CAMBRIDGE

Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

Part I

The Predictive View

> Prediction was the earliest and more prevalent form of statistical inference. This emphasis changed during the beginning of [the twentieth] century when the mathematical foundations of modern statistics emerged. Important issues such as sampling from well-defined statistical models and clarification between statistics and parameters began to dominate the attention of statisticians. This resulted in a major shift of emphasis to parametric estimation and testing.
>
> The purpose of this book is to correct this emphasis. The principal intent is to revive the primary purpose of statistical endeavor, namely inferring about reasonable values [that were] not observed based on values that were observed.
>
> Preface of Geisser (1993)

The first four chapters of this book are an exposition of the centrality of prediction. For instance, estimation, testing, and classification problems can be cast in predictive terms and, conceptually, the predictive view is as philosophically justified as any other established philosophy of statistics, perhaps more so. Of particular importance is the downgrading of the role of models to settings in which they are genuinely useful, a relatively narrow set of circumstances. The alternative is falling victim to the convenience of models and the feeling of understanding they give even when they are unreliable – which is essentially all the time, outside oversimplified contexts.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

1

Why Prediction?

> ... any model ... is merely a human attempt to describe or explain reality ... models are to be assessed in terms of their success at this task. It is misguided ... to believe in Nature as obeying some theory ... Even if we can find a completely successful theory, this does not mean we have identified Nature's true model – some other, distinct theory might be just as successful ... In this view, theories can only be distinguished by means of their predictions about observables ...
>
> Dawid (1992)

For centuries, perhaps millennia, people have tried all sorts of divination methods, from yarrow sticks to tarot cards, from the innards of animals to the positions of planets. With sufficiently skilled interpreters, these methods probably work a little: a tarot card reader may use the cards to evoke the frame of mind of the subject. To the extent that the future is implicit in a subject's frame of mind, the predictions may therefore be accurate. After all, if you know some one, even a little, you can predict some of their behavior. This is second nature for good salespeople, politicians, and others whose career success depends on detecting people's preferences. Arguably, this sort of procedure might even help with economic predictions that include market psychology. Note, however, that divination methods are rarely used to predict such outcomes as how much product a given chemical reaction will produce, or other outcomes that have essentially no element of human choice.

Here, by contrast, the goal is to make predictions by rules in such a way that evaluating how well the rules work will be unambiguous. The fortunate case occurs when the rules accurately reflect something about the mechanism used by the data generator (DG) to generate outcomes. This is the main goal of much of conventional science. However, there are vast classes of data where it is implausible to model the DG. As a slightly facetious example, one can treat *MacBeth* as a sequence of letters and try to predict the $(n+1)$th letter using the first n letters. A variant on this is predicting the $(n+1)$th nucleotide – or any finite sequence of nucleotides – on a chromosome, given the first n nucleotides in the chromosome. In both cases, the DG is so complex that detailed modeling for the purposes of prediction would be premature, to say the least. Indeed, if we want to make predictions, it's unlikely modeling will help much.

Worse, many DGs might not function by rules at all. The easiest way to think of this is that the outcome y_1 at a given time step is from one distribution, say Q_1, but the outcome y_2 from the next time step is from another distribution, Q_2, chosen by some agent who may not even know Q_1 or y_1 and chooses Q_2 using a hidden mechanism or no mechanism at all.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

Doing this repeatedly means there is nothing stable enough to model, so we cannot, even conceptually, use models to generate prediction rules. Another way to think of this is to ask whether more data can be generated – at least in principle – that would be informationally equivalent to the data we already have. This is not the same as asking whether an experiment is repeatable in practice – many aren't, as for example in econometrics – it only asks whether in principle we could generate further data sets of the same general form. Clearly, if the answer is no (think of *MacBeth*) then there may be no rule that the DG follows and hence it will be impossible to formulate a prediction rule that matches the DG. However, even when we admit that the DG does not follow rules, we may still want a well-defined prediction rule so that we can evaluate how good it is. Crazy as it sounds, this is not entirely impossible, as we shall see.

The stance of this book, stated concisely and unabashedly, is that predictive statistics proposes an alternative formulation of the paradigm statistical problem. The central feature of predictive statistics, as opposed to other schools of thought in statistics, is to use the data to predict ahead rather than try to find out what underlies the data generator, then try to model it, and finally use the model to make predictions. In either case, predictive or not, predictions must be compared with new data for validation. The question is when this comparison is made – is it before or after 'modelling' has been attempted? In predictive statistics, modeling does not start until good prediction has been achieved. This is the reverse of the conventional approach.

