Did Oldham Discover the Core After All? Handling Imprecise Historical Data with Hierarchical Bayesian Model Selection Methods

Jack B. Muir*[1] and Victor C. Tsai[2]

Abstract

Historical seismic data are essential to fill in the gaps in geophysical knowledge caused by the low rate of significant seismic events. Handling historical data in the context of geophysical inverse problems requires special care, due to the large errors in the data collection process. Using Oldham's data for the discovery of Earth's core as a case study, we illustrate how a hierarchical Bayesian model selection methodology using leave-one-out cross validation can robustly and efficiently answer quantitative questions using even poor-quality geophysical data. We find that there is statistically significant evidence for the existence of the core using only the *P*-wave data that Oldham effectively discarded in his discussion.

Cite this article as Muir, J. B., and V. C. Tsai (2020). Did Oldham Discover the Core After All? Handling Imprecise Historical Data with Hierarchical Bayesian Model Selection Methods, *Seismol. Res. Lett.* **91**, 1377–1383, doi: 10.1785/0220190266.

Supplemental Material

Introduction

Seismologists are highly motivated to study historical data due to the long timescales of geophysical processes compared to the human lifespan, and especially compared to the proliferation of modern digital instrumentation. A crucial consideration in seismology is that both the quantity and quality of seismic data are ever increasing, and that as the density of instrumentation increases, so too does our ability to accurately locate events in space and time and the number of useful recorded events greatly increase. When dealing with historical data, we must, therefore, unfortunately contend with the reversal of these trends, so that we are left with fewer data of poorer quality. Overcoming these deficiencies requires careful treatment of noise in the data. The required tools are provided by Bayesian analysis, which allows us to rigorously derive posterior probability distributions for models given observed data and explicitly quantified priors (Tarantola, 2005). The ability of Bayesian analysis to quantitatively encode *a priori* information is especially important for historical data, where the information provided by the data is relatively uninformative.

In this study, we focus on a particular type of data—pairs of station–receiver distance and travel times from earthquakes. These data have long been central to geophysical imaging, especially before the advent of computationally feasible waveform inversions. Because of the computational expense of simulating waveforms and the requirement that waveform methods have an accurate starting model, seismic tomography from travel-time data still holds a central position in the hierarchy of geophysical methods. When dealing with travel-time data, the tomographer's hope is that errors in the space–time location of an earthquake do not significantly contribute to the observed residuals used for inversion, or that they may at least be minimized by some relocation method. For historical data, the errors are often so large as to make this impossible, so any analysis requires that we explicitly handle errors in both distance and time. Classical regression methods such as orthogonal distance regression can handle this case when the model to be fitted is smooth and the ratios between the errors for distance and time are known (Boggs and Rogers, 1990). However, as the error ratio is generally not known, a full analysis requires marginalizing over all possible reasonable combinations of errors. An analytical solution to this problem for linear models and specific noninformative priors is given in a manuscript by Jaynes (1999), left, like much of his work, unfinished by his death. For nonlinear models and arbitrary priors, a numerical approach is required. For this study, we present a Bayesian analysis for nonlinear models of imprecise data using Markov chain Monte Carlo (MCMC) sampling and show how to incorporate it into a model selection framework. We apply the model selection framework to some of the most historically important data ever presented in seismology—the famous travel-time curve of Oldham (1906),

1. Seismological Laboratory, California Institute of Technology, Pasadena, California, U.S.A.; 2. Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, Rhode Island, U.S.A.

*Corresponding author: jmuir@caltech.edu

demonstrating how model selection can make a concrete case for the seismic observation of the Earth's core using only a subset of Oldham's data.

