Scoring Model Predictions using Cross-Validation*

Anna L. Smith Tian Zheng Andrew Gelman

Department of Statistics
Columbia University

Draft of October 26, 2018

Abstract

We formalize a framework for quantitatively assessing agreement between two datasets that are assumed to come from two distinct data generating mechanisms. We propose a methodology for prediction scoring which provides a measure of the distance between two unobserved data generating mechanisms (DGMs), along the dimension of a particular model. The cross-validated scores can be used to evaluate preregistered hypotheses and to perform model validation in the face of complex statistical models. Using human behavior data from the Next Generation Social Science (NGS2) program, we demonstrate that prediction scores can be used as model assessment tools and that they can reveal insights based on data collected from different populations and across different settings. Our proposed cross-validated prediction scores are capable of quantifying true differences between data generating mechanisms, allow for the validation and assessment of complex models, and serve as valuable tools for reproducible research.

Keywords: *complex models*; *cross-validation*; *model assessment*; *preregistration*; *reproducibility*.

1 Introduction

To begin, we provide a description of the motivation for our proposed methodology, stemming first from recent recommendations for more reproducible research and second from the question of how to appropriately validate complex models. We provide a working definition for data generating mechanisms, which we rely on throughout the paper, and

*This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) agreement number D17AC00001. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

differentiate between predictive and inferential DGMs. Finally, we provide a summary of existing methods for tackling these two issues.

1.1 Reproducible research

Open Science Collaboration (2015)'s attempt to replicate one hundred high-impact psychological studies marked the real beginning of the scientific community's latest struggle, the "replication crisis", although many had commented on similar phenomena previously. In essence, this article highlighted the unfortunate fact that many scientific studies fail to replicate in practice.

As a very simple example, suppose researchers at University A identify a positive effect for some new drug or treatment (e.g., they find evidence to reject a null hypothesis of no effect). Later, researchers at University B replicate University A's study – closely following University A's descriptions of subject recruitment (perhaps even increasing the sample size), experimental procedure, and analysis – but fail to identify a positive effect or even find the opposite, a negative effect. In cases like this one, we are faced with two separate studies that are concerned with the same scientific hypothesis, using the same experiments, but somehow reach different conclusions. This phenomenon has serious implications for the scientific community at large and for the general public; it impacts our ability to trust any single particular finding and results in the watering down of the credibility of science at large. Not surprisingly, this problem is not confined to one particular research domain (although much initial discussion focused on studies in psychology), but is a concern to scientific researchers in all fields.

Closely related to replicability is the issue of reproducibility. A study or experiment (including its specific analytic procedure) is said to be replicable if when the study is repeated, with fresh data or subjects, similar results are achieved. An analysis is said to be reproducible if when the same data is analyzed again, identical results are achieved.

Ideally, published scientific results should be the product of studies or experiments whose results can be independently replicated. This typically requires that the identified effect sizes are relatively large and can be accurately measured, samples sizes are relatively large, and the entire data collection and experimental procedures are well documented, as well as the results of the reproducible analyses (e.g., analytic procedures are well documented and explained and all code and data are made publicly available, if possible). In practice, ensuring replicability and reproducibility is often not straightforward. In fact, since the identification of this replication crisis, it has become clear that no simple solution exists. Instead, advances in this area will necessitate concerted efforts for methodological improvement and more research on reproducibility across many fields. Recently, a variety of new initiatives in this vein have been proposed. For example, some qualitative recommendations can be found in Spies (2018) and Stodden et al. (2016) provide suggestions for computational methods.

One recent recommendation involves the registration of hypothesis tests and scientific analyses prior to data collection (Humphreys et al., 2013; Gelman, 2013), so that one can

avoid the "garden of forking paths," a description offered by Gelman and Loken (2014) for the analytic pipeline in which decisions on data coding, exclusion, and analysis are made contingent on the data, thus inducing problems of multiple comparisons even if only one analysis is done on the particular data at hand. Ideally, prior to collecting any data whatsoever, researchers first prepare a preregistered plan of all data collection methods, modeling and analytic procedures, hypotheses, as well as plans for handling any unexpected deviations from this regime. This preregistration is made publicly available in some way, so that the researchers are held accountable to their preregistered plan. Just as simple random sampling is a prerequisite for classical interpretations of sampling probabilities, standard errors and point estimates, preregistration ensures that the classical interpretation of hypothesis tests and the resulting p-values is appropriate. It is worth pointing out that this sort of preregistration does not preclude further exploratory analyses; The point of preregistration is not to restrict analyses but rather to provide more structure to analyses that are already planned. For example, after data collection, a researcher may notice a pattern or posit a new explanation that motivates additional analyses. Such additional exploratory data analysis (beyond preregistered plans) are generally desirable as they can lead to new discoveries or hypotheses and even inspire additional confirmatory research. In some fields, researchers have the option of submitting such a preregistration to a scientific journal whom, if the submission is accepted, will agree to publish the research prior to any data collection or results. Such a manuscript is called a registered report and usually undergoes a round of peer review prior to the journal's agreement to publish. Not only do registered reports encourage reproducible research, but they also help journals avoid the negative impacts of publication bias.

1.2 Assessing preregistered hypotheses

One requirement of these reports (both in the case of *registered reports* or *preregistrations*) is that the researchers make predictions about the scientific hypotheses to be assessed and models to be fit once data collection is complete. Once the analysis is complete, the researchers are faced with a natural question: *how well do the preregistered predictions align with the observed data?* This question requires a methodology to *score* the predictions, in the face of the materialized observations, usually through some form of "prediction scoring." Further, prediction scoring represents one of the few *quantitative* recommendations for improving the reproducibility of scientific research.

The form of such predictions will largely impact the types of scientific insights one can gain, as well as impact the procedure for prediction scoring. We will discuss such impacts further in Section 3. For now, consider the case where the researcher is able to make predictions about the resulting data at the *observation-level* (i.e., rather than at some higher *summary* or *model estimate* level). In practice, such predictions could come in the form of data from a previously conducted closely related study, as pilot data, or from simulated data that represents *a priori* knowledge on the true data generating mechanism. We may then think of prediction scoring as a measure of the agreement or distance between the *assumed* data generating mechanism (DGM) behind the preregistered predictions and

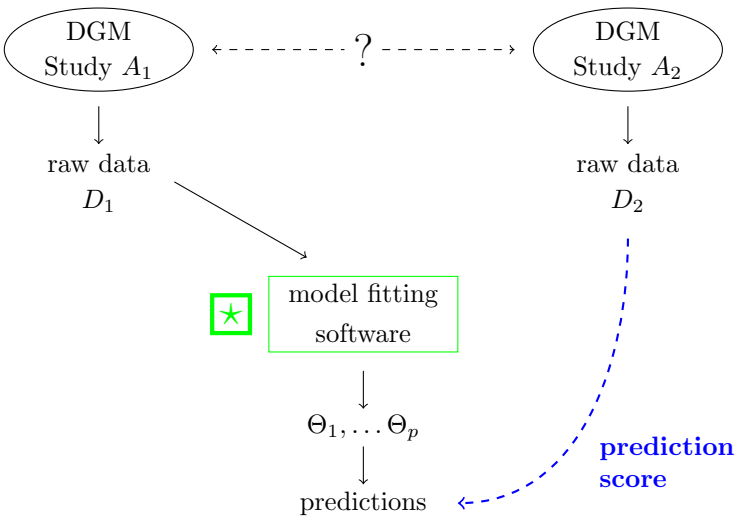


Figure 1: Prediction scores measure the agreement between predictions and realized data. In the case of preregistered predictions, Study A_1 represents a set of pilot data, or data from an existing study, or simulated data.

the *true* data generating mechanism (DGM) behind the realized observations, as in Figure 1.

In this sense, the replication crisis and the related push for more reproducible research motivate methodologies that can appropriately analyze and interpret scientific hypotheses or models across different settings. When performing a preregistered study, how can we quantitatively evaluate differences between the preregistered hypotheses or predictions and the observed data? More generally, when we have access to data that comes from different settings of the same experimental framework (i.e., can be viewed as replications of each other), can we quantify and evaluate differences across these settings? The first step we propose is to view each of these (the set of predictions and the observed data) as realizations from two distinct data generating mechanisms that describe the experimental framework—the preregistered hypotheses or predictions follow from a DGM based on our prior knowledge or pilot data and the observed experimental data follows from the true observed DGM. In this sense, the evaluation of the preregistered hypotheses comes down to identifying differences between the two DGMs.

