On the two-dataset problem

Steven N. MacEachern
Department of Statistics, The Ohio State University

Koji Miyawaki
School of Economics, Kwansei Gakuin University

November 4, 2019

Abstract

This paper considers the two-dataset problem, where data are collected from two potentially different populations sharing common aspects. This problem arises when data are collected by two different types of researchers or from two different sources. We may reach invalid conclusions if knowledge about the data collection process is ignored or assumptions about the process are wrong. To address this problem, this paper develops statistical models and proposes two prediction errors that can be used to evaluate the underlying data collection process. As a consequence, it is possible to discuss the heterogeneity/similarity of data in terms of prediction. Two real datasets are selected to illustrate our method.

Keywords: Data collection process; Random coefficients model; Bayesian model averaging.
1 Introduction

The data collection process is an important component of statistical analysis (see, e.g., Chapter 1 of Cox and Snell (1981)), where data are often assumed to be independently obtained from an identical population with appropriate precision for the analysis. In the literature on causal inference, this assumption is required for so-called internal validity (see, e.g., Shadish et al. (2002)). However, in practice, we encounter cases where this assumption does not hold because of the data collection process. This paper aims to develop statistical models that reflect the underlying data collection process and propose two prediction errors that can be used to evaluate the process. These two errors are closely related to the final prediction error (see Akaike (1969, 1970)), which is a precedent of the well-known Akaike information criterion.

Two typical data collection processes, neither of which satisfies the above assumption, are considered in this paper. The first situation concerns data quality. Except in cases of small datasets, data are typically collected by several researchers and are combined into one large dataset for analysis. Thus, it is reasonable to consider the heterogeneity in the degree of data collection skill among researchers. Even if the skill is homogeneous, the data may differ in quality due to the data collection environment, as we will see in Figure 1. Both scenarios yield a dataset that is a mix of low-quality and high-quality data. In this case, internal consistency in quality is an issue. If the quality is not consistent, using only the precise data may lead to a valid conclusion, but the conclusion may be inaccurate because it ignores information from the discarded data.

The second situation is related to the data source. To ensure a sufficient number of observations, data are collected from multiple sources, which may violate the assumption that the data are from the same population. For
example, medical data are often collected at several hospitals, but different hospitals are likely to have different treatments or patients. Thus, the analysis of data collected in this way may be invalid because the data are from different populations. However, as in the diabetes dataset discussed in Section 2, the two populations also share some common aspects. In practice, data are obtained by a mix of the two situations above.

To address the data collection process explicitly, we link the process to the linear regression model and derive two prediction errors, which are of Mallows’ $C_p$-type (see Mallows (1973, 1995)). The first evaluates the internal quality of the data. Furthermore, we investigate the heterogeneity/similarity among observations with the second error, which is addressed within the framework of Bayesian model averaging (see, e.g., Raftery et al. (1997); Brown et al. (2002); Steel (2019); Miyawaki and MacEachern (2019)).

The rest of this paper is organized as follows. After two real datasets used in this paper are illustrated in Section 2 Section 3 describes a general two-dataset problem and presents two prediction errors. We explain our approach to estimating these errors in Section 4. The sensitivity of our estimation approach to the prior distribution is also discussed in this section. Section 5 presents illustrative applications of our method using two real datasets. Finally, Section 6 concludes the paper.

2 Two motivating datasets

The geyser data used by Azzalini and Bowman (1990) are shown in Figure 1. Each observation is the repeated measurement of eruptions of the Old Faithful Geyser in the United States, and consists of two variables: the duration of the eruption and the waiting time between two consecutive eruptions. Of
the 298 observations, 77 have duration times of 2, 3, or 4, as depicted by the x marks in Figure 1. These values are rounded because they are collected at night, as noted in Azzalini and Bowman (1990). This is an example of heterogeneity in data quality.

Another example is the diabetes data analyzed by Efron et al. (2004), including diabetes progression measure (Y) as the response and ten predictors (age, sex, body mass index (BMI), average blood pressure (BP), and six blood serum measurements (S1 - S6)). This dataset is used to illustrate variable selection methods (e.g., Efron et al. (2004), Hahn and Carvalho (2015), Miyawaki and MacEachern (2019) among others). A close look suggests that the data may come from at least two sources, because the precision of the blood pressure and S4 (the fourth blood serum measurement) is different.
from that of the other measurements (see ID 95 and 99 in Table 1).