It is well known in the seismological community that Oldham provided the first strongly accepted seismic evidence for the Earth's core in his seminal paper *The Constitution of the Interior of the Earth, as Revealed by Earthquakes* (Oldham, 1906), for which he is generally credited with the seismic discovery of the core (Brush, 1980). Various geophysical arguments throughout the 1800s had suggested a core, most notably the arguments of Wiechert, which determined the parameters of a core model from geodetic observations combined with the calculated moments of inertia of the Earth (Wiechert, 1897). However, direct observation of the core was unavailable until the development of quantitative seismology. Oldham provided a travel-time curve for primary and secondary phases derived from teleseismic earthquake records and correctly postulated their mechanical behavior as being those of *P* and *S* waves, respectively. The curvature of travel time strongly suggested to him that the waves traveled deeply in the Earth and were therefore capable of informing us about properties far into the interior. In the travel-time curve, there is an apparent break in the behavior of the curves at around 120° epicentral distance, from which Oldham inferred the existence of the core. The change in character is much more apparent for the secondary arrivals than for the primary arrivals; indeed, Oldham states that it would have "probably remained undetected were it not for the very conspicuous alteration in the case of the second-phase waves." (Oldham, 1906, p. 471). As such, Oldham predicated most of his argument on the secondary arrivals. Unfortunately, immediately after the publication of the original paper it became apparent that the change in secondary arrival behavior was in fact caused by the difference between *S* and *SS* phases, and the apparent large jump in travel time was not due to transmission through the core —consequently Oldham has to a certain extent been lauded for his discovery of the Earth's core under false pretenses (Brush, 1980)! Figure 1 shows Oldham's data, taken from Oldham (1906, table 1) for averaged points and digitized from Oldham (1906, fig. 1) for nonaveraged points, and overlaid with modern travel-time curves from the ak135 model (Kennett *et al.*, 1995). It is apparent that the later primary phase data are likely core interacting *P* phases of the *PKP* family. However, the scatter is extreme for the primary arrivals, and it is difficult by eye to confidently claim that there is any meaningful change in the travel-time curve. It is, therefore, a point of historical interest whether it is in fact possible to deduce the seismic existence of the core using only the primary (*P*) data presented in Oldham's paper. If we can show that there is a statistically significant change in behavior of the *P* travel-time curve, then Oldham's deduction stands up even without the secondary (*S*, *SS*) data. Because of the highly imprecise nature of Oldham's data, this question provides an excellent case study for the handling of historical data using hierarchical Bayesian methods.

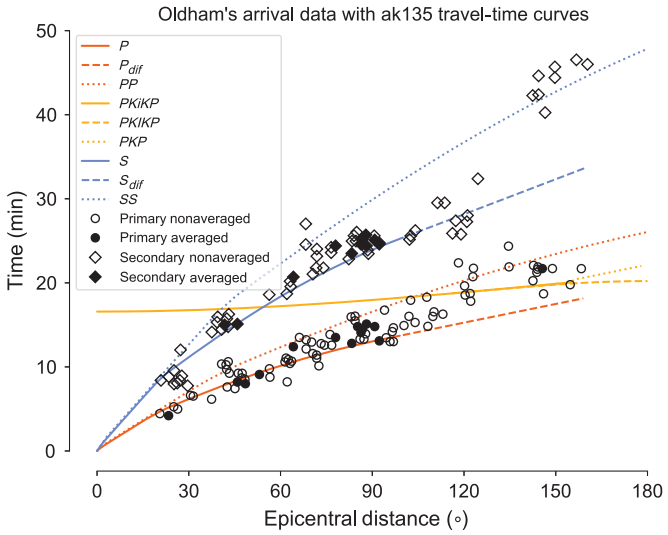

Figure 1. Data from Oldham (1906), with modern travel-time curves from ak135 overlaid (Kennett *et al.*, 1995). Assignments to primary and secondary arrivals are from Oldham.

Data and Methodology

Oldham's *P*-phase travel-time data consist of distances **d** and times **t**. To normalize the data to the interval (−1, 1) for curve fitting, we subtract the mean and divide by the range of both **d** and **t**.

We treat the question of detection of the core as one of quantitative Bayesian model selection. In particular, we propose several models for the data, some containing only one predicted travel-time curve, and some containing two. Following Oldham, we infer the presence of the core if a candidate model with two disjoint travel-time curves is significantly better at predicting the data than models with only one curve. This decision criterion (i.e., whether to choose the model that has the best predictive performance) is a philosophical choice—for instance, if the true data generating process was one of the candidate models, it is not guaranteed that we would recover it (Shao, 1993). For geophysical data, however, the true data generating process is almost always unavailable and not included as a candidate; selecting a model that best predicts the observations is often the most sensible choice from a practical standpoint. The predictive criterion is data driven and does not explicitly penalize model complexity, but instead relies on the tendency of unnecessarily complex models to overfit the data and, therefore, perform poorly at prediction for unseen data.