1.3 Complex data generating mechanisms

In our approach, we will use the term "data generating mechanism" to refer to a particular member of a family of probability distributions or equations that represent a set of (model) assumptions. In this sense, we can use subsets of this family to represent beliefs about the data generating mechanism that gave rise to the observed data, e.g., null and alternative hypotheses. For example, suppose we believe two sets of experimental data are exponentially distributed and we want to test hypotheses about the expected value of these two

distributions, represented by the parameter λ. The DGM family is a set of distributions for λ and, for example, two members of this family that we might be interested in evaluating are $p_1(\lambda \leq 5)$, representing a null hypothesis, and $p_2(\lambda > 5)$, representing an alternative hypothesis. Alternatively, we can think of the DGM family as a unifying experimental framework, such that experiments within this framework can be viewed as replications of each other. Continuing our previous example, perhaps two groups of researchers each ran an experiment where they recorded the waiting time between participants' incoming calls on their personal phones and we want to compare the average waiting time across these experiments.

In practice, we can categorize data generating mechanisms as falling within one of two distinct forms:

1. *predictive*, where the data generating mechanism can be represented by a conditional probability distribution, $p(y|x)$, or

2. *inferential*, where the data generating mechanism can be represented by a distribution for a parameter or set of parameters, $p(\theta)$.

For example, in regression-style analyses, we are most interested in the conditional distribution of some response, given fixed covariate or predictor values. Both parametric and non-parametric regression-style models and other forms of predictive analysis which focus on sampling values (from a probability distribution) for a response or outcome variable, y, given fixed covariate or predictor values, x, fall in the first case. On the other hand, many analyses focus on a particular marginal distribution, $p(\theta)$. For example, we may use linear regression to model some phenomenon but are only interested in one particular slope parameter (i.e., the effect size of one particular predictor). In the example with phone call waiting times mentioned above, we have conceptualized the data generating mechanism as a set of distributions for a model parameter, λ. However, for many complex models or datasets with highly nuanced features, choosing a single parameter or summary statistic upon which to base evaluation of differences between DGMs is not straightforward. In these cases, relying on a single summary measure to capture all relevant (unknown) differences between DGMs has the potential to oversimplify true differences or, in worst cases, fail to detect true differences altogether. For this reason, the prediction scoring methodology which we propose below is focused primarily on scoring differences between predictive DGMs.

1.4 Problem formulation

Under these settings, the natural *prediction scoring* question is translated to the following: are these experiments or realizations products of the same DGM or are they distinct in some way? While this is certainly a natural question, it is ill-posed for most experimental social science research settings. In almost all cases, the data generating mechanisms do in fact vary across experiments or settings, even if only slightly. Instead, we will focus on answering the following: How much do the data generating mechanisms differ across

settings (e.g., from preregistration to observed data) in a quantifiable way? Ultimately, we'd also like to be able to compare different ways that these differences across experimental frameworks (i.e., across experimental operationalizations of related scientific theories, ideas, or hypotheses) can be quantified.

In practice, just like any other statistical method based on sampled data, the observed differences between any two data generating mechanisms have two sources. The first is *by chance*, i.e., random variation in the data such as different samples of participants with different covariates and behaviors, or random variation in model estimates due to stochastic modeling algorithms. The second source is the *true differences* between the two DGMs. The latter is what we care to infer to derive scientific insight. Any methodology to compare these data generating mechanisms based on observed differences needs to be able to set apart differences due to these two sources. We will return to this point in Section 3.

Our proposed approach is to treat prediction scores as an *instrument* for quantifying the difference between our beliefs about the scientific process under study and reality, along a clearly specified "dimension". With this language, we hope to evoke the type of instruments used by social scientists, where a variable is constructed to measure something abstract or unobserved in some sense and carries the same meaning under a general setting. For example, to measure a subject's level of extraversion (which cannot be measured directly), a social scientist might design a survey that includes items that capture behavior indicative of extraversion. In a similar sense, we treat the true scientific process or data generating mechanism (DGM) as unobservable. This is a natural assumption, which suggests the construction of a numerical measure for the distance we are interested in (between our beliefs about that DGM and the true DGM) is not trivial and requires clear definitions and rigorous evaluation.

As shown in Figure 1, prediction scores shall directly measure the distance or difference, via a predictive model, between preregistered predictions and observations from experimental data. The preregistered predictions are conditioned on our beliefs about the DGM through the identification of pilot or sample data, and, often less discussed, through the choice of a particular model. The model and the form of the prediction jointly define an aspect of the DGM for which the prior belief will be evaluated against the reality using our proposed prediction scores. In this sense, prediction scores can be thought of as capturing two levels of scientific insight: (1) how well the predictions match the materialized observations and (2) how well our belief agrees with the reality in terms of the data generating mechanism.

2 Background

As best we can tell, current practice for prediction scoring in registered reports generally consists of making predictions in the form of directional hypotheses (in some cases, predictions for the relative effect size are also included) for model parameters or summary statistics and assessing these predictions by performing a corresponding hypothesis test and checking for a significant effect (for some examples of published registered reports, see

the Zotero library maintained by the Center for Open Science: Mellor, 2018). Our proposed methodology will advance these methods in two main directions. First, it is general enough to accommodate parameter-level hypotheses (or other, higher-level summary statistics) as well as predictions at the individual-level or lowest level of analysis. Individual-level predictions will allow for a more fine-grained assessment of the agreement between our prior beliefs and the true underlying DGM. We will also strongly encourage that these predictions incorporate appropriate measures of uncertainty, such as in the form of probabilistic forecasts. Second, our methodology will provide prediction scores on a continuous scale, which can be viewed as estimates of the distance between our prior beliefs and reality. Thus, we provide a quantitative measure of prediction performance rather than the simple binary detection of a significant effect.

In order to provide these advances, we pull ideas and insights from related research in the statistical literature; Below, we briefly describe statistical methodology for the evaluation of probabilistic forecasts, Bayesian software-checking procedures, and approximate cross-validation (a more thorough discussion is available in the Supplementary Materials).

2.1 Diagnostic plots for probabilistic forecasts and scoring rules

In the statistical literature, perhaps the most applicable line of research to inform prediction scoring for preregistered hypotheses is the evaluation of probabilistic forecasts and the theoretical development of *scoring rules*. Gneiting and Katzfuss (2014) provide a nice summary of recent research in this area. First, let us point out that a scoring function measures the agreement between a point prediction and an observation while a scoring rule measures the agreement between a probabilistic forecast (a predictive probability distribution over future quantities or events of interest, such as a posterior predictive density from a Bayesian analysis) and an observation. Naturally, a probabilistic forecast contains much more information than a simple point prediction and, most importantly, provides a suitable measure of the uncertainty associated with the predictions. For this reason, we will focus on probabilistic forecasts (for a review of issues with point forecasts and scoring rules, see Gneiting, 2011). The importance of probabilistic forecasts as a tool for statistical inference is well-motivated by Dawid (1984)'s framework for prequential analysis, which frames the creation of sequential probability forecasts (over time) as the true focus and underlying motivation for classical statistical concepts and theory. Of course, with the advent of rapidly increasing computational power, MCMC and other estimation techniques have greatly increased analysts' ability to create probabilistic forecasts. In fact, in the past, much of the literature surrounding the evaluation of probabilistic forecasts came out of weather forecasting research. Currently, probabilistic forecasts have been used in applications ranging from climate models, flood risk, seismic hazards, renewable energy availability, economic and financial risk management, election outcomes, demographic and epidemiological projections, health care management, and preventive medicine.

Gneiting et al. (2004) and Gneiting et al. (2007) define two important characteristics of probabilistic forecasts: sharpness and calibration. In this context, sharpness is (solely) a property of the predictive distribution and refers to the concentration of the distribution.