One of the key arguments for predictive statistics at this time of writing, and for the foreseeable future, is that so much of the statistical world has changed. Volumes of data have massively increased, preprocessing techniques for raw data (e.g., in the 'omics world) have increased and become more diverse, multitype data is more prevalent than before (and often very hard to model), and the complexity of data streams that confront the Statistician is nearly overwhelming. Together, these features make modeling difficult, if not infeasible. Indeed, often only a small fraction of the available data can be used in an analysis. Taking a predictive approach, and hence achieving good prediction, is likely to be better in the long run than direct modeling for understanding a data generator. Even where modeling is infeasible, good prediction is the sine qua non of a good theory. The reason is that modeling requires dealing with model uncertainty and misspecification and these can be extraordinarily difficult.

1.1 Motivating the Predictive Stance

It may seem strange to ask, but it's important to answer the question 'Why is prediction so important?' First, one obvious answer is that sometimes the goal really is to know what the next outcome is likely to be: prediction may be the goal of the statistical analysis. For instance, it might be helpful to predict who will get post-traumatic stress disorder (PTSD). That way, to prevent PTSD, or minimize its effects, a physician would want to know who it is most important to treat prophylactically. Sometimes the goal is prediction even when it's not phrased predictively. For instance, one may estimate a probability of recurrence of cancer (with a standard error), but it would be more informative to give a point predictor for when a patient will get a recurrence along with an assessment of the variability of that prediction. Aside from being the information that a patient or physician wants, a prediction

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

interval is less abstract and more intelligible than a probability, let alone a confidence or credible interval for a probability. Of course, people might want to predict the weather, the economy, the response to a new treatment, and so forth. Who hasn't wanted to know the future for some purpose, base or laudable?

A second answer (that is less obvious) is that most other goals of statistical analysis can be subsumed within prediction. What, after all, are the main goals of statistical analyses? Any list would have to include (1) model identification, (2) decision making, and (3) answering a question about a population – even if there is some overlap among these goals as stated. For instance, identifying a model may amount to making a decision: in classification, model identification amounts to identifying a classifier, and using a classifier amounts to deciding the rule by which one will assign a future subject to a class. In general, it's hard to find a statistical problem that doesn't have a direct connection with prediction.

Let's start with *model identification* – which is essentially estimation in one guise or another. This includes, among other possibilities, parametric estimation, classification, regression, nonparametric estimation, and model selection. It also includes some hypothesis testing. A simple versus simple test such as $\mathcal{H}_0 \colon P = P_1$ versus $\mathcal{H}_1 \colon P = P_2$, where P_1 and P_2 are the only two candidate probabilities for P, is an obvious example. In general, one can use a series of goodness-of-fit tests to determine models that can't be rejected. Moreover, tests such as whether a given variable should be in a regression function must also be included as part of model identification. Even though such a test does not by itself identify a model, it constitutes an effort to reduce the class of models that must be considered and is therefore a step toward model identification.

In all these cases, how can one know in reality that a model has been successfully identified without using it to generate accurate predictions? Even more, how can one know that another model with an equally good fit, possibly using different variables, can be ignored if it is not predictively discredited – for instance, on the grounds of high bias or high variance? Loosely, outside very simple problems, model identification without predictive verification is little more than conjecture. Put otherwise, whenever a model is selected, or a parameter estimated, a predictor is formed and, if it doesn't perform well, the model it came from is discredited.

Essentially this means that the search for a good model is merely a special case of the search for a good predictor. The substitution is thought worthwhile because a scientist can bring 'modeling information' into the search for a predictor. The problem is that modeling information is usually not itself predictively verified and hence is often of dubious value. Thus, taking a purely predictive view and treating modeling information as likely to be unreliable guards against the use of such suspect 'information'.

Let us now turn this around. Just as a credible model, when it exists, can often be used to generate predictions, a predictor can sometimes be used to identify a model. In the simplest case, it is assumed that there is a parametric family $\mathcal{P} = \{p(\cdot|\theta)|\ \theta \in \Omega\}$, where $\Omega \subset \mathbf{R}^d$ for some integer $d \geq 0$, equipped with a prior on Ω having a density with respect to μ, say. Then, the predictive distribution for a random variable Y_{n+1} with outcomes y_{n+1} given $Y^n = (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n) = y^n$ is

$$m(y_{n+1}|y^n) = \int p(y_{n+1}|\theta) w(\theta|y^n) \mu(\mathrm{d}\theta),$$

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

where $w(\theta|y^n)$ is the posterior density. It is easy to verify that $m(y_{n+1}|y^n)$ is optimal in a relative entropy sense (see Clarke *et al.* (2014)) and that when $Y_i \sim P_{\theta_0}$, with density p_0, is independent and identically distributed (IID) for all i that

$$m(y_{n+1}|y^n) \to p_{\theta_0} \tag{1.1}$$

pointwise in y_{n+1} in distribution, as $n \to \infty$. So, one could fix a distance d on densities for Y_{n+1} and choose a model based on $\theta^* = \theta^*(y^n)$ satisfying

$$p_{\theta^*}(\cdot) = \arg\min_{\theta} d(m(\cdot|y^n), p_\theta(\cdot)).$$

Not all predictors can be so obviously converted into models, just some of the good ones. Moreover, (1.1) only holds under highly restricted conditions.