As mentioned before, Oldham's data are extremely imprecise both in time and epicentral distance compared to modern standards. As such, treating the data properly requires that we handle unknown errors on both axes. As earlier mentioned, frequentist methods such as orthogonal distance regression can fit curves to data with errors in both independent and dependent variables but require *a priori* knowledge of the

relative error. In contrast, hierarchical Bayesian methods allow us to set up the unknown error standard deviations σ_d and σ_t in both distance and time, respectively, as parameters that are inverted, for along with the parameters describing the travel-time model—as these parameters describe the form of the model likelihood and prior distributions they are referred to as hyperparameters in the Bayesian geophysical literature (Malinverno and Briggs, 2004). The model parameters are therefore $\mathbf{m}$, describing the form of the fitted travel-time curves, and σ_d and σ_t. The functional form of the travel-time curve, including any jumps, is given by $f(d, \mathbf{m})$. We assume that the model parameters $\mathbf{m}$ are *a priori* independent from σ_d and σ_t because making observations should not impact the type of models we propose for the travel-time structure, and we also assume that σ_d and σ_t are independent because the scales of the errors in distance and travel time are not correlated.

To compare the predictive performance of different models and determine if there is enough evidence in Oldham's *P* arrival to indicate the existence of the core, we must first derive the posterior distribution for the parameters of the different models given the observed data. From Bayes' rule, the posterior distribution is given by

$$\begin{aligned} p(\mathbf{m}, \sigma_d, \sigma_t|\mathbf{d}, \mathbf{t}) &\propto p(\mathbf{d}, \mathbf{t}|\mathbf{m}, \sigma_d, \sigma_t)p(\mathbf{m}, \sigma_d, \sigma_t) \\ &= p(\mathbf{d}, \mathbf{t}|\mathbf{m}, \sigma_d, \sigma_t)p(\mathbf{m})p(\sigma_d)p(\sigma_t), \end{aligned} \tag{1}$$

assuming independent priors for $\mathbf{m}$, σ_d, and σ_t. Both $\mathbf{t}$ and $\mathbf{d}$ have significant noise, so we represent the relation between them as

$$\mathbf{t} = f(\mathbf{d} + \mathbf{e}_d, \mathbf{m}) + \mathbf{e}_t, \tag{2}$$

in which $\mathbf{e}_d$ and $\mathbf{e}_t$ are the unknown measurement errors in distance and time, respectively. This formulation implicitly assumes errors in distance and time are independent, which given the majority of earthquake origins in Oldham (1906) are from local reports, rather than by triangulation from travel time, is not unreasonable. As we are ranking different functional forms f, we do not include model uncertainty in this analysis.

To write out the posterior, we introduce dummy variables $\mathbf{D} = \mathbf{d} + \mathbf{e}_d$ and $\mathbf{T} = f(\mathbf{D}, \mathbf{m})$; from a Bayesian standpoint, $\mathbf{D}$ represents the unknown "true" distances in Oldham's data and $\mathbf{T}$ the corresponding "true" times predicted by the travel-time curve. $\mathbf{T}$ is a deterministic function of $\mathbf{D}$ and $\mathbf{m}$ so $p(\cdot|\mathbf{D}, \mathbf{m}) = p(\cdot|\mathbf{T})$. Assuming that $\mathbf{D}$ is independent of $\mathbf{m}$ and given our earlier assumption that uncertainties in $\mathbf{d}$ and $\mathbf{t}$ are independent, we can write the likelihood as a marginalized distribution over $\mathbf{D}$:

$$p(\mathbf{d},\mathbf{t}|\mathbf{m},\sigma_d,\sigma_t) = \int p(\mathbf{d},\mathbf{t}|\mathbf{D},\mathbf{m},\sigma_d,\sigma_t)p(\mathbf{D})d\mathbf{D} \tag{3a}$$

$$= \int p(\mathbf{d}|\mathbf{D},\mathbf{m},\sigma_d)p(\mathbf{t}|\mathbf{D},\mathbf{m},\sigma_t)p(\mathbf{D})d\mathbf{D} \tag{3b}$$