For a real-valued variable, we could measure the sharpness of the probabilistic forecast by considering the average width of prediction intervals. On the other hand, calibration is a property of both the predictive distribution and the materialized observations or events. A probabilistic forecast is calibrated if the distributional forecast is statistically consistent with the observations; In other words, the observations should be indistinguishable from random draws from the predictive distribution. Gneiting et al. (2007) outline various levels of calibration—probabilistic, exceedance, and marginal (listed here from most to least strict)—as well as provide diagnostic tools for identifying these properties in practice. It should be noted that calibration is defined in terms of asymptotic consistency between random variable representations for the probabilistic forecast and the true underlying distribution for the observations (i.e., F is a CDF-valued random variable representing the probabilistic forecast and G is a CDF-valued random variable representing the true data generating mechanism). Thus, in practice, these random variables are themselves unobservable and diagnostic approaches using sample versions (using empirical CDFs) are necessary to assess the calibration of a particular forecast.

To check for probabilistic calibration, histograms (or empirical CDFs, if the sample size is small) of the PIT (probability integral transform) values can be verified for uniformity (this idea can be traced as far back as Rosenblatt, 1952; Pearson, 1933, and perhaps earlier). In meteorological research, Talagrand et al. (1997) proposed a verification rank histogram or Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997) to assess the calibration of ensemble forecasts and Shephard (1994, page 129) has used a similar diagram to sess samples from an MCMC algorithm. However, in the introduction of their paper, Gneiting et al. (2007) demonstrate that merely checking for the uniformity of PIT values is insufficient for distinguishing the ideal forecaster from three (poorer) competitor forecasts. Instead of relying solely on the PIT diagnostic, the authors highlight additional diagnostics (described below) and advocate maximizing the sharpness of the predictive distribution, subject to calibration, as mentioned above. To check for marginal calibration, Gneiting et al. (2007) suggest plotting differences between the average predictive CDF and the empirical CDF for the observations versus x. If the probabilistic forecast is marginally calibrated, we would expect to see only minor fluctuations about zero. Exceedance calibration does not allow for an obvious sample analogue.

Additionally, scoring rules allow us to assess calibration and sharpness simultaneously. Taking a decision theoretic perspective, we can think of a scoring rule as a loss function. In this sense, we can interpret the scores as penalties that the forecaster wishes to minimize. In terms of the choice of a particular form for a scoring rule, one natural restriction is that the truth or true forecast should receive an optimal score. This is precisely what is meant by proper scoring rules (some examples include the logarithmic score, the quadratic score, the spherical score, the continuous ranked probability score, and the Brier score). In fact, Gneiting and Raftery (2007) point out that the log Bayes factor is equivalent to a logarithmic scoring rule in the no-parameter case (i.e. forecasts do not depend on parameters to be estimated from the data). This implies that the log Bayes factor can be used to compare competing forecasting rules, and not only to compare models. When the forecasting rules are specified only up to unknown parameters which will be estimated from

the data, the authors outline a variation of cross-validation that could be used to replace the logarithmic score with other proper scoring rules, to estimate a predictive Bayes factor of some kind. While there are some connections to Bayesian methods, the literature on scoring rules and the evaluation of probabilistic forecasts generally assumes a frequentist or classical perspective. While the discussion is typically focused on predictions for continuous variables, Czado et al. (2009) provide extensions of many of these ideas for count variables.

This literature provides a sound framework for comparing probabilistic forecasts or predictions (such as from preregistration materials) to observed data, where each competing forecast could correspond to different modelling choices or assumptions. The diagnostic tools and recommendations for scoring rules outlined above allow this comparison to be nonparametric and thus, enable the comparison of non-nested, highly diverse models. However, each of these diagnostic measures is necessarily model-based in that any diagnostic plot or set of scoring rules depends on the model assumptions used to create the probabilistic forecast. This complicates the interpretation of the scores or diagnostics themselves, as they measure not only differences between our prior beliefs and the realized data (i.e. between the preregistered predictions and observations) but also any differences between the modelling choices and the true underlying data generating mechanisms. We will propose a prediction scoring framework that uses cross-validation to remove the dependence on model-based differences which enables us to quantitatively measure true differences between our prior beliefs and the realized data.

2.2 Bayesian software-checking

Although perhaps not obvious at first glance, recent proposals for algorithm-checking of Bayesian model fitting software (Cook et al., 2006; Talts et al., 2018) can also provide interesting insights in the prediction scoring setting. These proposals recommend simulating fake data conditional on random draws from the prior distribution, running the model fitting software to obtain draws from the posterior distribution, and using a summary measure to diagnose the alignment between the draws from the posterior distribution and the random draws from the prior distribution. Based on the self-consistency property of the marginal posterior and the prior distribution, these draws should be indistinguishable from one another. To diagnose this alignment, Cook et al. (2006) suggest computing empirical quantiles, comparing the random draw form the prior distribution to the posterior distribution based on that particular draw. The authors suggest looking at histograms of these quantiles, demonstrating that if the software is working correctly, the quantiles should be approximately uniformly distributed. Talts et al. (2018) point out that the empirical quantiles are necessarily discrete and that artifacts of this discretization can lead to misleading diagnostic quantile histograms. Instead, the authors suggest computing rank statistics which will follow a discrete uniform distribution, if the software is correct. Additionally, Talts et al. (2018) provide a nice summary of the types of expected deviations from uniformity that one might observe in the diagnostic histograms with corresponding explanations of modelling choices or software errors that could lead to such deviations.

In terms of the prediction scoring setting, we can think of this software-checking

methodology as a special case where the chosen modelling strategy matches the underlying DGM exactly. We will borrow ideas from this methodology, such as the use of empirical quantiles and rank statistics and the self-consistency properties, to motivate our proposed prediction scoring framework.

2.3 Bayesian model selection and approximate cross-validation

As briefly mentioned previously, our proposed prediction scoring framework will utilize cross-validation to separate true DGM differences from purely model-based differences. Cross-validation, particularly for Bayesian analyses, has been a very active research area in recent years. First, we should point that many Bayesian model comparison summary statistics (such as AIC, DIC, WAIC) can be motivated by the estimation of out-of-sample predictive accuracy (see Vehtari et al., 2012, for a thorough review, from a formal decision theoretic perspective), which of course is one of the goals of cross-validation as well. Gelman et al. (2014) provide a nice review of these model comparison summary measures. As opposed to exact leave-one-out cross-validation (LOOCV), each of the Bayesian model summary statistics utilize the full predictive density and perform an adjustment (e.g., importance sampling, or division by an appropriate variance) to remove the effect of over-fitting, since no data was actually held out. The authors conclude the paper by citing cross-validation as their preferred method for model comparison, despite its high computational cost and requirement that data can be easily partitioned (i.e., partitioning is often not straight forward for dependent data). In this line of thought, Vehtari et al. (2017) develop an approximate version of leave-one-out cross-validation which implements Pareto-smoothing of the importance sampling weights to improve robustness to weak priors or influential observations. Li et al. (2016) develop a version of cross-validation that can be applied to models with latent variables, which relies on an integrated predictive density. In application with competing probabilistic forecasts, Held et al. (2010) compare software fitting algorithms using approximate cross-validation and many of the diagnostic plots mentioned by Gneiting et al. (2007). Finally, Wang and Gelman (2014) and Millar (2018) address the problem of appropriate data partitioning and out-of-sample prediction error estimation for multilevel or hierarchical model selection using cross-validation and predictive accuracy. Wang and Gelman (2014) highlight the fact that model selection can be largely based on the size and structure of the hierarchical data.

This line of research, and its proposed improvements and extensions of cross-validation in various Bayesian settings, can certainly be incorporated in the prediction scoring methodology that we propose. Our contribution will be to expand this literature, from the perspective of the registered reports setting as well as from the unique perspective offered by the set of NGS2 experiments (described in greater below). We formalize the use of cross-validation to appropriately adjust agreement measures between preregistered predictions and realized observations. In other words, we will recommend a unique combination of cross-validation *and* external validation to provide meaningful prediction scores and to enable nonparametric model assessment. Further, in the application to NGS2, we will demonstrate how these cross-validated prediction scores can be used to assess scientific hypotheses across distinct

experiments and data in a nonparametric way.

3 Cross-validated prediction scoring

In this section, we provide a general framework for our proposed prediction scoring methodology. Our goal is to formalize the problem and provide concrete procedures that are general enough to be applicable to a variety of statistical models and analytic procedures.

