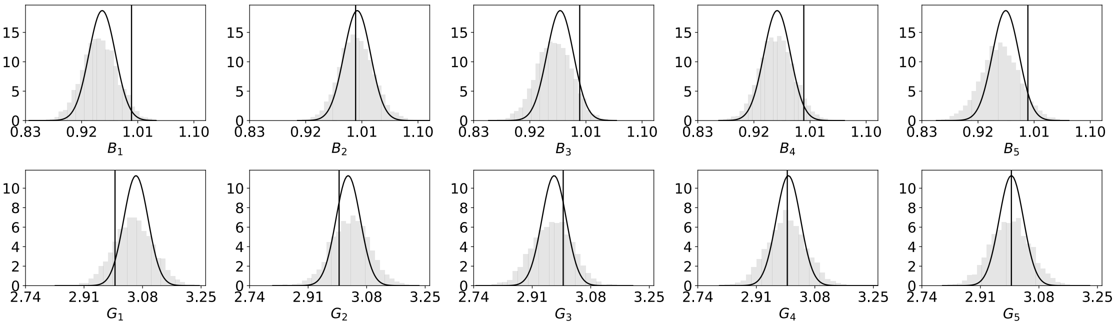

Figure C.3: Simulations VIII. The legend is the same as in Fig. C.1.

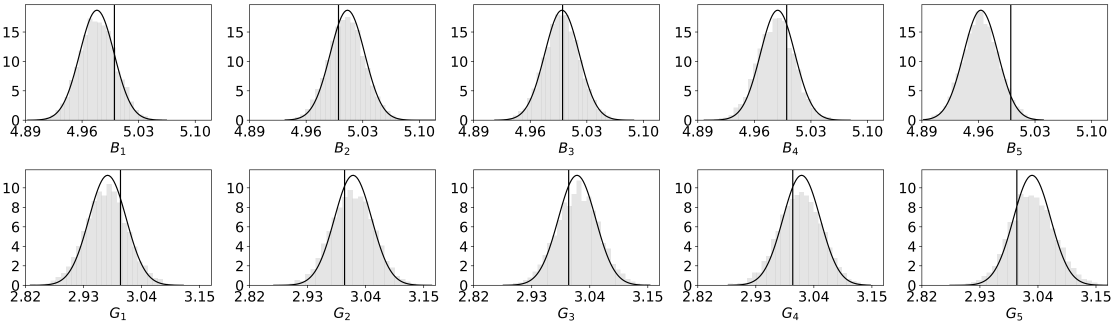

Figure C.4: Simulations IX. The legend is the same as in Fig. C.1.

C.5 Results for fractions of prior information

We display the proportion of prior information for the real data sets (Section 3.3) in Tables C.1 and C.2.

Table C.1: Proportion of prior influence for E0102 data in Section 3.3.2.1.

Instrument	O		Ne	
	$\tau = 0.025$	$\tau = 0.05$	$\tau = 0.025$	$\tau = 0.05$
RGS1	0.570	0.205	0.063	0.016
MOS1	0.279	0.077	0.075	0.019
MOS2	0.355	0.065	0.077	0.017
pn	0.250	0.041	0.620	0.218
ACIS-S3	0.218	0.040	0.270	0.088
ACIS-I3	0.906	0.640	0.099	0.026
HETG	0.648	0.341	0.129	0.034
XIS0	0.180	0.051	0.069	0.018
XIS1	0.298	0.078	0.071	0.019
XIS2	0.463	0.140	0.063	0.016
XIS3	0.772	0.364	0.062	0.018
XRT-WT	0.726	0.278	0.154	0.026
XRT-PC	0.934	0.235	0.906	0.017

Table C.2: Proportion of prior influence for data in Section 3.3.2.2 and 3.3.2.3.

Data Name	$\tau_i = 0.025$			$\tau_i = 0.05$		
	pn	mos1	mos2	pn	mos1	mos2
hard band 2XMM	0.093	0.075	0.082	0.025	0.020	0.022
medium band 2XMM	0.250	0.216	0.222	0.076	0.065	0.067
soft band 2XMM	0.093	0.075	0.069	0.025	0.020	0.018
hard band	0.010	0.019	0.031	0.003	0.005	0.008
medium band	0.023	0.016	0.028	0.006	0.004	0.007
soft band	0.021	0.011	0.007	0.005	0.003	0.002

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.

[2] Blyth, S. (1994). Local divergence and association. *Biometrika*, (pp. 579–584).

[3] Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391), 580–598.

[4] Chandra X ray observatory (2009). E0102-72.3: adding a new dimension to an old explosion. http://chandra.harvard.edu/photo/2009/e0102/.

[5] De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., & De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

[6] Dierckx, P. (1995). *Curve and surface fitting with splines*. Oxford University Press.

[7] Doksum, K., Blyth, S., Bradlow, E., Meng, X.-L., & Zhao, H. (1994). Correlation curves as local measures of variance explained by regression. *Journal of the American Statistical Association*, 89(426), 571–582.