$$= \int p(\mathbf{d}|\mathbf{D},\sigma_d)p(\mathbf{t}|\mathbf{T},\sigma_t)p(\mathbf{D})d\mathbf{D}, \tag{3c}$$

allowing us to write a fully decoupled marginal posterior

$$\begin{aligned} &p(\mathbf{m}, \sigma_d, \sigma_t|\mathbf{d}, \mathbf{t}) \\ &\propto \int p(\mathbf{d}|\mathbf{D}, \sigma_d)p(\mathbf{t}|\mathbf{T}, \sigma_t)p(\mathbf{m})p(\sigma_d)p(\sigma_t)p(\mathbf{D})d\mathbf{D}. \end{aligned} \tag{4}$$

The full posterior, including the dummy variables $\mathbf{D}$, can be written by demarginalizing equation (4) and applying Bayes theorem to obtain

$$p(\mathbf{m}, \sigma_d, \sigma_t, \mathbf{D}|\mathbf{d}, \mathbf{t}) \propto p(\mathbf{D}|\mathbf{d}, \sigma_d)p(\mathbf{t}|\mathbf{T}, \sigma_t)p(\mathbf{m})p(\sigma_d)p(\sigma_t). \tag{5}$$

The hierarchical parameterization used in this study may cause difficulties in efficient MCMC sampling due to the structure of the posterior—we discuss a method to avoid this issue in the supplemental material. The inclusion of noise in the independent variable $\mathbf{d}$ means that the final inverse problem has a free parameter corresponding to every data pair $(\mathbf{d}_i, \mathbf{t}_i)$, plus those used to specify the error scales σ_d and σ_t and the model variables $\mathbf{m}$, meaning that the problem is fundamentally underdetermined and requires careful selection of priors. In addition, as the posterior is relatively high-dimensional, explicit integration over it is intractable. We use MCMC to calculate integrals with respect to the posterior, specifically using Hamiltonian Monte Carlo (HMC) to sample the high-dimensional posterior efficiently (Neal, 2011).

Once the posterior distributions for the candidate models are determined, a metric for comparing them for model selection must be defined. To determine the relative performance of the candidate models, we use leave-one-out cross validation (LOO-CV). LOO-CV estimates the predictive performance of a model by removing datums from the observations one at a time, fitting the model, and then testing the left out datum against the model predictions. The posterior predictive distribution for the *i*th left out datum is $p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i})$. The LOO-CV estimate for N_{data} data points is given by

$$\text{LOO} - \text{CV} = \sum_{i}^{N_{\text{data}}} \log p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i}). \tag{6}$$

For each left out datum, we use MCMC sampling to draw N_{MCMC} samples from the marginal posterior $p(\mathbf{m}, \sigma_d, \sigma_t|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i})$—using MCMC sampling we avoid explicitly integrating over the nuisance parameters $\mathbf{D}_{j\neq i}$. By writing $p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i})$ as a marginalization of the posterior predictive $p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{m}, \sigma_d, \sigma_t)$ with respect to the held out data, we can estimate $p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i})$ using the MCMC draws

$$p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i}) = \int p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{m}, \sigma_d, \sigma_t)p(\mathbf{m}, \sigma_d, \sigma_t|\mathbf{d}_{j\neq i}, \mathbf{t}_{j\neq i})d\mathbf{m}d\sigma_d d\sigma_t \quad (7)$$

$$\approx \frac{1}{N_{\mathrm{MCMC}}} \sum_{n=1}^{N_{\mathrm{MCMC}}} p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{m}_n, \sigma_{d,n}, \sigma_{t,n}), \quad (8)$$

in which $\mathbf{m}_n$, $\sigma_{d,n}$, and $\sigma_{t,n}$ notate the nth MCMC sample of the posterior for the model parameters. For each MCMC sample, we can then calculate

$$p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{m}_n, \sigma_{d,n}, \sigma_{t,n}) = \int p(\mathbf{d}_i, \mathbf{t}_i|\mathbf{m}_n, \sigma_{d,n}, \sigma_{t,n}, \mathbf{D}_i)p(\mathbf{D}_i)d\mathbf{D}_i \quad (9)$$

by explicit numerical integration. Higher values of the LOO-CV score indicate better predictive performance. We, therefore, infer the presence of the core from Oldham's data if a model with a jump in the P travel time has a LOO-CV score at least one standard error higher than all models without a jump. We chose LOO-CV as its estimates of predictive performance are robust and unbiased (Vehtari and Ojanen, 2012). LOO-CV is quite computationally intensive, as it requires an independent MCMC run for each datum, which could motivate the use of less expensive methods such as k-fold cross validation for large data sets. However, Oldham's data consist of only 90 points, so explicit LOO-CV is feasible—we further discuss these convergence performance considerations in the supplemental material.