3.1 General framework

For any family of data generating mechanisms, we will be interested in estimating the distance between different members of the same family. The assumption of a meaningful distance between DGMs is an essential element of this methodology; in order to make quantitative comparisons between DGMs, or between experimental settings, or between preregistered and confirmatory hypotheses, we need to define a distance between DGMs.

Definition 3.1. For a particular family of data generating mechanisms, let the distance between any two members of the family be given by

$$\Delta_{DGM} = f\left(p_i, p_j\right)$$

where p_i and p_j are the ith and jth members of the particular DGM family and the choice of the function f is motivated by the form of the DGM family.

Specifying the form of this distance is not straightforward. For example, consider the case where we are interested in measuring the distance between two straight lines in a two-dimensional Euclidean space. Candidate measures might include calculating the difference in the slope or calculating the Euclidean distance within some window. Each of these measures is a sensible candidate but could result in wildly different conclusions. The issue of choosing an appropriate distance metric is not unique to the example of lines in Euclidean space; a variety of candidate measures exist for assessing the distance or disagreement between sets, or network objects, or points in space, or shapes, etc. Instead, we argue that the form of this distance in the prediction scoring framework should be motivated by the form of the data generating mechanism family. For example, for predictive data generating mechanisms, we might consider conditional KL-divergence (also called relative conditional entropy), whereas for the inferential case, L_p distance is a more natural metric. Recall from Section 1, we want to move away from the simple binary question of disagreement across DGMs (i.e., are the two DGMs different?) and instead promote the quantification of a distance between them (i.e., how far apart are the two DGMs?).

As mentioned previously, we will treat the prediction scores as an instrument for estimating this unobservable distance between data generating mechanisms. In essence, the prediction scores compare the difference between model-based predictions and real-world observations, and in many ways, can be viewed as a validation procedure. Traditionally, model validation is used to assess the predictive ability of the model. In this setting,

Algorithm 1 Prediction Scoring for Predictive Inference

procedure CROSS-VALIDATION
  for $k = 1, \dots K$ **do**
    $x_{1,-k} \leftarrow$ dataset x_1 with kth observation(s) removed
    $\hat{\theta}|x_{1,-k}, y_{1,-k} \leftarrow$ estimate using model fitting software, p_*
    $\hat{y_{1k}} \leftarrow$ prediction, given $x_{1,-k}, \hat{\theta}|x_{1,-k}, y_{1,-k}$
    $q_{1k} \leftarrow g\left(\hat{y}_{1k}, y_{1k}\right)$
procedure VALIDATION
  for $k = 1, \dots K$ **do**
    $\hat{\theta}|x_1, y_1 \leftarrow$ estimate using model fitting software, p_*
    $\hat{y_{2k}} \leftarrow$ prediction, given $x_2, \hat{\theta}|x_1, y_1$
    $q_{2k} \leftarrow g\left(\hat{y}_{2k}, y_{2k}\right)$
procedure PREDICTION SCORING
  $\Delta_{pred} \leftarrow h\left(q_1, q_2\right)$

we are less interested in the fit of any particular model and more interested in learning about potential differences in the data generating mechanism(s) across experiments or settings. Most importantly, note that validation captures differences due to *both* random noise and true differences. Instead of relying solely on validation measures, we propose using cross-validation to properly calibrate the measurements from validation (see Algorithm 1 and Figure 2 for a description of our proposed methodology). In this way, we can separate the differences due to random variation (as measured by cross-validation) from any true differences between the the data generating mechanisms. Further, note that any decisions or conclusions based on validation or cross-validation results alone include the assumption that the researcher's chosen model is correct. In this sense, any observed (apparent) differences between the data generating mechanisms could be due solely to an inadequate model. Instead, comparing results across validation and cross-validation avoids this issue. Because both routines rely on the same model fitting software, comparisons across these routines should be less sensitive to poor modelling choices. In this sense, we are using cross-validation to calibrate the results of the validation procedure.

For DGMs belonging to the same family, let x_1, x_2 represent datasets corresponding to DGMs one and two, respectively. Let z represent the quantity of inference and p_* be the model fitting software, described by model parameters, θ. As described in Algorithm 1 and Figure 2, the prediction scores are calculated as a difference between the distribution of prediction (dis)agreement measures across cross-validation and validation. For both cross-validation and validation procedures, we can define prediction (dis)agreement statistics as follows:

$$q_{jk} = g\left(z_{jk}, \hat{\boldsymbol{z}}_{jk}\right)$$

where z_{jk} is the kth observation (or set of observations) for the jth dataset, $\hat{\boldsymbol{z}}_{jk}$ is a set of predictions for this observation(s), and g is the (dis)agreement measure. For cross-validation, $\hat{\boldsymbol{z}}_{jk}$ is estimated from a model that uses the jth dataset with kth observation (or set of observations removed). For validation, $\hat{\boldsymbol{z}}_{jk}$ is estimated from a model that uses

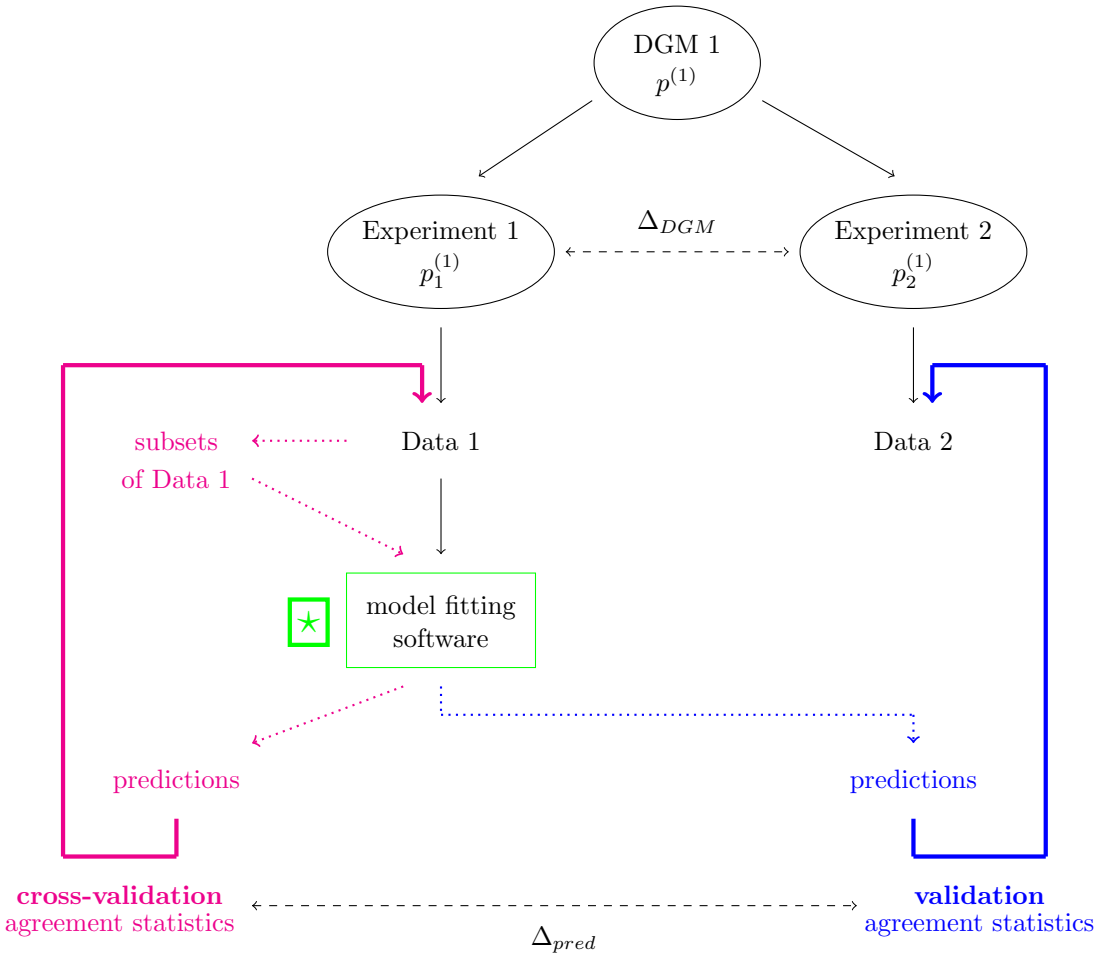


Figure 2: General outline of the proposed prediction scoring methodology for generic data generating mechanisms.