[8] Efron, B. & Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311–319.

[9] Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, (pp. 89–102).

[10] Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.

[11] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, (pp. 1–67).

[12] Friedman, J. H. & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1), 3–21.

[13] Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, (pp. 733–760).

[14] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

[15] Genest, C. & Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2), 335–369.

[16] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar), 723–773.

[17] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* (pp. 63–77).: Springer.

[18] Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1), 81–102.

[19] Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

[20] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.

[21] Heller, R., Heller, Y., Kaufman, S., Brill, B., & Gorfine, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29), 1–54.

[22] Hoeffding, W. (1948). A non-parametric test of independence. *The annals of mathematical statistics*, (pp. 546–557).

[23] Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.

[24] Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, 38(4), 2282.

[25] Hušková, M. & Meintanis, S. G. (2008). Testing procedures based on the empirical characteristic functions i: Goodness-of-fit, testing for symmetry and independence. *Tatra Mt. Math. Publ*, 39, 225–233.

[26] IACHEC (2017). International Astronomical Consortium for High Energy Calibration. http://web.mit.edu/iachec/.

[27] Kankainen, A. & Ushakov, N. G. (1998). A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciences*, 89(5), 1486–1494.

[28] Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430), 773–795.

[29] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6), 066138.

[30] Lin, Y., Zhang, H. H., et al. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5), 2272–2297.

[31] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag New York, Inc.

[32] Marshall, H., Kashyap, V., Chen, Y., Meng, X.-L., & Wang, X. (2017). Calibration concordance. *In Preparation*.

[33] Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, (pp. 1142–1160).

[34] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), 1087–1092.

[35] Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381), 47–55.

[36] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1), 141–142.

[37] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo*. CRC press.

[38] Plucinsky, P. P., Beardmore, A. P., Foster, A., Haberl, F., Miller, E. D., Pollock, A. M. T., & Sembay, S. (2017). SNR 1E 0102.2-7219 as an X-ray calibration standard in the 0.5-1.0 keV bandpass and its application to the CCD instruments aboard Chandra, Suzaku, Swift and XMM-Newton. *Astronomy and Astrophysics*, 597, A35.

[39] Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009–1030.

[40] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, 334(6062), 1518–1524.

[41] Reshef, D. N., Reshef, Y. A., Sabeti, P. C., & Mitzenmacher, M. M. (2015a). An empirical study of leading measures of dependence. *arXiv preprint arXiv:1505.02214*.

[42] Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., & Mitzenmacher, M. (2016). Measuring dependence powerfully and equitably. *Journal of Machine Learning Research*, 17(212), 1–63.

[43] Reshef, Y. A., Reshef, D. N., Sabeti, P. C., & Mitzenmacher, M. M. (2015b). Equitability, interval estimation, and statistical power. *arXiv preprint arXiv:1505.02212*.

[44] Ritzema, H. (2006). *Drainage principles and applications*. Number 16. ILRI.

[45] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, (pp. 34–58).

[46] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464.

[47] Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), 2263–2291.

[48] Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, (pp. 898–916).

[49] Stan Development Team (2015). *Stan Modeling Language User's Guide and Reference Manual, Version 2.10.0*.

[50] Stan Development Team (2016). *PyStan: the Python interface to Stan*.

[51] Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, (pp. 689–705).

[52] Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, (pp. 590–606).

[53] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794.

[54] Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *The annals of applied statistics*, 3(4), 1236–1265.

[55] Wand, M. & Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2), 179–198.

[56] Wang, X., Jiang, B., & Liu, J. (2017). Generalized R-squared for detecting dependence. *Biometrika*, 104(1), 129–139.

[57] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, (pp. 359–372).

[58] Watson, M. G., Schröder, A. C., Fyfe, D., Page, C. G., Lamer, G., Mateos, S., Pye, J., Sakano, M., Rosen, S., Ballet, J., Barcons, X., Barret, D., Boller, T., Brunner, H., Brusa, M., Caccianiga, A., Carrera, F. J., Ceballos, M., Della Ceca, R., Denby, M., Denkinson, G., Dupuy, S., Farrell, S., Fraschetti, F., Freyberg, M. J., Guillout, P., Hambaryan, V., Maccacaro, T., Mathiesen, B., McMahon, R., Michel, L., Motch, C., Osborne, J. P., Page, M., Pakull, M. W., Pietsch, W., Saxton, R., Schwope, A., Severgnini, P., Simpson, M., Sironi, G., Stewart, G., Stewart, I. M., Stobbart, A.-M., Tedds, J., Warwick, R., Webb, N., West, R., Worrall, D., & Yuan, W. (2009). The XMM-Newton serendipitous survey. V. The Second XMM-Newton serendipitous source catalogue. *Astronomy and Astrophysics*, 493, 339–373.

[59] XMM Catalogue public pages (2008). XMM-Newton serendipitous source catalogue: 2XMM. http://xmmssc-www.star.le.ac.uk/Catalogue/xcat_public_2XMM.html.

[60] Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

[61] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.