Table 1: Ten observations selected from the diabetes data

| ID† | Age | Sex | BMI | BP  | S1  | S2  | S3  | S4  | S5  | S6  | Y   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |
| 91  | 36  | 1   | 21.9| 89  | 105.2| 68  | 3   | 4.3694| 96  | 111 |
| 92  | 52  | 1   | 24  | 83  | 167  | 71  | 2   | 3.8501| 94  | 98  |
| 93  | 61  | 1   | 31.2| 79  | 235  | 47  | 5   | 5.0499| 96  | 164 |
| 94  | 43  | 1   | 26.8| 123 | 193  | 67  | 3   | 4.7791| 94  | 48  |
| 95  | 35  | 1   | 20.4| 65  | 187  | 67  | 2.79| 4.2767| 78  | 96  |
| 96  | 27  | 1   | 24.8| 91  | 189  | 69  | 3   | 4.1897| 69  | 90  |
| 97  | 29  | 1   | 21  | 71  | 156  | 97  | 4   | 4.654 | 90  | 162 |
| 98  | 64  | 2   | 27.3| 109 | 186  | 38  | 5   | 5.3083| 99  | 150 |
| 99  | 41  | 1   | 34.6| 87.33| 205 | 41  | 5   | 4.6728| 110 | 279 |
| 100 | 49  | 2   | 25.9| 91  | 178  | 52  | 3   | 4.5747| 75  | 92  |

† The variable ID indicates the number of rows in the data file, which does not appear in the original.

As a preliminary analysis of the diabetes data, we separated observations into two sets based on precision and applied a multivariate linear regression model to each dataset. The prior distribution was specified as noted in Section 4. The marginal likelihood was maximized with \((x_2, x_3, x_9, x_{10})\) when only precise data were used and with \((x_2, x_3, x_4, x_7, x_9)\) when only imprecise data were used. The results suggest that the data are obtained from two different populations, keeping in mind that the two populations also have common aspects: they are collected from diabetes patients. How the populations are different/similar will be investigated in Subsection 5.2.
3 Two-dataset problem

Suppose $n$ observations are independently drawn from one of two populations. These two populations may be the same or different. We know which observation comes from which population in terms of, say, precision, but have no knowledge of whether the two populations are identical. In this paper, this is called the two-dataset problem.

Let $S_0$ and $S_1$ be sets of observation indices of samples drawn from the first and second populations, respectively. Then, $\{1, \ldots, n\} = S_0 \cup S_1$ and $S_0 \cap S_1 = \emptyset$. In the context of our application, we use $S_0$ for low-quality data and $S_1$ for high-quality data. Specifically, $|S_0| = 77$ and $|S_1| = 221$ in the geyser dataset, while $|S_0| = 377$ and $|S_1| = 65$ in the diabetes dataset.

We apply linear regression to this type of dataset to infer the underlying mechanism between the response and predictors. If the assumption that the data are from the identical population is the state of nature, estimation methods such as ordinary least squares with the entire dataset are asymptotically correct. If the data are from different sources, pooled estimation is not appropriate, and estimation with separate datasets is preferable. However, as in the diabetes dataset, the data may be from different populations that share some aspects to a certain degree. To address this issue, we consider two aspects of the model: regression coefficients and a set of predictors. When these factors are taken into account, we can obtain a criterion that sheds light on the two-dataset problem.

The general prediction error considered in this paper is the following. Let $\hat{f}_0(x)$ be a prediction based on data for $S_0$. Then, the prediction error of data $S_1$ is given by

$$E\left[ L \left\{ Y_1 - \hat{f}_0(x_1) \right\} \right],$$
where \((Y_1, x_1)\) is the response and predictors for \(S_1\) and \(L(\cdot)\) is the loss function. In this paper, we use the squared loss. The expectation is over \(Y_1\) and is conditional on data for \(S_0\) and \(x_1\). Conditional arguments are suppressed for simplicity. This expression is the error of data \(S_1\) given \(S_0\). Similarly, we define the error of data \(S_0\) given \(S_1\). The following two subsections apply this framework to the linear regression model, in which the prediction error is specified by the model parameters and observed data.

### 3.1 Random coefficients model

A possible specification of the two-dataset problem is the random coefficients model (see, e.g., Griffiths et al. [1979]). For each dataset, we consider a linear regression model with different regression coefficients, which is given by

\[
Y_i = \begin{cases} 
  x'_{0i}\beta_0 + \epsilon_{0i}, & \text{for } i \in S_0, \\
  x'_{1i}\beta_1 + \epsilon_{1i}, & \text{for } i \in S_1,
\end{cases}
\]

where \(Y_i\) is the response for observation \(i\) \((i = 1, \ldots, n)\), \(x_{ji}\) is a vector of predictors in model \(j\), and the error term \(\epsilon_{ji}\) has mean 0 and variance \(\sigma_j^2\) for \(j = 0, 1\). In this specification, the predictors are common and arranged in the same order across the two datasets. Throughout this paper, the predictors are standardized within each dataset. When data for \(S_j\) are stacked by rows, the above model can be represented in matrix form as

\[
y_j = X_j\beta_j + \epsilon_j, \text{ for } i \in S_j,
\]

where \(y_j = (Y_i)_{i \in S_j}\), \(X'_j = (x_{ji})_{i \in S_j}\), and \(\epsilon_j = (\epsilon_{ji})_{i \in S_j}\) for \(j = 0, 1\).