We use low-degree Chebyshev polynomials of the first kind T_i to define the model travel-time curves, following Oldham's expectation that individual travel-time curves for a single phase should be smooth. For models that contain a jump in the travel-time curve, we use two polynomials to represent the curve before and after the changepoint. We notate the models as $(a, -)$ for single travel-time curves of degree a with—signifying no second travel-time curve for P arrivals, and (a, b) for double travel-time curves of degrees a and b with a jump in travel time. For a single travel-time curve of degree a,

$$f(x, \mathbf{m}) = \sum_{i=0}^{a} \mathbf{m}_i T_i(x). \quad (10)$$

For double travel-time curves, the model parameter vector m contains two sets of polynomial coefficients and the location of the changepoint D_j. The HMC method requires that the posterior be continuously differentiable, so a model containing two travel-time curves with a true discontinuity between them cannot be used. To model the jump in travel time between curves, we instead use a hyperbolic tangent to transition between two polynomials f_1 and f_2 so that

$$f(\mathbf{D}, \mathbf{m}) = \frac{f_1(\mathbf{D}, \mathbf{m})(1 - \tanh(k(\mathbf{D} - \mathrm{D}_j)) + f_2(\mathbf{D}, \mathbf{m})(1 + \tanh(k(\mathbf{D} - \mathrm{D}_j))}{2}, \quad (11)$$

in which the factor of $k = 1000$ ensures that the jump is very sharp, but still continuous.

We use products of univariate normal distributions for $p(\mathbf{d}|\mathbf{D}, \sigma_d)$ and $p(\mathbf{t}|\mathbf{T}, \sigma_t)$, which is appropriate given we expect the data to be independent, and the distribution of residuals is approximately normal. For the prior $p(\mathbf{m})$, we also use a product of normals with large standard deviation ($\sigma = 10$); the purpose of this choice is to constrain $\mathbf{m}$ to reasonable parameter ranges for Chebyshev polynomials on the interval (–1, 1), which is important for the two travel-time-curve cases where the second polynomial may rely on very little data. This choice of prior contains all physically reasonable travel-time curves, and so is only as informative as is required to make the posterior sensible. For the dummy variables $\mathbf{D}$, we use an uninformative uniform prior on the whole real line as the distribution of measurment distances is *a priori* unknown. Based on visual inspection of the data, we use a uniform prior on the range (90°, 130°) for the changepoint $\mathbf{D}_j$. Finally, setting the priors for σ_d and σ_t requires special attention. Because the data are highly scattered, the error parameters trade off very strongly with one another, which can lead to parts of the posterior distribution being so highly curved that MCMC sampling is not feasible despite the rescaling mentioned earlier and further discussed in the supplemental material. These situations have unreasonable choices of the parameter space, with σ_d being almost zero whereas σ_t is extremely large, or vice versa. Solving this issue requires putting an informative prior in the error parameters that stops either σ from being very small or very large, which is justified as it is *a priori* clear that the errors in both distance and time are significant but finite. We use inverse-gamma priors tuned so that $P(\sigma_d \leq 1.38°) = P(\sigma_d \geq 13.8°) = 0.01$ and $P(\sigma_t \leq 0.202\,\mathrm{min}) = P(\sigma_t \geq 2.02\,\mathrm{min}) = 0.01$, which are ranges we feel are reasonable for the error standard deviations given Oldham's description of the data.

