Octorber, 2015. Incomplete Draft. Please do not cite without permission.

Insensitivity of Predictive Accuracy for Selecting among Multilevel Models

Wei Wang
Columbia University
ww2243@columbia.edu

As a simple and compelling approach for estimating out-of-sample prediction error, cross-validation naturally lends itself to the task of model comparison. However, even with moderate sample size, it can be surprisingly difficult to compare multilevel models based on predictive accuracy. Using a hierarchical model fit to large survey data with a battery of questions, we demonstrate that even though cross-validation might give good estimates of pointwise out-of-sample prediction error, it is not always a sensitive instrument for model comparison.

Models selection is an integral part of any data analysis. In an ideal world, iteratively improving and comparing model fits of different specifications should be the routine of all statistical procedures, especailly when developments in methodology and computation facilitate evermore sophisticated and complex models. Often, the most important question is not that whether a more complicated model is computational tractable, but why this model is an improvement over the older and simpler ones. Multilevel models (also known as Hierarchical Models) are an example of modern statistical models, which specifically handles data with group structure, for example, a national survey data with geographic and demographic information or an educational intervention applied to different schools and neighborhoods.

The gold standard of model comparison is out-of-sample prediction accuracy, i.e., in the hypothetical case of more observations coming in, which model gives the best prediction of new case of outcomes based on new cases of predictors. Cross-validation is a perhaps the most widely-used method for estimating out-of-sample prediction

error and comparison of statistical models. By fitting the model on the training data set and then evaluating it on the hold-out testing set, the over-optimism of using data twice is avoided. Furthermore, attempts have been made to use cross-validated objective functions for statistical inference (Craven and Wahba 1978; Seeger 2008), thus integrating out-of-sample prediction error estimation and model selection into one step.

In this chapter, I will discuss several challenges I encounter in using cross-validation predictive accuracy in evaluating and selecting among multilevel models, specifically in binary classification models. The first challenge is the lack of clear protocol for the cross-validation procedure: to truly test the model, the holdout set cannot be a simple random sample of the data but instead needs to have some multilevel structure itself, so that entire groups as well as individual observations are held out. Hierarchical cross-validation can be performed in the context of particular applications (Price, Nero, and Gelman 1996) but it is not clear how best to subsample structured data for cross-validation in a general way. The second challenge is that, in multilevel models, the observed loss function for data-level cross-validation can be so close to flat that the cross-validation estimates of prediction errors under candidate models can be swamped by random fluctuations.

I focus on the second of these concerns, demonstrating the limitations of prediction error in the context of a set of multilevel models fit to a large cross-tabulated national survey. An innovative aspect of our analysis is that we evaluate separately on 71 different survey responses, taking each in turn as the outcome in a comparison of regression models. This allows us to construct a relatively large corpus of data out of a single survey.

This chapter is a joint work with Andrew Gelman, and mostly based on (Wang and Gelman 2014).

MULTILEVEL MODELS AND SURVEY RESEARCH

There are two types of survey researchers, as identified by the classic book "Survey Erros and Survey Costs" (Groves 2004), the *describers*, who "use surveys to descxribe characteristics of a fixed population", and the *modelers*, who "seek to identify causes of phenomena constatly occuring in a society". The latter group developed models to genertae less biased estimates, as a result of using more data and handling more inherent structure within the data. Multilevel models, an example of the *modeler* approach, are effective in survey research, as partial pooling can yield accurate state-level estimates from national polls (Gelman and Hill 2007). Multilevel models have been successfully applied both to representative and nonrepresentative surveys to obtain accurate small-area estimation and prediction (Fay and Herriot 1979; Ghitza

and Gelman 2013; Lax and Phillips 2009; Wang et al. 2014), and the practical application of such methods is currently being actively discussed in social science research (Buttice and Highton 2013; Lax and Phillips 2013). In the present paper, we conduct model selection procedures based on k-fold cross-validation and find that under this framework, the improvement of multilevel models over classical models is surprisingly small when measured on the scale of prediction error. Furthermore, we demonstrate that this lack of notable improvement is related to the sample size and data structure by repeating the analysis on simulated data sets that vary in terms of these two factors.