To exploit the similarity across datasets, variation in the regression coefficients is explicitly considered, so that the random coefficients assumption
is introduced; that is, 

\[ \beta_1 = \beta_0 + \eta, \]

where \( \eta \) has mean 0 and variance \( \sigma^2_\eta \). Independence between \( \eta \) and \( \epsilon_j \) is assumed \((j = 0, 1)\).

Under the above setting, the following proposition holds.

**Proposition 1.** Let \( \hat{y}_{1|0} \) be the prediction of \( y_1 \) using the least squares estimate of \( \beta_0 \). Then, the expectation of the average squared prediction error is given by

\[
\frac{1}{n_1} E (y_1 - \hat{y}_{1|0})' (y_1 - \hat{y}_{1|0}) = \sigma^2_1 + k \sigma^2_\eta + \frac{\sigma^2_0}{n_1} \text{tr}\{ (X'_0 X_0)^{-1} X'_1 X_1 \},
\]

where \( n_j = |S_j| \) and \( k \) is the number of predictors.

**Proof.** Because the least squares estimate of \( \beta_0 \) is \((X'_0 X_0)^{-1} X'_0 y_0\), the prediction based on data for \( S_0 \) is given by

\[
\hat{y}_{1|0} = X_1 (X'_0 X_0)^{-1} X'_0 y_0 = X_1 (X'_0 X_0)^{-1} X'_0 (X_0 \beta_0 + \epsilon_0)
\]

\[ = X_1 \beta_1 - X_1 \eta + X_1 (X'_0 X_0)^{-1} X'_0 \epsilon_0. \]

The independence of error terms leads to

\[
E (y_1 - \hat{y}_{1|0})' (y_1 - \hat{y}_{1|0}) = E (\epsilon'_1 \epsilon_1) + E (\eta' X'_1 \eta)
\]

\[ + E \left\{ \epsilon'_0 X'_0 (X'_0 X_0)^{-1} X'_1 X_1 (X'_0 X_0)^{-1} X'_0 \epsilon_0 \right\}. \]

After exchanging the trace and expectation operators, the right-hand side becomes

\[
n_1 \sigma^2_1 + n_1 k \sigma^2_\eta + \sigma^2_0 \text{tr}\{ (X'_0 X_0)^{-1} X'_1 X_1 \}.
\]

The second term is obtained because the predictors are standardized, i.e.,

\[
\frac{1}{n_1} \sum_{i \in S_1} x^2_{1,il} = 1 \text{ for all } l = 1, \ldots, k, \text{ where } x_{1,il} \text{ is the } l\text{-th predictor for observation } i \text{ in } S_1. \]

Dividing both sides by \( n_1 \) yields the result. \( \square \)
When the predictors are orthogonal to each other, the third term reduces to \( k\sigma_0^2/n_0 \).

### 3.2 A generalization of the random coefficients model

The random coefficients model assumes that both datasets follow the same model but can differ in their marginal effects. However, the model is also likely to vary across datasets. In the context of linear regression, different models equate to different sets of predictors. Therefore, we have a generalization of the random coefficients model, which is given by

\[
\begin{align*}
\mathbf{y}_0 &= \mathbf{Z}_0\mathbf{\alpha}_0 + \mathbf{X}_0\mathbf{\beta}_0 + \mathbf{\epsilon}_0, \\
\mathbf{y}_1 &= \mathbf{W}_1\mathbf{\alpha}_1 + \mathbf{X}_1\mathbf{\beta}_1 + \mathbf{\epsilon}_1.
\end{align*}
\]

As in the previous model, \( \mathbf{X}_0 \) and \( \mathbf{X}_1 \) have the same predictors in their corresponding columns, and the homoskedastic random coefficients structure is retained in \( \mathbf{\beta}_j \)s. The error term also has the same mean and variance as in the random coefficient model. The remaining terms, \( \mathbf{Z}_0 \) and \( \mathbf{W}_1 \), may or may not share predictors, and we do not impose any specific structure on their regression coefficients \( \mathbf{\alpha}_j \)s.