Results

We used the STAN HMC sampler (Carpenter *et al.*, 2017) to calculate the LOO-CV score for seven models, detailed in Table 1. About 5000 samples were generated for six chains, with the first 2500 discarded, and the chains were compared to ensure convergence. The best-performing model was (1, 1), which fits a quadratic to the first part of the data and a line to the data after the jump in travel times. The difference in LOO-CV scores for all models relative to (1, 1) is given in Table 2. Both models with two travel-time curves separated by a jump are favored over all models without a jump by at least five times the standard error, indicating that within the

TABLE 1
Catalog of Models Used to Test Oldham's Data

Model	Two Travel-Time Curves?	Total Free Parameters
(1, –)	No	94
(2, –)	No	95
(3, –)	No	96
(4, –)	No	97
(5, –)	No	98
(1, 1)	Yes	97
(2, 1)	Yes	98

context of the models chosen, there is a very significant change in the behavior of the travel-time curve despite the large scatter in the data. As such, the *P*-phase arrival data alone are sufficient to support Oldham's arguments as to the existence of the core in Oldham (1906). However, the data are not sufficient to distinguish between a quadratic or a line for the first part of the travel-time curve, as the difference between models (1, 1) and (2, 1) is not significant. We show model (3, –), which is the best-performing model with only one travel-time curve, in Figure 2a and model (1, 1), the best-performing model with two travel-time curves, in Figure 2b. The mean travel-time curve shown for Figure 2b is smoother than any individual sample of model (1, 1), which has a sharp jump between the two travel-time branches. The mean model predictions for both (3, –) and (1, 1) fall between the P/P_{dif} and *PP* phases of the ak135 model for distances less than 120°, suggesting that Oldham's data are potentially a mix of these phases; for greater distances, the better performing (1, 1) model sits between ak135 *PKiKP* and *PP* at around 120° before moving toward

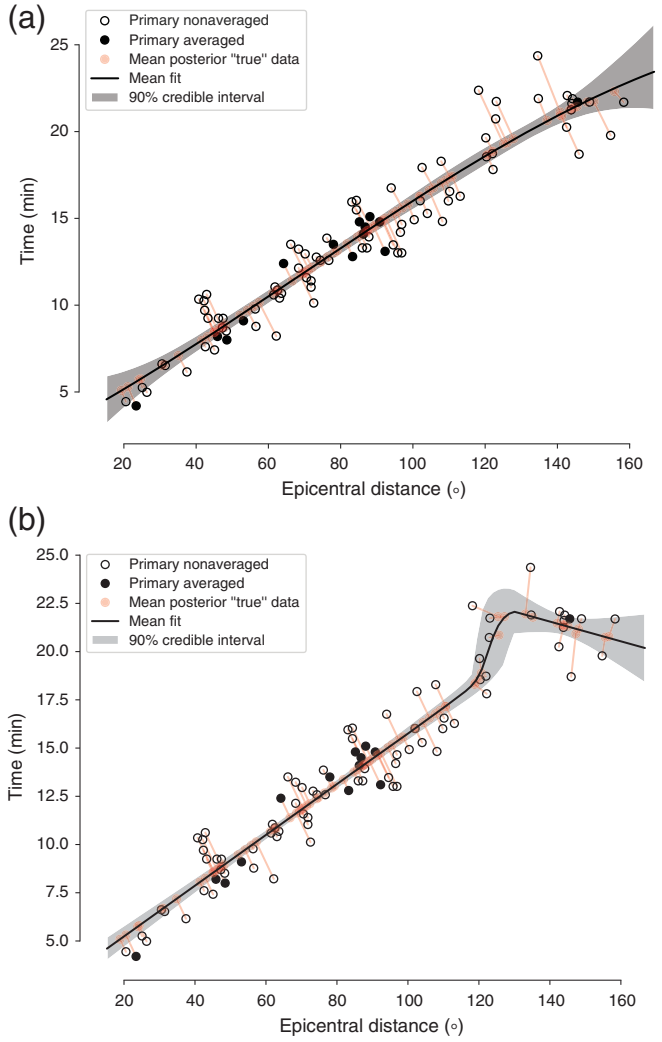

Figure 2. (a) Posterior distribution for model (3, –), the best-performing single travel-time curve model. (b) Posterior distribution for model (1, 1), the best-performing model. Note that the mean shown for (b) is smoother than an individual sample of model (1, 1).