Our results illustrate that under multilevel structure, it could be tricky to use cross-validation in model selection, as the size of the data and how balanced the structure is heavily affect the relative performance of the models.

In the next section, I will present a fully Bayesian model comparison framework, a preparation for the real data analysis.

MODEL ASSESSMENT AND SELECTION VIA CROSS-VALIDATION

Predictive Loss

I start with a loss function $l(\tilde{y}, a)$ corresponding to the inferential action a_M based on a model M, in face of future observations $\tilde{y}$. The available data, typically consisting of predictors x and outcomes y, are labeled as D. The corresponding predictive loss is then,

$$PL(p^t, M, D) = E_{p^t} l(\tilde{y}, a_M) = \int l(\tilde{y}, a_M) p^t(\tilde{y}) d\tilde{y} \tag{1}$$

where $p^t(\cdot)$ is the true distribution from which the future observations $\tilde{y}$ are generated.

The predictive loss is affected by the form of the action a_M, the loss function l, and the data D. For example, a_M could be the mean of the posterior predictive distribution and l the mean square error loss. However, it is often convenient and theoretically desirable to use the whole posterior predictive distribution as the inferential action and a logarithmic loss function. In addition, using the whole posterior predictive distribution has a Bayesian justification, as it reflects the full inferential uncertainty conditional on the model (Vehtari and Ojanen 2012). Substituting the choice of a_M and l into (1) yields,

$$\begin{aligned} PL(p^t, M, D) &= E_{p^t}[-\log p(\tilde{y}|D, M)] \\ &= -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} \end{aligned} \tag{2}$$

This quantity is central to predictive model selection. The fundamental difficulty in estimating it is that the true distribution $p^t(\cdot)$ is unknown.

Another important quantity arises when we approximate the true distribution with the empirical distribution, which gives the training loss,

$$\begin{aligned} TL(M, D) &= -\int \log p(y|D, M) d\hat{F}(y) \\ &= -\frac{1}{N} \sum_{y \in D} \log p(y|D, M). \end{aligned} \tag{3}$$

The training loss uses the same data for both estimation and evaluation and so in general underestimates prediction error.

Prediction Error

With (2), the model selection task is straightforward. Among the candidate models, the best model under this framework is the one that minimizes the predictive loss:

$$-\min_M \int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y}, \tag{4}$$

which has a lower bound, $-\int p^t(\tilde{y}) \log p^t(\tilde{y}) d\tilde{y}$, which is the entropy of the true distribution. It is often more informative to look at the excess of the predictive loss over this lower bound, as shown in (5). I label this quantity as the prediction error. Conceptually, the prediction error indicates how far the posterior predictive distribution is from the oracle, and it is the Kullback-Leibler divergence between the posterior predictive distribution of the candidate model and the true generative model. As its form suggests, the prediction error is the difference between log posterior predictive density and log true predictive density, averaged over the true predictive distribution,

$$\begin{aligned} &PE(p^t, M, D) = PL(p^t, M, D) - LB(p^t) \\ &= -\int p^t(\tilde{y}) \log p(\tilde{y}|D, M) d\tilde{y} + \int p^t(\tilde{y}) \log p^t(\tilde{y}) d\tilde{y}. \end{aligned} \tag{5}$$

So to estimate the prediction error, we need to estimate the two terms in (5).

k-fold Cross-Validation for Estimating Predictive Loss

In the predictive framework, the central obstacle of estimating the predictive loss (2) is that the future observations are not available. One thread of research attempts to

estimate and correct the bias introduced by reusing the sample and thus gives rise to various information criteria, whose validity hinges on a number of assumptions and simplifications. Another thread of research is to use hold-out data for testing, thus making training and testing data independent. This leads to a variety of resampling procedures, including leave-one-out cross-validation, k-fold cross-validation, Monte Carlo cross-validation, and bootstrapping. In practice, k-fold cross-validation is popular due to its computational convenience and stability (Kale, Kumar, and Vassilvitskii 2011). Formally, the k-fold cross-validation of the predictive loss is given by

$$\begin{aligned} \widehat{PL}^{\text{CV}}(M, D) &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in \text{test}_k}\log p(y_i|D^k, M) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|D^{(\setminus i)}, M), \end{aligned} \tag{6}$$

where D^k represents the k^{th} training set, test_k represents the k^{th} testing set under the random partition and $D^{(\setminus i)}$ denotes the training set that excludes the i^{th} observation. Because k-fold cross-validation does not use all the data, the prediction error estimates are biased, but in the cases where there are relatively few predictors, this bias is small (Burman 1989).