The prediction in this model is given by the least squares estimate of \( \mathbf{\beta}_0 \) and \( \mathbf{\alpha}_1 \). More precisely,

\[
\hat{\mathbf{y}}_{1|0} = \mathbf{W}_1\hat{\mathbf{\alpha}}_1 + \mathbf{X}_1\hat{\mathbf{\beta}}_0,
\]

where

\[
\begin{align*}
\hat{\mathbf{\beta}}_0 &= \left( \mathbf{\tilde{X}}_0'\mathbf{\tilde{X}}_0 \right)^{-1} \mathbf{\tilde{X}}_0'\mathbf{y}_0, \\
\hat{\mathbf{\alpha}}_1 &= (\mathbf{W}_1'\mathbf{W}_1)^{-1} \mathbf{W}_1' \left( \mathbf{y}_1 - \mathbf{X}_1\hat{\mathbf{\beta}}_0 \right), \\
\mathbf{\tilde{X}}_0 &= \mathbf{M}_0\mathbf{X}_0, \\
\mathbf{M}_0 &= \mathbf{I} - \mathbf{Z}_0 (\mathbf{Z}_0'\mathbf{Z}_0)^{-1} \mathbf{Z}_0'.
\end{align*}
\]

The matrix \( \mathbf{M}_0 \), which is called the annihilator in econometrics, is the difference between the unit matrix \( \mathbf{I} \) and the projection matrix. The above result
is a direct application of the Frisch-Waugh theorem (see Frisch and Waugh (1933) for this theorem).

Further calculation leads to the following proposition, which is a generalization of the previous one.

**Proposition 2.** The expectation of the average squared prediction error under the model and prediction described above is given by

\[
\frac{1}{n_1} E \left( y_1 - \hat{y}_{1|0} \right) \left( y_1 - \hat{y}_{1|0} \right) = \frac{n_1 - k_1}{n_1} \sigma^2_1 + \frac{n^2}{n_1} \text{tr} \left( \tilde{X}_1' \tilde{X}_1 \right) + \frac{\sigma_0^2}{n_1} \text{tr} \left\{ \left( \tilde{X}_0' \tilde{X}_0 \right)^{-1} \tilde{X}_1' \tilde{X}_1 \right\},
\]

where \( k_1 \) is the number of predictors in \( W_1 \), \( \tilde{X}_1 = M_1 X_1 \), and \( M_1 = I - W_1 (W_1' W_1)^{-1} W_1' \).

The proof is very similar to the previous one, and is thus left to Appendix A. When the constant regressor is explicitly included in the model and is in \( X_j \)s, \( n_1 \) and \( \frac{n_1}{n_0} \) are added to the first and second trace terms, respectively. In this paper, we always include the constant regressor in the common predictors \( (X_j)s \).

### 4 Estimation of the prediction error

Three parameters \( (\sigma_j^2, \sigma^2_\eta) \) need to be established in order to estimate the prediction error. We take the Bayesian approach to their estimation (see, e.g., Berger (1985) for its foundations).

The variances of the regression errors are estimated as their posterior means, given the models. In the following analyses, we assume the \( g \)-prior on the regression parameters (see Zellner (1986); Zellner and Siow (1980)). The hierarchical prior on \( g \) is introduced, and we use the hyper-\( g \) prior,
setting its hyperparameter as \( a = 3 \) (see Liang et al. (2008)). The remaining parameters are assumed to follow noninformative prior distributions: the prior for the intercept is proportional to a constant, and the priors for the variances of the regression errors are \( \pi(\sigma_j^2) \propto \sigma_j^{-2} \) \( (j = 0, 1) \). With this prior specification, these posterior means are analytically tractable (see, e.g., Miyawaki and MacEachern (2019)).

The variance of the random coefficients structure, \( \sigma^2_\eta \), is estimated as the sample variance of the posterior means of the regression coefficients without correcting for the degrees of freedom. This method is a convenient approximation of the estimate from the model specifying standard priors on all model parameters. We note that our choice can be viewed as another Bayesian approach with a specific prior on \( \sigma^2_\eta \). We will not pursue this specification further in this paper. The performance of this method will be discussed in the next subsection.

### 4.1 Estimation of \( \sigma^2_\eta \)

This subsection evaluates how the estimation approach of \( \sigma^2_\eta \) affects the final result. One possible alternative is to assume a standard prior on \( \sigma_\eta \) and estimate the model parameters in a Bayesian manner. We assume the half standard Cauchy prior distribution on \( \sigma_\eta \) (see, e.g., Gelman (2006) for a discussion of this prior specification). The other priors remain the same.