Model	f	g	h
linear regression[1]	conditional KL-divergence	empirical quantiles	KL-divergence
logistic regression[2]	L^p distance		
logistic regression[3]	-	ROC curves	visual inspection
GP model[3]	-	MSE	difference

Table 1: Examples of choices of f, g and h used in this and related work. [1]Section 3.2; [2]Section 3.3; [3]Smith et al. (2018)

the $(j-1)$th dataset, and plugs in any covariates or predictor variables observed in the jth dataset. The choice of g should be motivated by the model fitting software, $p_\star^{(i)}$, chosen by the researcher. For example, when using a linear regression model in focusing on inference for the conditional distribution $p(y|x)$, the predictions will be continuous and so quantiles are a natural choice. However, for logistic regression in the same predictive setting, the predictions will be probabilities (between 0 and 1) while the observations are binary. Some variant of the area under the curve (AUC) statistic would be a better choice for g.

Finally, with these sets of (dis)agreement measures, we can compute the prediction score:

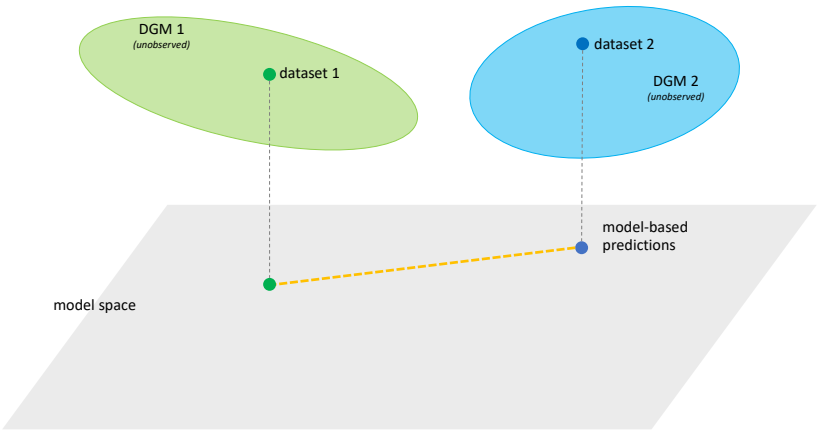
$$\Delta_{pred} = h\left(q_{j_1}, q_{j_2}\right)$$

where $q_{j_1}^{(i)}$ is the vector of cross-validation (dis)agreement statistics and $q_{j_2}^{(i)}$ is the vector of validation (dis)agreement statistics for the ith experimental framework.

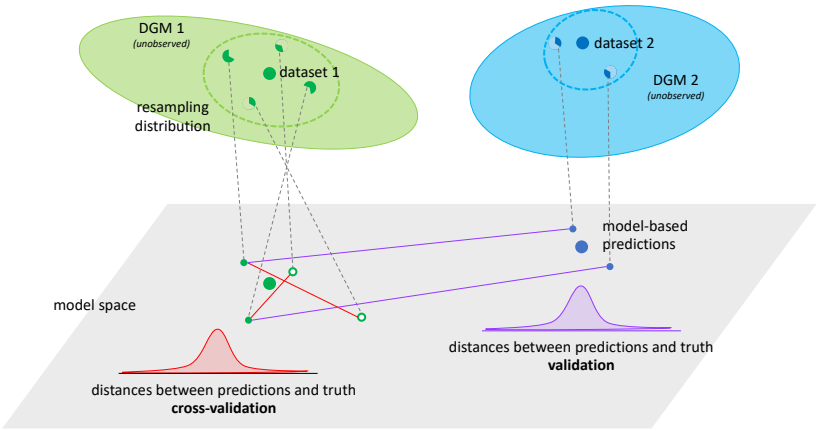
Note that for each particular application, appropriate choices for f (measure of the true difference between the data generating mechanisms), g ((dis)agreement statistic for the cross-validation and validation predictions), and h (measure of the difference between the distributions of (dis)agreement statistics) must be made. As we have suggested above, these choices should be well motivated by the particular application. More specifically, f should be motivated by the form of the family of data generating mechanisms being considered, and g should be motivated by the researcher's model fitting software. Additionally, the choice of h should be motivated by both of these considerations and the subsequent choices for f and g. Although this methodology would be simpler if f, g and h were universally specified, it is important that they appropriately capture the important features of the data generating mechanisms and are suitable to whatever model fitting software is chosen by the researcher (see Table 1 for some specific examples). Further, note that this sort of conditional specification is not unlike the choice of an appropriate link function for generalized linear models. Appropriate forms of f, g, and h may be derived for more complex settings (e.g, dependent data, such as networks or time series) in the future.

3.2 Example: Linear regression

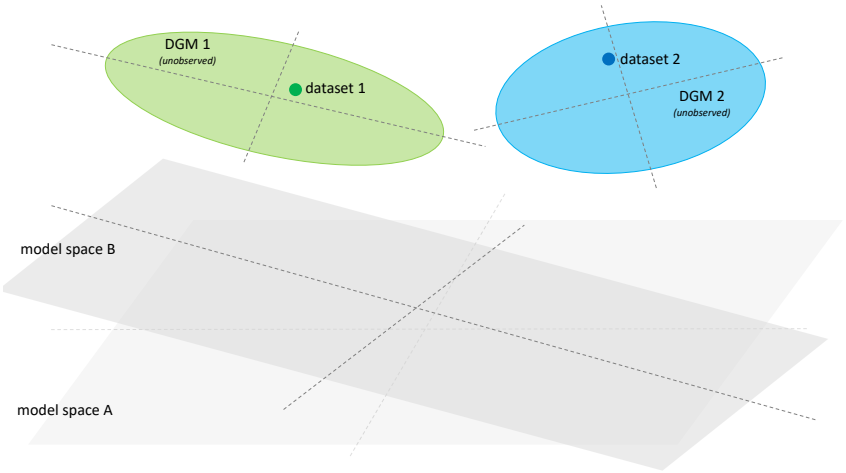
To better understand this methodology, we turn now to an example in the predictive case, a Bayesian linear regression model, documented in Figure 4. In this setting, $p_j^{(i)} = p_j^{(i)}(y|x)$ is the conditional distribution of the outcome or response variable,y, given fixed values of the



(a) Differences between unobserved DGMs can be measured indirectly through model-based predictions.



(b) Cross-validation and validation account for sampling variability and separate model fit issues from true DGM differences.



(c) Prediction scores measure the distance between DGMs along the dimension of the model used to make predictions.

Figure 3: Geometric interpretation of the prediction scores.

predictors or covariates, x. The true difference between the data generating mechanisms is the conditional KL-divergence. In this Bayesian setting, predictions are draws from the posterior predictive distribution. For cross validation, this distribution, $p^{(1)}_{*|-k}$ is conditioned on the set of data (covariates and responses) from experiment 1 with the kth subset removed and provides a prediction for the kth subset of responses, corresponding to the kth set of covariates in experiment 1. For validation, the posterior predictive distribution, $p^{(1)}_{*|1}$ is conditioned on the entire set of data (covariates and responses) from experiment 1 and provides a prediction for the kth response, corresponding to the kth covariate in experiment 2. Finally, we estimate the true difference between the data generating mechanisms by calculating the KL-divergence between the distributions of (dis)agreement statistics across cross-validation, $\boldsymbol{q}^{(1)}_1$, and validation, $\boldsymbol{q}^{(1)}_2$. Motivated by the Bayesian software-checking approaches of Cook et al. (2006) and Talts et al. (2018), a natural choice for the (dis)agreement statistics might be empirical quantiles or rank statistics.

3.3 Example: Logistic regression

To understand how this methodology can be used for inferential DGMs, consider the case where we assume the underlying process follows a simple logistice regression.

4 Probabilistic behavior of prediction scores

To understand how these prediction scores behave in practice and to get a sense of their asymptotic behavior, we have designed a simulation study that utilizes a simplified experimental design and models the outcome of interest with logistic regression. Many aspects of this simulation study were designed to complement related research that examines prediction scores for human behavior data in experimental social science research (Smith et al., 2018). We summarize the set up of this simulation study below and will detail how this study has been extended here to better examine the general probabilistic behavior of our prediction scoring methodology.