The practical impediment of using cross-validation is the computational burden: with k-fold cross-validation, we need to fit the model k times. However, in many cases it is possible to perform the k steps in parallel.

The problem remains of estimating the second term in (5), namely the lower bound of predictive loss. In this paper, we use the in-sample training loss $TL(M_{\text{s}}, D)$ of the saturated model M_s as the surrogate for the lower bound. So the estimated prediction error is

$$\begin{aligned} &\widehat{PE}(M, D) = \widehat{PL}^{\text{CV}}(M, D) - TL(M_{\text{s}}, D) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|D^{(\setminus i)}, M) + \frac{1}{N}\sum_{y\in D}\log p(y|D, M_{\text{s}}). \end{aligned} \tag{7}$$

Cross-Validation of Structured Data

Standard cross-validation assumes that data are independent and with no distributional differences between the training and testing sets. For structured data, it is not always clear how best to perform this partition. (Burman, Chow, and Nolan 1994) discusses a modification of ordinary cross-validation procedure for stationary time series. In

this paper, we focus on the cross-tabulated structure, which is the characteristic of survey data with discrete responses. In an unbalanced cross-tabulated data set, simple random sampling might result in undersampling of small cells. Thus, we adopt a stratified sampling approach to guarantee that each cell is partitioned into a training part and a testing part. Another possibility is to perform a cluster sampling and train the model on some cells and test the fitted model on others. This approach is related to transfer learning (Pan and Yang 2010). In the analysis of survey data, the focus is mostly on the existing cells rather than on hypothetical new cells, and so we only discuss cross-validation using stratified sampling on structured data.

COMPARING MULTILEVEL MODELS FOR BINARY SURVEY OUTCOMES

The 2006 Cooperative Congressional Election Survey, the example data set in this paper, is a national stratified sample of size 30,000 that includes a wide variety of response outcomes, %(a sample of the questions is listed in Figure~**??**) thus providing an ideal setting to evaluate cross-validation. Although various demographic predictors are available in this data set, we keep our model simple by using only two predictors, state and income. Under this setting, the multilevel model is the preferred model over no pooling (saturated model) or complete pooling (additive model). On one hand, the saturated model will trigger overfitting. On the other hand, income and state are known to have strong interactions when predicting electoral choice (Gelman et al. 2009), so the additive model must be substantively inadequate.

Complete Pooling, No Pooling, and Partial Pooling Models

Bayesian multilevel modeling is a natural choice for analyzing cross-tabulated data. When the data provide many explanatory variables, and thus a potentially complex cross-tabulated structure, it is difficult to model the interactions among explanatory variables in classical models, since each single cell is getting sparser and the estimates become unstable. By borrowing strength across cells, a multilevel model (or, alternatively, some other structured model such as a Gaussian process) can produce stable estimates even for cells that have few observations and thus can be viewed as a multivariate regression or interpolation procedure..

We develop our model on a simple two-way cross-tabulation of survey data, with state and income as the two explanatory variables, having J_1 and J_2 levels respectively.[1] We assume no continuous predictors in our model. Let N be the

[1]For the 2006 Cooperative Congressional Election Survey data set, there are 50 states ($J_1 = 50$), and 5 income levels ($J_2 = 5$), including less than \$20,000, \$20,000-\$40,000, \$40,000-\$75,000, \$75,000-

total sample size of the survey, then the array of cell counts follows a multinomial distribution,

$$\boldsymbol{N} \sim \text{Multinomial}(N, \boldsymbol{p})$$

, where

$$\begin{aligned}\boldsymbol{N} &= (N_{j_1 j_2})_{J_1 \times J_2},\\ \boldsymbol{p} &= (p_{j_1 j_2})_{J_1 \times J_2}.\end{aligned}$$

The population is thus divided into $J_1 \times J_2$ cells. We constrain our discussion to binary outcomes. Then for a respondent in cell (j_1, j_2), the probability that he or she gives a positive response is $\pi_{j_1 j_2}$, which is modeled using logistic regression:

$$\text{logit}(\pi_{j_1 j_2}) = \boldsymbol{Z}\boldsymbol{\beta},$$

in which $\boldsymbol{Z}$ is the covariate vector and $\boldsymbol{\beta}$ includes the main and interaction effects. Since our goal of inference is on cell proportions $\pi_{j_1 j_2}$ rather than cell assignment probabilities $p_{j_1 j_2}$, we treat $p_{j_1 j_2}$ as fixed throughout.