The posterior means under this prior setting are not analytically tractable, so we use the Markov chain Monte Carlo method to draw samples from the posterior and estimate the posterior means. The full conditional distributions are derived in Appendix B. The following comparison is performed model by model, and we restrict our attention to models in which the number of common predictors (not including the constant regressor) is greater than seven.
The diabetes data are used for the comparison, and the resulting number of models is 436.

First, we compare the posterior means of $\sigma_{\eta}^2$, of which the differences for the 436 models are summarized in the second row of Table 2. The $\hat{\sigma}_{\eta}^2$s based on the approximated approach are larger than those estimated from the alternative Bayesian approach, which is partly due to the prior specification. In addition, $\sigma_j^2$s ($j=0,1$) are affected in a similar manner, although their differences are relatively small (see the last two rows of this table).

This alternative specification has two disadvantages. First, the specification is computationally intensive, because we need to use the Markov chain Monte Carlo method to estimate the posterior means, which, for this subset of models, takes approximately 20 minutes with the Fortran language and a 3.4 GHz Intel Core i7 computer. The number of models grows rapidly when we investigate the heterogeneity/similarity among observations (see Subsection 5.2), which makes this specification infeasible. Second, the estimation is sensitive to the prior when the number of common predictors is small. When only the constant regressor is common, as an extreme case, the full conditional for $\sigma_{\eta}^2$ does not have a mean, which produces a larger posterior mean. Future research will involve a search for a prior that is both robust to the number of predictors and simple for estimation.

Table 2: Difference between the alternative Bayesian and approximated approaches

|               | Minimum | 1st Quartile | Median | Mean (SD)          | 3rd Quartile | Maximum |
|---------------|---------|--------------|--------|--------------------|--------------|---------|
| $\hat{\sigma}_{\eta}^2$ | -0.088  | -0.014       | -0.0063| -0.012 (0.015)     | -0.0035      | 0.00040 |
| $\hat{\sigma}_0^2$    | -0.00078| -0.00037     | -0.00025| -0.00015 (0.00037)| -0.00011     | 0.0016  |
| $\hat{\sigma}_1^2$    | -0.021  | -0.0066      | -0.0038| -0.0042 (0.0063)   | -0.0018      | 0.023   |
5  Illustrative applications

5.1  Internal quality

This subsection focuses on the problem raised by data quality and applies Proposition 1 to the geyser data. Because they are repeated observations, we are sure that they are from the same population. The response is the logarithm of waiting time at time $t + 1$, and two potential predictors are considered: the duration at time $t$ (denoted by $x_1$), and an indicator variable that is one if the duration is less than or equal to 3 and is zero otherwise (denoted by $x_2$).

Table 3 presents the prediction errors estimated with the approximated approach when changing a set of predictors. The first column indicates which predictors are included in the model, noting that the constant regressor is always included. The last column gives the estimated prediction error, which is the sum of the preceding three columns (the second to fourth columns), as presented in Proposition 1.

That the prediction errors of $S_0 | S_1$ are smaller than those of $S_1 | S_0$ suggests that precise data are useful for developing better models at the cost of collecting high-quality data. However, because the difference is small, it is reasonable to pool the entire set of data to estimate the model parameters of interest. Furthermore, the binary variable ($x_2$) is comparable to the continuous variable ($x_1$) in terms of prediction, indicating that the rounded measurements at night are sufficient for prediction based on linear regression.

5.2  Similarity search

This subsection considers the two-dataset problem characterized by different data sources. Proposition 2 can be used to evaluate the heterogene-
Table 3: Internal quality of the geyser data

| Predictors | $S_0$ | $S_1$ | $S_1$ | $S_0$ |
|------------|-------|-------|-------|-------|
|            | Term 1 | Term 2 | Term 3 | Prediction error |
| $\emptyset$ | 0.032  | 0      | 0.00021 | 0.032 |
| $x_1$ | 0.0068  | 0.0020 | 0.00008 | 0.0089 |
| $x_2$ | 0.0078  | 0.00010 | 0.00008 | 0.0080 |
| $(x_1, x_2)$ | 0.0069 | 0.0028 | 0.00009 | 0.0098 |

In general, when we have two possible model specifications, the model with the smaller prediction error is the better model, in terms of prediction. In the two-dataset problem, we obtain two prediction errors by changing 0 and 1. It is therefore reasonable to select the model with the smallest sum of the prediction errors. Alternatively, the model with the smallest maximum can be selected, though we do not take this specification in this paper. This process is a specific form of cross-validation when two datasets are drawn from the identical population. In the two-dataset problem, however, we allow for the datasets to be drawn from potentially different populations. Even when this is the case, a smaller prediction error based on the generalized random
coefficients model indicates a better model.