TABLE 2
Difference in Leave-One-Out Cross-Validation (LOO-CV) Score Relative to (1, 1), the Best-Performing Tested Model

Model	LOO-CV Score	Difference in LOO-CV	Standard Error in Difference
(1, –)	23.99	–3.91	0.44
(2, –)	24.03	–3.87	0.40
(3, –)	23.12	–3.78	0.38
(4, –)	23.14	–4.76	0.69
(5, –)	22.55	–5.35	0.95
(1, 1)	27.90	—	—
(2, 1)	27.85	–0.06	0.14

what are likely to be *P* core phases at higher distances. The presence of some *PP* data may explain why the posterior mean of model (1, 1) has negative slowness after the jump in travel time, although the majority of post jump data is closer to the modern core phase times.

From examining the spread of the 90% credible interval (i.e., the area between the 5% and 95% quantiles of the posterior model distribution at each epicentral distance), we can see that (3, –) is more tightly constrained than (1, 1), at the expense of fitting the data significantly worse. For both models, the average correction increases as a function of distance, as is seen in the length of the red connecting lines in Figure 2, indicating that measurements were generally worse fit by a single travel-time curve at longer epicentral distances. The LOO-CV score

balances these concerns and strongly favors models with two travel-time curves. The LOO-CV score degrades substantially from model (3, –) to model (4, –) and model (5, –), which indicates that further higher degrees would perform yet worse in predicting held out data because the single travel-time curve models are strictly nested (i.e., (5, –) contains all of (4, –), which contains all of (3, –) as a special case). Overfitting becomes significant even for simple polynomials due to the high scatter in the data.

Discussion and Conclusions

Model selection, provided by the LOO-CV score, strongly supports there being enough evidence solely in Oldham's *P*-wave data to support two apparently distinct travel-time curves, which leads to Oldham's argument for the core. Although the scientific question presented in this study has not been in question for more than 100 yr, the robust statistical tools required to analyze the problem fully have only recently become available. Model selection, in particular, remains at the forefront of statistical research and has great implications for both traditional inverse theory and newer techniques such as machine learning (e.g., Rasmussen and Williams, 2006; Wit *et al.*, 2012; Claeskens, 2016). The problem of how to perform model selection is unfortunately less resolved than that of how to sample from the Bayesian posterior, for which MCMC sampling, and in particular HMC, has emerged as the clearly preferred technique (Neal, 2011; Betancourt, 2017; Fichtner *et al.*, 2019). Model selection, in contrast, has a plethora of related techniques, ranging in complexity from penalized fits to the maximum *a posteriori* point such as the Akaike and Bayesian information criteria (Claeskens, 2016), cross-validation methods such as that presented here (Vehtari and Ojanen, 2012), and extending to full calculation of the Bayesian evidence. The Bayesian evidence or Bayes factor calculation, in particular, has received attention in geophysics and astronomy because it can be cleanly derived from Bayes' theorem as explicitly comparing the probability of two models given the data. Unfortunately, estimating the Bayesian evidence is highly nontrivial and is typically only shown in the literature for low-dimensional models due to convergence difficulties, which limits its utility for realistic geophysical problems (Friel and Wyse, 2012; Vehtari and Ojanen, 2012). The LOO-CV method used here proved to be tractable for models with ∼100 parameters; however, it does suffer from computational difficulties as the number of data points becomes large. Vehtari *et al.* (2017) give a method for using importance sampling on the posterior MCMC samples using *all* data to approximate the results from held out data, which is promising for large geophysical datasets.

In this study, we have tested only simple functional forms for travel-time curves with and without jumps, with the further restriction that after the jump point all data are assigned to the second travel-time curve. This restriction means that there are no overlapping travel-time curves at any epicentral distance, which restricts our analysis to data sets for which there are not multiple groups of phases observed at a particular distance. Visual inspection of Oldham's data suggests that this simple model is the highest level of complexity warranted by the data. With modern seismic data, however, it is likely that observations at a particular distance will contain multiple phases that need to be classified into different classes. In this case, more advanced modeling strategies that allow the expression of uncertainty as to what phase is being observed, such as Gaussian Mixture Modeling, may be useful (e.g., Grana *et al.*, 2017).