Under this setup, we consider three models:

- Complete pooling of interactions:

$$\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}}\right)$$

- No pooling:

$$\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}} + \beta_{j_1 j_2}^{\text{state*inc}}\right)$$

- Partial pooling:

$$\pi_{j_1 j_2} = \text{logit}^{-1}\left(\beta_{j_1}^{\text{state}} + \beta_{j_2}^{\text{inc}} + \beta_{j_1 j_2}^{\text{state*inc}}\right)$$

  with $\beta_{j_1 j_2}^{\text{state inc}} \overset{i.i.d.}{\sim} \text{N}(0, \sigma^2)$, where the scale parameter σ is estimated from the data (with a separate value for each survey outcome).

Although nonparametric multilevel modeling, both in the Bayesian (Hjort 2010) and the frequentist (Ruppert, Wand, and Carroll 2003) perspectives, have been under rapid development, we adopt a linear parametric specification for the multilevel model, because linear parametric models are still the standard specification, and software that

---

$150,000, and $150,000+.

fit the routine linear parametric models are widely available and easily accessible to practitioners. In the remaining sections of this paper, we compare the prediction error of these three models under various real data and simulation settings.

We recognize that multilevel models in big-data applications can be much more complicated (Ghitza and Gelman 2013); we use a relatively simple example here to explore the basic ideas.

Computation

Ideally we want to do full Bayesian inference on our model, but for computational reasons we are currently using an approximate marginal posterior mode estimate provided by blme (Dorie 2013) in R, which is an extension of the widely-used lme4 (Bates, Maechler, and Bolker 2013) package. The lme4 package approximately integrates out the random effects to obtain an approximate marginal MLE of the scale parameter and the fixed effects. However, modal estimates can end up on the boundary due to sampling variability (Chung et al. 2013), which in our case makes the partial pooling model reduce to complete pooling. In blme, the scale parameter σ is also given a gamma prior with shape parameter 2.5 and rate parameter 0. The gamma prior is used to regularize the prior of the scale and pull the estimates of the interactions away from zero, a situation that often happens in modal estimation. We have developed an R package, mrp (Gelman et al. 2012), to streamline the multilevel model fitting and cross-validation procedure.

Estimation Procedure

For each outcome, we fit a multilevel logistic regression model, with additive, fully-interacted, and multilevel models. We use 5-fold cross-validation to estimate predictive loss (using more folds gives essentially identical results). We estimate the lower bound using the training loss of the saturated model.

Under the aforementioned setting, the cross-validation loss estimate is,

$$\widehat{PL}^{\text{CV}}(M, D) = -\frac{1}{N}\sum_{k=1}^{K}\sum_{j\in \text{test}_k} \log p(y_j|D^k, M)$$

$$= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i,j}[y_{ij}^{\text{test}_k}\log\hat{\pi}_{ij}^{D^k} + (n_{ij}^{\text{test}_k} - y_{ij}^{\text{test}_k})\log(1-\hat{\pi}_{ij}^{D^k})]$$

$$= -\frac{1}{N}\sum_{i,j}\sum_{k=1}^{K}[y_{ij}^{\text{test}_k}\log\hat{\pi}_{ij}^{D^k} + (n_{ij}^{\text{test}_k} - y_{ij}^{\text{test}_k})\log(1-\hat{\pi}_{ij}^{D^k})]$$

$$= -\frac{1}{N}\sum_{i,j}\left[y_{ij}\overline{\log\hat{\pi}_{ij}} + (n_{ij}-y_{ij})\overline{\log(1-\hat{\pi}_{ij})}\right]$$