One application of this approach is to find common predictors across datasets. Assume first that one predictor is common across datasets in the state of nature. We then have two specifications: one with the predictor in the model and the other without the predictor. A smaller prediction error in the former is an indication that the predictor is common across datasets and a support for the assumption. When several specifications share predictors of interest (or several specifications do not share them), a possible approach is to take average possible models, which is reasonable under the squared loss function (see also Miyawaki and MacEachern (2019)).

Because the Bayesian approach is cohesive framework for estimating model parameters and model uncertainly, we use it to estimate the prediction errors and average them with the posterior model probability, which is proportional to the product of the marginal likelihood and prior model probability given a model. When the uniform prior model probability is used, as in this paper, the result is proportional to the marginal likelihood.

The selection process in this context is summarized in the following five steps.

**Step 1.** Select a set of predictors assumed to be common across datasets.

**Step 2.** Given the models represented by Equation (1) (one for $S_0$ and the other for $S_1$), estimate the prediction error.

**Step 3.** Average the error over possible models based on their posterior model probabilities.

**Step 4.** Repeat the three steps above by changing the labels 0 and 1; the prediction error for this set of common predictors is the sum of the two errors.
Step 5. Choose the set of predictors that minimizes the prediction error as common predictors across datasets.

The diabetes data are used to illustrate the application of our selection of common predictors. The response is the log of the diabetes progression measure, and the ten predictors explained in Section 2 are used. The maximum marginal likelihood criterion indicates that $x_2$, $x_3$, and $x_9$ are common across datasets (see Section 2). Miyawaki and MacEachern (2019) report that $x_3$, $x_5$, and $x_9$ give the least mean squared loss when using the same model across all observations.

We performed our selection process as described above. The selection based on the prediction error is summarized in the left panel of Figure 2. The top plot draws the prediction error in ascending order, and the bottom map shows the corresponding common predictors. When the row for a predictor—for example, $x_1$—is black, it is selected as a common predictor.

Figure 2: Selection of common predictors with the diabetes data.
We note two findings observed in this figure. First, the prediction error increases rapidly at the right side of the figure. This increase appears to be closely related to the inclusion of $x_5$ and $x_6$ as common predictors. Further, they are not included as common predictors in the top one hundred models, as found in the right panel of Figure 2. There results suggest that these predictors vary across datasets. The smallest prediction error is achieved with only $x_2$ as a common predictor.

The prediction error is decomposed into the three terms in Proposition 2, as shown in Figure 3. The sharp increase in the prediction error is mainly due to the second term. Although this result is for a specific dataset, the random coefficients structure is useful for model selection.

6 Conclusion

When data are obtained from two potentially different sources, the issue of whether we should pool the data to estimate the parameters of interest
is an important one. One conservative choice is to separate the data and perform estimation using different models for different datasets, at the cost of inaccurate estimates due to reduced information from the subset of data. As we found for the geyser dataset, the difference between the two datasets can be small in terms of prediction. However, in general, the difference can be large, so that different models are required. The second prediction error or other variable selection method such as the one proposed by Miyawaki and MacEachern (2019) is useful for finding common predictors to aid the model building process. The former is demonstrated with the diabetes dataset.

A possible empirical application of our method is the determinant of growth, which is an important question in economics that has been analyzed in a number of papers (see, e.g., Fernández et al. (2001) and references therein). However, the conclusion may change when we take the two-dataset problem into account. Economic theory typically assumes one common mechanism of growth. This assumption may be true, but heterogeneity in data quality exists, and data may differ across countries for several reasons. Our approach is useful in either case.

We would like to note two future issues to address. The first is the extension of the case, where we have more than two datasets. In this case, the computational burden would increase by a power of more than two, making our estimation infeasible. Improvement in the methodology and/or computation would be necessary to reduce the computational burden.