In Smith et al. (2018), we consider $K = 5$ settings of a public goods game in which each participant has the opportunity to contribute ("cooperate") or not ("defect") to a set of pooled resources that will be multiplied and shared among all participants. Additionally, we imagine that some percentage of the total number of players, $\pi = \{0, 0.25, 0.50, 0.75, 1\}$, are in fact bot participants whose behavior is strictly specified according to some set of algorithmic rules. The goal of these hypothetical experiments is to understand the ways in which participants' decisions to cooperate are influenced by the presence of bots.

True DGM. Let y_{ijkt} be the decision to cooperate ($y_{ijkt} = 1$) or defect ($y_{ijkt} = 0$) for the ith individual in the jth cohort of the kth experimental setting for round t. Additionally, let z_{ijk} be an indicator of whether the ith participant in the jth cohort of the kth round is a human participant ($z_{ijk} = 1$) or a bot (z_{ijk}). We will assume the true underlying data

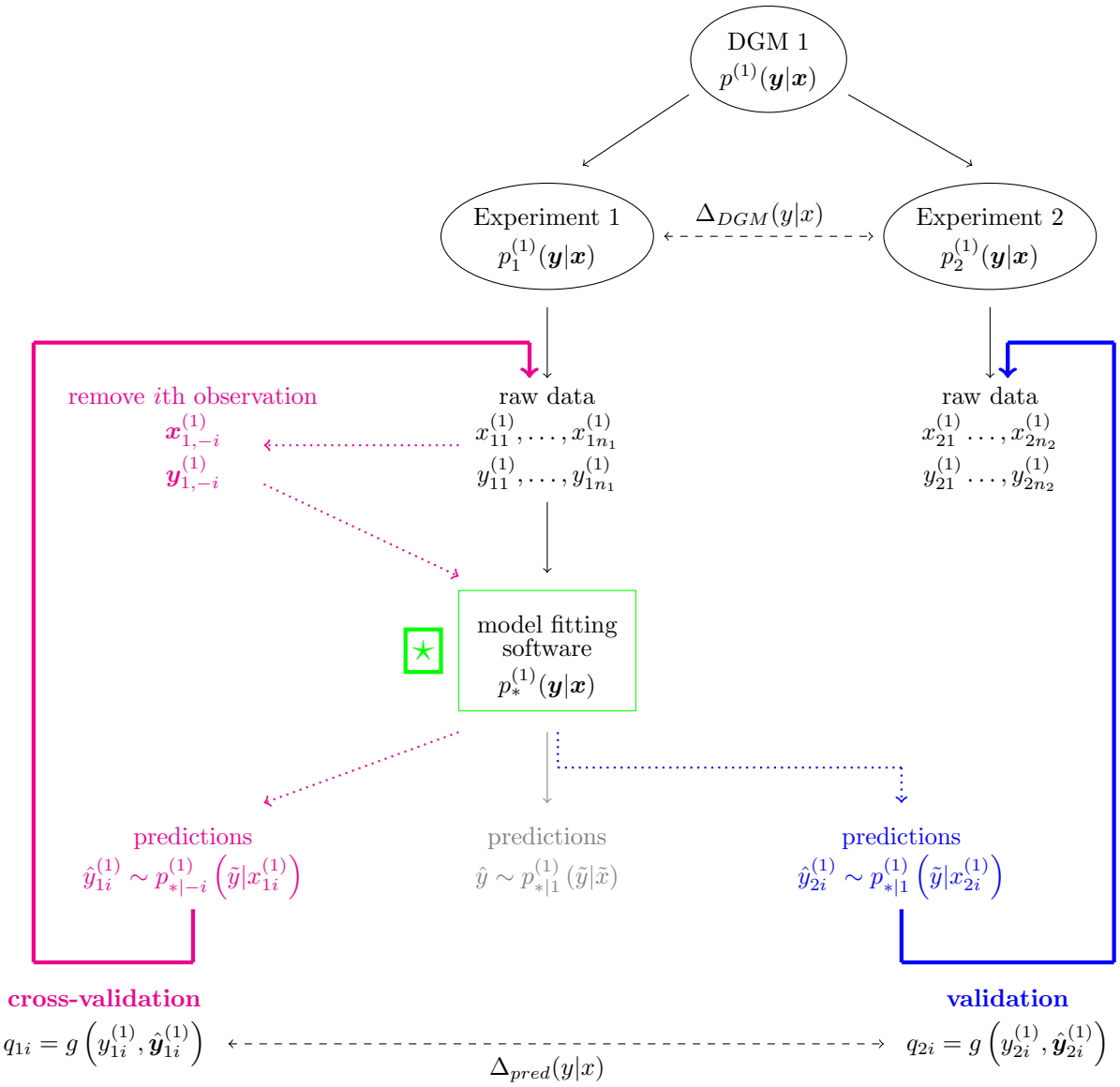


Figure 4: Outline of the procedure for a predictive data generating mechanism, such as linear regression.

generating mechanism is given by the following:

$$
\begin{aligned}
& & z_{ijk} &\overset{iid}{\sim} \text{Bernoulli}(\pi_k) \\
\text{Model 0:} & & \text{logit}^{-1}\left[P(y_{ijkt}=1|z_{ijk}=1)\right] &= \beta_0 + \beta_1 t + \beta_2 y_{ijk,t-1} + \beta_3 \bar{y}_{\cdot jk,t-1} \\
& & \text{logit}^{-1}\left[P(y_{ijkt}=1|z_{ijk}=0)\right] &= \beta_0' + \beta_2' y_{ijk,t-1}
\end{aligned}
$$

where π_k is the percentage of bots in the kth round, β_0 and β_0' are baseline tendencies to cooperate, β_1 captures any trend across the rounds, β_2 and β_2' capture the tendency to switch between behaviors, and β_3 represents the influence of team members' decisions. Values for these parameters for the simulated data are provided and motivated in Smith et al. (2018).

Prediction scoring details. In this setting, the DGMs being compared are predictive conditional distributions which we can refer to by $p(\boldsymbol{y}_k|\boldsymbol{x}_k, \pi_k)$. We perform this analysis in a Bayesian setting, so that predictions are draws from the posterior predictive distribution.

To compare these predictions to the set of true observations, we compute receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) statistics (Davis and Goadrich, 2006). These measures are very popular model fit assessment tools for logistic regression. In order to comptue these measures, we use L-fold cross-validation where L is chosen such that each partition contains roughly 500 observations.

Researcher models. To uncover true differences across the experimental settings, we consider the following three researcher models:

$$
\begin{aligned}
\text{Model 1:} \quad \text{logit}^{-1}\left[P(y_{ijkt}=1)\right] &= \gamma_0 + \gamma_1 t, \\
\text{Model 2:} \quad \text{logit}^{-1}\left[P(y_{ijkt}=1)\right] &= \gamma_0' + \gamma_2 y_{ijk,t-1}, \\
\text{Model 3:} \quad \text{logit}^{-1}\left[P(y_{ijkt}=1)\right] &= \gamma_0'' + \gamma_3 \bar{y}_{\cdot jk,t-1},
\end{aligned}
$$

where γ_0 is a baseline tendency to cooperate, γ_1 can capture some trends across the rounds, γ_2 represents the influence of of the most recent decision, and γ_3 represents the influence of team members' decisions.