$$= -\sum_{i,j}\frac{n_{ij}}{N}\left[\pi_{ij}\overline{\log\hat{\pi}_{ij}} + (1-\pi_{ij})\overline{\log(1-\hat{\pi}_{ij})}\right],$$

in which $n_{ij}^{\text{test}_k}$ is the number of respondents in cell (i, j) of the k-th testing set, $y_{ij}^{\text{test}_k}$ is the number of respondents who answered yes in cell (i, j) of the k-th testing set, correspondingly, n_{ij} and y_{ij} are the numbers of total respondents and respondents who answered yes in cell (i, j), $\hat{\pi}_{ij}^{D^k}$ is the estimated π_{ij} using the k-th training data set, and $\overline{\log\hat{\pi}_{ij}}$ is the weighted average log posterior proportion from each fold, $\left(\sum_{k=1}^{K} y_{ij}^{\text{test}_k}\log\hat{\pi}_{ij}^{D^k}\right)/y_{ij}$, and $\overline{\log(1-\hat{\pi}_{ij})}$ has the similar form. The cross-validation loss estimate is approximately a measure of loss under cell proportion distribution $(\exp(\overline{\log\hat{\pi}_{ij}}), \exp(\overline{\log(1-\hat{\pi}_{ij})}))$ (here we say "approximately" because these two probabilities do not in general add up to 1). The quick calculation in section 1.2 suggests that we should expect to see only small improvements in cross-validation loss even from substantively important model improvements.

BIBLIOGRAPHY

Bates, Douglas, Martin Maechler, and Ben Bolker. 2013. *Lme4: Linear Mixed-Effects Models Using S4 Classes*. http://CRAN.R-project.org/package=lme4.

Burman, Prabir. 1989. "A Comparative Study of Ordinary Cross-Validation, V-Fold Cross-Validation and the Repeated Learning-Testing Methods." *Biometrika* 76(3): 503–14.

Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81(2): 351–58.

Buttice, Matthew K., and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21(4): 449–67.

Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika* 78(4): 685–709.

Craven, Peter, and Grace Wahba. 1978. "Smoothing Noisy Data with Spline Functions." *Numerische Mathematik* 31(4): 377–403.

Dorie, Vincent. 2013. *Blme: Bayesian Linear Mixed-Effects Models*. http://CRAN.R-project.org/package=blme.

Fay, R. E., and R. A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–77.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, Andrew, Michael Malecki, Daniel Lee, Yu-Sung Su, and Wei Wang. 2012. *Mrp: Multilevel Regression and Poststratification*. http://CRAN.R-project.org/package=mrp.

Gelman, Andrew, David K. Park, Boris Shor, and Jeronimo Cortina. 2009. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do, Second Edition*. Princeton University Press.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57(3): 762–76.

Groves, Robert M. 2004. 536 *Survey Errors and Survey Costs*. John Wiley & Sons.

Hjort, Nils Lid. 2010. *Bayesian Nonparametrics*. Cambridge University Press.

Kale, Satyen, Ravi Kumar, and Sergei Vassilvitskii. 2011. "Cross-Validation and Mean-Square Stability." In *Innovations in Computer Science,* Tsinghua University Press, 487–95.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53(1): 107–21.

———. 2013. "How Should We Estimate Sub-National Opinion Using MRP? Preliminary Findings and Recommendations." *Presented at Midwest Political Science Association*.

Pan, Sinno Jialin, and Qiang Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22(10): 1345–59.

Price, Phillip N., Anthony V. Nero, and Andrew Gelman. 1996. "Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties." *Health Physics* 71: 922–36.

Ruppert, David, Matt P Wand, and Raymond J Carroll. 2003. *Semiparametric Regression*. Cambridge University Press.

Seeger, Matthias W. 2008. "Cross-Validation Optimization for Large Scale Structured Classification Kernel Methods." *Journal of Machine Learning Research* 9: 1147–78.

Vehtari, Aki, and Janne Ojanen. 2012. "A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison." *Statistics Surveys* 6: 142–228.

Wang, Wei, and Andrew Gelman. 2014. "Difficulty of Selecting Among Multilevel Models Using Predictive Accuracy." *Statistics and Its Interface* 7: 1–8.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting.*: Forthcoming.