The second issue is the prior on $\sigma^2$. For its square root, the half Cauchy distribution is used in this paper, but this specification requires the Markov chain Monte Carlo sampling. To make our method scalable, a reasonable prior with a shorter computational time must be developed.
A Proof of Proposition 2

The prediction error is given by

\[ y_1 - \hat{y}_{1|0} = y_1 - Z_1 \hat{\alpha}_1 - X_1 \hat{\beta}_0 = M_1 y_1 - M_1 X_1 \hat{\beta}_0 = M_1 (y_1 - Qy_0), \]

where \( Q = X_1 (X_0'X_0)^{-1} X_0' \). Linear models for \( y_j (j = 0, 1) \) and the random coefficients structure give the right-hand side as

\[ M_1 (\epsilon_1 + X_1 \eta - Q \epsilon_0). \]

We note that \( QX_0 = X_1 \) and \( QZ_0 = O \), where \( O \) is the null matrix. The independence between errors gives the expected sum of squared prediction errors as

\[ E (y_1 - \hat{y}_{1|0})' (y_1 - \hat{y}_{1|0}) = (n_1 - k_1) \sigma_1^2 + \sigma_0^2 \text{tr} (X_1' M_1 X_1) + \sigma_0^2 \text{tr} (Q'M_1 Q). \]

Because the annihilator is idempotent, the first trace term is \( \text{tr}(X_1'X_1) \). The second trace term then becomes

\[
\text{tr} (Q'M_1 Q) = \text{tr} \left\{ (X_0'X_0)^{-1} X_1' X_1 \right\} - \text{tr} \left\{ (X_0'X_0)^{-1} X_1' P_1 X_1 \right\} \\
= \text{tr} \left\{ (X_0'X_0)^{-1} X_1' X_1 \right\},
\]

where \( P_1 = W_1 (W_1' W_1)^{-1} W_1' \). Collecting terms and dividing both sides by \( n_1 \) yields the result.

B Full conditional distributions for the random coefficients model

The model is represented by Equation (1) and the random coefficients structure. Because we always include the intercept and assume the random coef-





ficients structure on it, the model specification in this context is given by

\[
y_0 = Z_0\alpha_0 + \xi_0 + X_0\beta_0 + \epsilon_0,
\]

\[
y_1 = Z_1\alpha_1 + \xi_1 + X_1\beta_1 + \epsilon_1,
\]

\[
\begin{pmatrix}
\xi_1 \\
\beta_1
\end{pmatrix} = \begin{pmatrix}
\xi_0 \\
\beta_0
\end{pmatrix} + \eta.
\]

For brevity, we replace \(W_1\) with \(Z_1\) in this appendix. Let

\[
\Lambda_j^{-1} = \left(Z_j'X_j\right) \left(Z_j'X_j\right)' , \quad (j = 0, 1).
\]

The prior distributions are specified in Section 4.

Then, the combination of the \(g\)-prior and the random coefficients structure results in the following conditional distribution,

\[
\begin{pmatrix}
\xi_j \\
\alpha_j \\
\beta_j
\end{pmatrix} \mid \xi_{1-j}, \beta_{1-j} \sim N \left\{ \begin{pmatrix}
\xi_{1-j} \\
\alpha_j \\
\beta_j
\end{pmatrix} , \begin{pmatrix}
\Sigma_j \\
0 \\
\sigma^{-2}_{\eta}\beta_{1-j}
\end{pmatrix} , \begin{pmatrix}
\sigma^2_{\eta} & 0' \\
0 & \Sigma_j
\end{pmatrix} \right\},
\]

where \(0\) is a vector of zeros, \(N(\mu, \Sigma)\) denotes the multivariate normal distribution with mean vector \(\mu\) and covariance matrix \(\Sigma\), \(g_j\) is the hyperparameter in the \(g\)-prior, and

\[
\Sigma_j^{-1} = g_j^{-1}\sigma^{-2}_{\eta}\Lambda_j^{-1} + \begin{pmatrix}
O \\
O \\
O \\
\sigma^{-2}_{\eta}I
\end{pmatrix},
\]

for \(j = 0, 1\).

**The full conditional for** \(\xi_j, \alpha_j, \beta_j\) \((j = 0, 1)\).

It is the multivariate normal distribution with mean \(m_j\) and covariance matrix \(S_j\); that is,

\[
\begin{pmatrix}
\xi_j \\
\alpha_j \\
\beta_j
\end{pmatrix} \mid \xi_{1-j}, \beta_{1-j} \sim N \left( m_j, S_j \right),
\]
where

\[ S_j^{-1} = \begin{pmatrix} \sigma_\eta^{-2} + \sigma_j^{-2}n_j & 0' \\ 0 & \tilde{\Sigma}_j^{-1} \end{pmatrix}, \]

\[ m_j = \begin{pmatrix} \sigma_\eta^{-2} + \sigma_j^{-2}n_j \left( \sigma_\eta^{-2} \xi_{1-j} + \sigma_j^{-2}n_j \bar{y}_j \right)^{-1} (\sigma_\eta^{-2} \xi_{1-j} + \sigma_j^{-2}n_j \bar{y}_j) \\ \tilde{\Sigma}_j \left( \sigma_j^{-2} Z_j' y_j - \sigma_j^{-2} \beta_{1-j} + \sigma_j^{-2} X_j' y_j \right) \end{pmatrix}, \]

\[ \tilde{\Sigma}_j^{-1} = (1 + g_j^{-1}) \sigma_j^{-2} \Lambda_j^{-1} + \begin{pmatrix} O & O \\ O & \sigma_\eta^{-2} I \end{pmatrix}. \]

The full conditional for \( \sigma_j^2 (j = 0, 1) \).