Our study shows how to set up a model that marginalizes over multiple potential sources of error and can be efficiently sampled using HMC. We have shown how careful specification of the prior is especially important in the historical context, where the scale of the data errors are unknown and multiple sources of error may trade off. Together, hierarchical Bayesian modeling and model selection give us a powerful toolbox to explore poor-quality historical data and derive robust conclusions about geophysical processes. In the context of Oldham's travel-time data, it allows us to marginalize out the large errors associated with both the distances and travel times to conclude that there is sufficient evidence contained in the *P* arrivals alone to indicate the existence of the core.

Data and Resources

Historical data were taken from Oldham (1906), either from the reported tables of averaged events or by digitizing the presented travel-time curves. All calculations were performed using the PyStan wrapper of the Stan statistical software package (Carpenter *et al.*, 2017). The supplemental material contains inversion results for the five models not presented in the article (Figs. S1–S5). Additional discussion regarding hierarchical Markov chain Monte Carlo (MCMC) sampling and leave-one-out cross validation (LOO-CV) versus *k*-fold CV are also present in the supplement.

Acknowledgments

The authors would like to thank two anonymous reviewers for their feedback and the *SRL* Editor-in-Chief Allison Bent for managing the review process. The authors would also like to thank Luis Rivera for providing an internal review of this article. J. B. M. would like to thank the General Sir John Monash Foundation and the Origin Energy Foundation for financial support during his graduate studies.

References

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo, available at https://arxiv.org/abs/1701.02434 (last accessed November 2019).

Boggs, P. T., and J. E. Rogers (1990). Orthogonal distance regression, *Contemp. Math.* **112,** 183–194.

Brush, S. G. (1980). Discovery of the Earth's core, *Am. J. Phys.* **48,** no. 9, 705–724.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language, *J. Stat. Software* 76**,** no. 1, doi: 10.18637/jss.v076.i01.

Claeskens, G. (2016). Statistical model choice, *Annu. Rev. Stat. Appl.* **3,** no. 1, 233–256.

Fichtner, A., A. Zunino, and L. Gebraad (2019). Hamiltonian Monte Carlo solution of tomographic inverse problems, *Geophys. J. Int.* **216,** no. 2, 1344–1363.

Friel, N., and J. Wyse (2012). Estimating the evidence—A review, *Stat. Neerl.* **66,** no. 3, 288–308.

Grana, D., T. Fjeldstad, and H. Omre (2017). Bayesian Gaussian mixture linear inversion for geophysical inverse problems, *Math. Geosci.* **49,** no. 4, 493–515.

Jaynes, E. T. (1999). Straight line fitting—A Bayesian solution, presented at the *Tenth Annual MAXENT Workshop*, University of Wyoming, July 1990, available at https://bayes.wustl.edu/etj/articles/leapz.pdf (last accessed October 2019).

Kennett, B. L. N., E. R. Engdahl, and R. Buland (1995). Constraints on seismic velocities in the Earth from traveltimes, *Geophys. J. Int.* **122,** no. 1, 108–124.

Malinverno, A., and V. A. Briggs (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes, *Geophysics* **69,** no. 4, 1005–1016.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, Chapter 5, S. Brooks, A. Gelman, G. l. Jones, and X. Meng (Editors), Chapman & Hall/CRC Press, London, United Kingdom, 50 pp.

Oldham, R. D. (1906). The constitution of the interior of the Earth, as revealed by earthquakes, *Q. J. Geol. Soc.* **62,** nos. 1/4, 456–475.

Rasmussen, C. E., and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, MIT Press, Cambridge, Massachusetts.

Shao, J. (1993). Linear model selection by cross-validation, *J. Am. Stat. Assoc.* **88,** no. 422, 486–494.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, Pennsylvania.

Vehtari, A., and J. Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison, *Stat. Surv.* **6,** 142–228.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Stat. Comput.* **27,** no. 5, 1413–1432.

Wiechert, E. (1897). Ueber die massenverteilung im inneren der erde, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, *Math. Phys. Kl.* **1897,** no. 3, 221–243 (in German).

Wit, E., E. van den Heuvel, and J.-W. Romeijn (2012). 'All models are wrong…': An introduction to model uncertainty, *Stat. Neerl.* **66,** no. 3, 217–236.

Manuscript received 20 September 2019
Published online 15 January 2020