Smith et al. (2018) provide interpretations of visual differences in the ROC curves across the different models and experimental settings. To summarize these results, the prediction scores behave as expected; they appear similar when comparing data generated from the same DGM and appear more different as the distance between DGMs (here, measured simply in terms of $|\pi_i - \pi_j|$) increases. This demonstrates that prediction scores can be used to uncover features of the DGM that vary across experimental settings, in a way that properly accounts for sampling variability. Additionally, the results of the simulation study indicate that Model 1 is the most sensitive to differences across the experimental settings. This is well-aligned with boxplots of the cooperation rate by round across each setting. In other words, when the model is aligned with true differences between the data

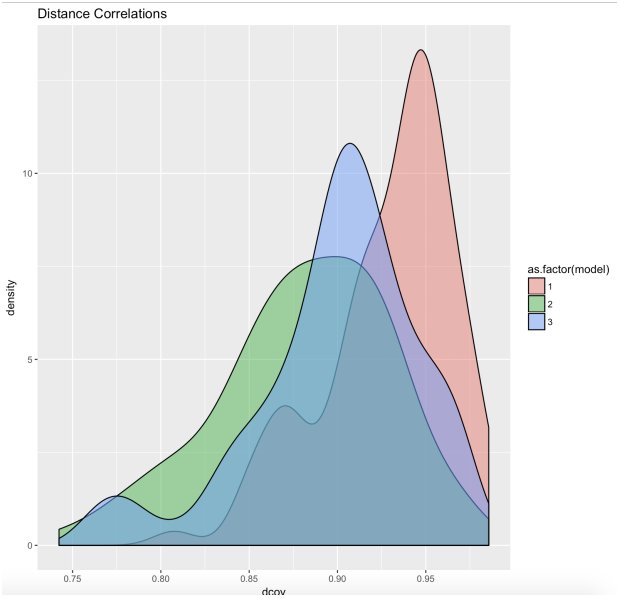


Figure 5: Distance correlations for prediction scores.

generating mechanisms, the distance between the cross-validation and validation statistics reflects the true distance between the DGMs. In practice, relevant data patterns may be much more nuanced (i.e., not obvious from simple summary plots) and the true data generating mechanisms may be much more complex (i.e., it may be much more difficult to specify a model that predicts well).

Extension to study probabilistic behavior . In order to get a sense of how these prediction scores behave asymptotically, we repeat the above simulation study many times and examine the relationship between the true distance between DGMs and our prediction scoring estimates of that distance. This requires defining a true distance between the data generating mechanisms. In this extended simulation, we consider two measures: (1) the difference between the percentage of bots, $|\pi_i - \pi_j|$, and (2) the conditional KL-divergence, calculated as follows:

$$KL(p_1, p_2) =$$

To evaluate whether or not the prediction scoring estimates are well-aligned with these measures of the true underlying distance, we calculate distance covariances (Székely et al., 2007). A distance covariance is a measure of dependence between two paired vectors that is capable of detecting both linear and nonlinear associations. If the vectors are independent, then the distance covariance is zero. We can treat each repetition of the above simulation study (where we compute prediction scores across all possible pairs of π) as a sample which gives rise to a vector of prediction scoring distance estimates. Then we examine the distribution of distance covariances, as a function of the (researcher) model used to make predictions. After repeating this simulation 1000 times, we plot the distance covariances in Figure 5.

5 Network experiments in cooperative games

The experiments proposed by the research teams in the (currently ongoing) Next Generation Social Science (NGS2) program present a great opportunity to evaluate the proposed prediction scoring framework. This program funds multiple research teams over two cycles of experiments and is designed as a methodologically-focused effort to develop a fundamental reimagining of the social science research cycle (Nosek et al., 2018). During each cycle, each research team will conduct distinct experimental social science studies regarding a shared research question. Prior to any data collection, each team will complete preregistration materials, which includes predictions for study outcomes. In the following, we briefly describe the Gallup teams' experiments for the first cycle of the program (for more detailed descriptions of each team's planned and completed research, see the preregistration materials which have been made publicly available on the Open Science Framework Nosek et al., 2018).

In the first cycle of the NGS2 program, the Gallup team provided an excellent application for our proposed prediction scoring methodology since their preregistered materials included pilot data from a previous study which informed their study hypotheses. This allows for an intuitive application of our proposed prediction scoring framework where we can compare the agreement between predictions for experimental data (based on the Gallup team's proposed modeling strategy and their identified pilot data) and the materialized observations from the experiment itself.

Experimental setting The first cycle of the NGS2 program focused on identifying pathways towards the formation of collective identity and cooperative decisions. To address this research question, the Gallup team considered the role of social networks in the development of large-scale cooperation among individuals in an economic game. They used a logistic regression model to examine individuals decisions (cooperation or defection) and showed that social networks which can be frequently updated by participants (rather than fixed throughout the course of the game or randomly updated) foster cooperative decisions in this setting. The Gallup teams experiments were designed to mimic the experiments performed by Rand et al. (2011), and whose data can serve as a set of preregistration data.

Experimenters randomly assigned participants to one of four conditions (see below) in a series of realizations of network experiments. In all conditions, subjects play a repeated cooperative dilemma (each game/session consists of multiple rounds) in an artificial social network created in the virtual laboratory. During each round of the game, each player can choose one of the following two actions: (1) cooperation: donate 50 units per neighbor, resulting in each neighbor actually gaining 100 units and (2) defection: donate nothing, resulting in neighbors getting nothing. After each round, players learn about the decisions of their neighbors and their own payoff. Additionally, the experimenters considered the following possible link-updating regimes for the social network in the game: (1) static or fixed links, (2) random link updating, where the entire network is regenerated at each round, (3) strategic link updating, where a randomly selected actor of a randomly selected pair may change the link status of that pair. The strategic link updating condition was

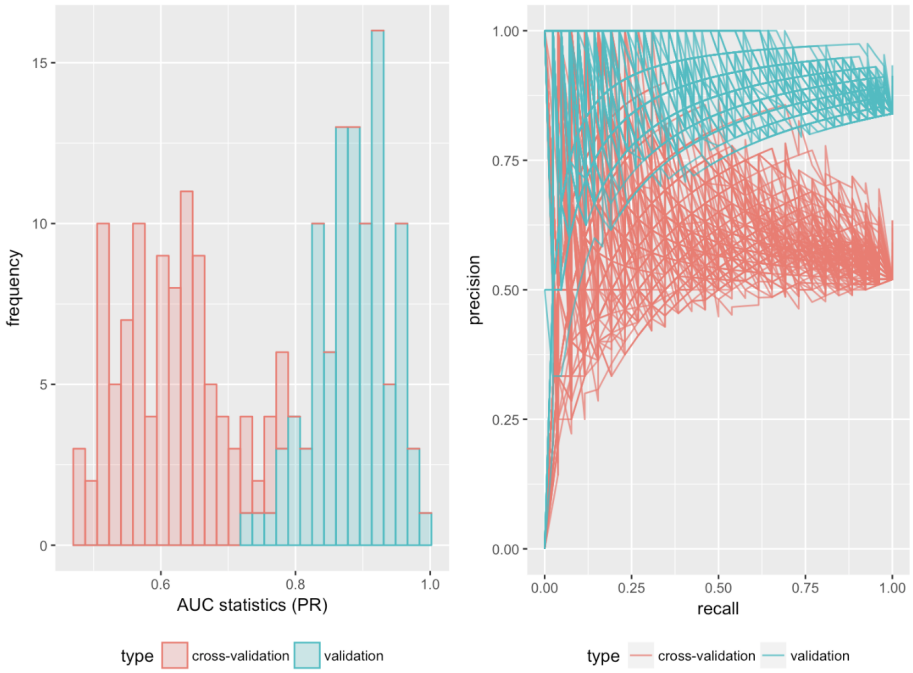


Figure 6: Prediction scores for Gallup's Cycle 1 Hypothesis 1.4: rapidly updating networks support cooperation, relative to all other conditions.

further split into two categories: (a) viscous, where 10% of the subject pairs were selected and (b) fluid, where 30% of the subject pairs are selected.

Prediction scoring Recall, that we have suggested using quantiles or rank statistics as a disagreement statistic in our proposed prediction scoring methodology. However, for logistic regression, observations and predictions will be collections of 0s and 1s. Thus, using quantiles doesnt make sense in this setting. Instead, we can use the ROC (receiver operating characteristic) curve or precision-recall curve and the AUC (area under the curve) statistic to measure the agreement between observations and predictions. These measures are very popular model fit assessment tools for logistice regression. Thus, rather than comparing quantile distributions across cross-validation and validation, we will compare the distribution of AUC statistics across these settings. For this particular dataset, we will calculate the AUC statistic for the precision-recall curve. Generally, the precision-recall curve is preferred over the ROC curve whenever the data is imbalanced (see Davis and Goadrich, 2006, for more discussion). Finally, we need to point out that the AUC statistic is not defined for a single data point. Thus, we can not use leave-one-out cross-validation in our prediction scoring routine. Instead we partition the dataset into k subsets and use k-fold cross-validation, resulting in k AUC statistics. Similarly, when performing validation, we must again partition the data into subsets.