It is the inverse gamma distribution; that is,

\[ \sigma_j^2 | \xi_j, \alpha_j, \beta_j \sim IG \left( \frac{n_j}{2}, \frac{e_j^' e_j}{2} \right), \]

where \( IG(a, b) \) is the inverse gamma distribution with shape parameter \( a \) and scale parameter \( b \) and \( e_j = y_j - Z_j \alpha_j - \xi_j - X_j \beta_j \).

The full conditional for \( g_j (j = 0, 1) \).

It is a nonstandard distribution with a conditional density function proportional to

\[ g_j^{-k/2} (1 + g_j)^{-a/2} \exp \left( -\frac{\beta_j' X_j' X_j \beta_j}{2\sigma_j^2} \right), \]

where \( k \) is the number of predictors in \( X_j \) and \( a \) is the parameter in the prior distribution on \( g_j \), which is equal to 3 in our applications. The Metropolis-Hasting algorithm with the inverse gamma proposal is used to draw a sample from this density. The proposal distribution is specified as

\[ IG \left( \frac{k}{2}, \frac{\beta_j' X_j' X_j \beta_j}{2\sigma_j^2} \right). \]

The full conditionals for \( \sigma_\eta^2 \) and the auxiliary variable.
As shown in [Wand et al. (2011)] and [Makalic and Schmidt (2016)], the half standard Cauchy distribution can be represented as a mixture of two inverse gamma distributions: when $U \sim IG(1/2, 1)$ and $\sigma^2_\eta \mid U \sim IG(1/2, U^{-1})$, $\sigma_\eta$ follows the half standard Cauchy distribution.

By means of this fact, the generation of $\sigma^2_\eta$ is augmented by the auxiliary variable $U$. Their full conditionals are $IG\left(\frac{k+2}{2}, \frac{S_\eta^2}{2} + U^{-1}\right)$, where $S_\eta = (\xi_1 - \xi_0)^2 + (\beta_1 - \beta_0)'(\beta_1 - \beta_0)$ for $\sigma^2_\eta$ and $IG(1, 1 + \sigma_\eta^{-2})$ for $U$.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21(1), 243–247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22(1), 203–217.

Azzalini, A. and A. W. Bowman (1990). A look at some data on the Old Faithful Geyser. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 39(3), 357–365.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer series in statistics. New York: Springer-Verlag.

Brown, P. J., M. Vannucci, and T. Fearn (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(3), 519–536.

Cox, D. and E. Snell (1981). *Applied Statistics: Principles and Examples*. Boca Raton: CRC Press.
Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Fernández, C., E. Ley, and M. F. Steel (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*(5), 563–576.

Frisch, R. and F. V. Waugh (1933). Partial time regressions as compared with individual trends. *Econometrica 1*(4), 387–401.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*(3), 515–534.

Griffiths, W. E., R. G. Drynan, and S. Prakash (1979). Bayesian estimation of a random coefficient model. *Journal of Econometrics 10*(2), 201–220.

Hahn, P. R. and C. M. Carvalho (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association 110*(509), 435–448.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423.

Makalic, E. and D. F. Schmidt (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters 23*(1), 179–182.

Mallows, C. L. (1973). Some comments on $c_p$. *Technometrics 15*(4), 661–675.

Mallows, C. L. (1995). More comments on $c_p$. *Technometrics 37*(4), 362–372.
Miyawaki, K. and S. N. MacEachern (2019). Economic variable selection. Available at https://arxiv.org/abs/1903.02136.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437), 179–191.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Steel, M. F. (2019). Model averaging and its use in economics. Forthcoming in *Journal of Economic Literature*. Available at https://arxiv.org/abs/1709.08221.

Wand, M. P., J. T. Ormerod, S. A. Padoan, , and R. Frühwirth (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* 6(4), 847–900.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Volume 6 of *Studies in Bayesian Econometrics and Statistics*, Chapter 15, pp. 233–243. Amsterdam: North-Holland/Elsevier.

Zellner, A. and A. Siow (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, pp. 585–603. Valencia: University Press.