As an example, consider Hypothesis 1.4 from the Gallup team's preregistration materials. They hypothesized that rapidly updating networks would support cooperation more

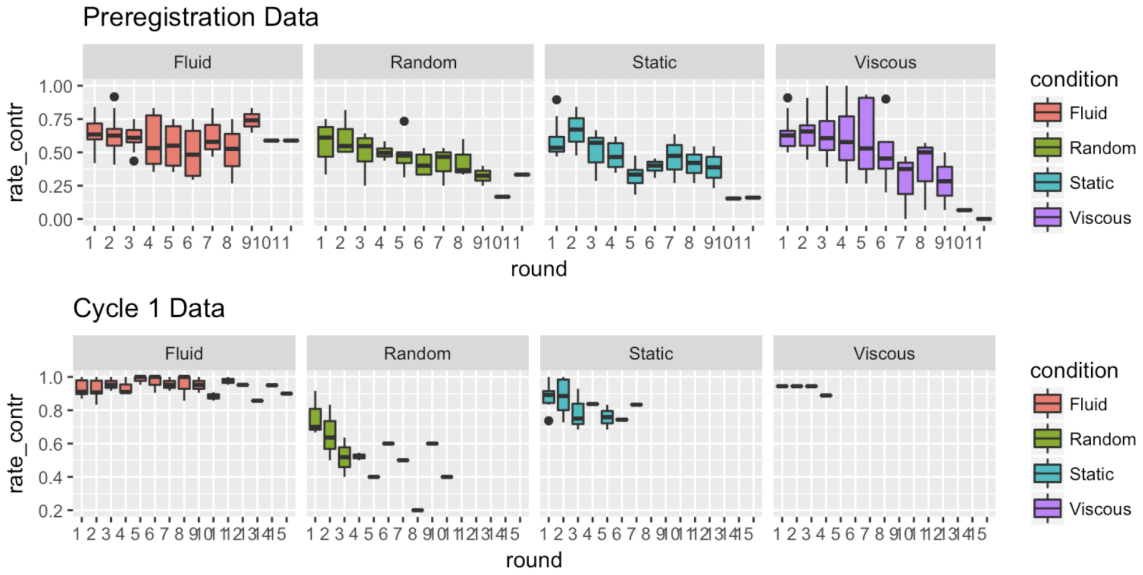


Figure 7: Boxplots of average cooperation levels across rounds of Gallup's Cycle 1 games.

than any other condition. To evaluate the prediction scores, we have compared the distribution of AUC statistics from cross-validation to those from validation as well as plotted the corresponding precision-recall curves from each of the k subroutines of cross-validation and validation (see Figure 6). The validation statistics measure differences due to both random noise and true differences between the DGMs (i.e., between our prior beliefs about the preregistered data and reality), while the cross-validation statistics only capture differences due to random noise or disagreement between the underlying DGM and the chosen model. In this case, we observe larger AUC statistics and better ROC curves in the validation routine. This indicates that there is less variability in individuals' behavior in the experimental data, than in the preregistration data. In a sense, subjects in the Gallup experiment are acting in more predictable ways than the subjects from the previous experiment. And in fact, if we simply examine summary statistics of the in-game decisions themselves, we can see the same type of pattern. In Figure 7, we provide boxplots of individuals' average cooperation levels across rounds of the game, where each color corresponds to a different link-updating experimental condition. Comparing the preregistration data (top row) to the experimental data (bottom row), we see that the boxplots are drastically narrower, indicating that there is less variability in participant behavior.

This application serves as an illustration of how our prediction scoring can enable interesting scientific insights. Further, we demonstrated how this methodology can be adapted to appropriately address modelling choices (i.e., using distributions of AUC statistics, rather than quantiles) and demonstrates the type of diagnostic plots that can be used to interpret the resulting prediction scores.

References

Anderson, J. L. "A method for producing and evaluating probabilistic forecasts from ensemble model integrations." *Journal of Climate*, 9(7):1518–1530 (1996).

Cook, S. R., Gelman, A., and Rubin, D. B. "Validation of software for Bayesian models using posterior quantiles." *Journal of Computational and Graphical Statistics*, 15(3):675–692 (2006).

Czado, C., Gneiting, T., and Held, L. "Predictive model assessment for count data." *Biometrics*, 65(4):1254–1261 (2009).

Davis, J. and Goadrich, M. "The relationship between Precision-Recall and ROC curves." In *Proceedings of the 23rd international conference on Machine learning*, 233–240. ACM (2006).

Dawid, A. P. "Present position and potential developments: Some personal views: Statistical theory: The prequential approach." *Journal of the Royal Statistical Society. Series A (General)*, 278–292 (1984).

Gelman, A. "Preregistration of studies and mock reports." *Political Analysis*, 21(1):40–41 (2013).

Gelman, A., Hwang, J., and Vehtari, A. "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6):997–1016 (2014).

Gelman, A. and Loken, E. "The statistical crisis in science." *American Scientist*, 102:460 – 465 (2014).

Gneiting, T. "Making and evaluating point forecasts." *Journal of the American Statistical Association*, 106(494):746–762 (2011).

Gneiting, T., Balabdaoui, F., and Raftery, A. E. "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268 (2007).

Gneiting, T. and Katzfuss, M. "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1:125–151 (2014).

Gneiting, T., Raftery, A., Balabdaoui, F., and Westveld, A. "Verifying probabilistic forecasts: Calibration and sharpness." In *Preprints, 17th Conf. on Probability and Statistics in the Atmospheric Sciences, Seattle, WA, Amer. Meteor. Soc*, volume 2 (2004).

Gneiting, T. and Raftery, A. E. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477):359–378 (2007).

Hamill, T. M. and Colucci, S. J. "Verification of Eta–RSM short-range ensemble forecasts." *Monthly Weather Review*, 125(6):1312–1327 (1997).

Held, L., Schrödle, B., and Rue, H. "Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA." In *Statistical modelling and regression structures*, 91–110. Springer (2010).

Humphreys, M., Sanchez de la Sierra, R., and Van der Windt, P. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis*, 21(1):1–20 (2013).

Li, L., Qiu, S., Zhang, B., and Feng, C. X. "Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC." *Statistics and Computing*, 26(4):881–897 (2016).

Mellor, D. "Registered reports library." www.zotero.org/groups/479248/osf/items/collectio nKey/KEJP68G9. Center for Open Science (2018).

Millar, R. B. "Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation." *Statistics and Computing*, 28(2):375–385 (2018).

Nosek, B. A., Spitzer, M., Russell, A., Tully, E., Rajtmajer, S., Ahn, S.-H., Zheng, T., Foy, D., Kluch, S. P., Stewart, C., and et al. "NGS2 DARPA Program." (2018).

Open Science Collaboration. "Estimating the reproducibility of psychological science." *Science*, 349(6251):aac4716 (2015).

Pearson, K. "On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random." *Biometrika*, 379–410 (1933).

Rand, D. G., Arbesman, S., and Christakis, N. A. "Dynamic social networks promote cooperation in experiments with humans." *Proceedings of the National Academy of Sciences*, 108(48):19193–19198 (2011).

Rosenblatt, M. "Remarks on a multivariate transformation." *The annals of mathematical statistics*, 23(3):470–472 (1952).

Shephard, N. "Partial non-Gaussian state space." *Biometrika*, 81(1):115–131 (1994).

Smith, A., Zheng, T., and Gelman, A. "Evaluating drivers of human behaviors in experimental social science using prediction scoring." (2018).

Spies, J. R. "Reproducibility Rubric." (2018).

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P., and Taufer, M. "Enhancing reproducibility for computational methods." *Science*, 354(6317):1240–1241 (2016).

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. "Measuring and testing dependence by correlation of distances." *The annals of statistics*, 35(6):2769–2794 (2007).

Talagrand, O., Vautard, R., and Strauss, B. "Evaluation of probabilistic prediction systems, paper presented at ECMWF Workshop on Predictability, Eur. Cent. for Med. Range Weather Forecasts." *Reading, UK* (1997).

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." *arXiv preprint arXiv:1804.06788* (2018).

Vehtari, A., Gelman, A., and Gabry, J. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5):1413–1432 (2017).

Vehtari, A., Ojanen, J., et al. "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6:142–228 (2012).

Wang, W. and Gelman, A. "Difficulty of selecting among multilevel models using predictive accuracy." *Statistics at its Interface*, 7(1):1–88 (2014).