Decision making can also be subsumed into prediction. Suppose that there is a prior, a parametric family, data, and a loss function and that the task is to seek a decision rule minimizing the Bayes risk. In these cases, the challenge is to verify that the decision rule gives a good performance; at root this depends on whether some element of the parametric family matches the DG. For instance, if the decision regards which parameter value to choose then there is a model that can be used for prediction. If the decision regards which stock to buy on a given day – i.e., an action – then the gain or loss afterward gives an assessment of how good the decision is; this evaluation is disjoint from the procedure that generated the decision or action in the first place. Another common decision problem is to decide which treatment is best for a given patient or for a given patient population. Again, one is selecting an action. One can make a decision based on data, but it is only when the predictions following from that decision are tested that one can be sure the decision was the best possibile. That is, the goal in decision making is, at root, to predict the action that will be most advantageous in some sense. Otherwise stated, the merit in a given decision is determined by how well it performs, and taking empirical performance into account (since that's what's important) makes decision making merely a way to choose a predictor. Therefore, essentially, a decision problem is a one-stage prediction problem, i.e., there is one prediction, not a sequence of predictions, so there is no chance to reformulate the predictor.

As before, in some cases decision making procedures can be turned into predictors. Indeed, the decision problem may be to find a good predictor. However, even when the decision problem is not directly about prediction, it has a predictive angle. For instance, consider the frequentist test of $\mathcal{H}_0 \colon F \neq G$ versus $\mathcal{H}_1 \colon F = G$ using IID data y^n from F and z^n from G. This is sometimes called an equivalence test. The two-sample Kolmogorov–Smirnovtest would be one of the natural test statistics if the hypotheses were reversed. To address the test, choose a distance d, write $\mathcal{H}_0$ as $\mathcal{H}_0 \colon d(F, G) > \delta$, and let $\{(F, G)|\, d(F, G) > \delta\} = \cup_k S_k$, where the S_k are sets of pairs of distribution functions, $k = 1, \ldots, K$, and the diameter of S_k is small in terms of d. Then, $\mathcal{H}_0$ vs. $\mathcal{H}_1$ is equivalent to the K tests $\mathcal{H}_{0,k} : S_k$ vs. $\mathcal{H}_1$. If the S_k are small enough, they can be approximated by their midpoints, say s_k. This gives the approximate simple-versus-simple when testing the problems $\mathcal{H}'_{0,k} \colon (F, G) = s_k$ vs. $\mathcal{H}_1 \colon F = G$. Now, in principle these tests can be done, using a multiple comparisons correction, and a single approximate model (or a small collection of approximate models) can be given from which to make predictions. Note that this is not modeling, and in fact nonparametric approaches can be used in a decision problem to generate predictions. (This approach will arise in Sec. 1.2.2.)

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

Answering a question about a *population*, or, more generally, understanding the structure of a data set, is a more nebulous goal. However, it may be regarded as trying to identify some feature of a population, or of an individual within a population, that is not obviously expressible in model identification terms. As an example, imagine trying to identify which dietary supplements the residents of a city buy or determining whether two random variables are associated. The predictive angle in these cases is one of confirmation. De facto, the prediction is that residents of the city use dietary supplements from a given list. So, if a resident of the city is chosen, does the resident use one of the identified supplements or not? If predictions are made assuming the independence of two random variables, are these predictions noticeably worse than including a dependence between them? Equally important, it is relatively rare that the final end point of an analysis is the description of a data set or the answering of a question about a population. Usually, one is doing this sort of task with a greater goal in mind, such as deciding whether to offer a new supplement for sale or classifying subjects into two classes on the basis of covariates.

Since this class of problems is less well defined, it is not obvious how to convert methods from it generically into a predictive interpretation beyond what has already been discussed. It is enough to be aware of the centrality of prediction among the various statistical goals subsumed under the term population description.

For the sake of completeness, recall that there are other statistical goals such as data presentation (graphics), data summarization, and the design of experiments. These too are generally in the service of some greater goal. Data presentation may be used to explain a statistical finding to non-statisticians, but these people generally have a reason why they want the analysis and a goal that they want fulfilled, which is usually predictive. Similarly, data summarization is rarely an end in itself but a subsidiary goal towards some other presumably greater goal. The design of experiments is done before data is collected, and its primary goal is to ensure that the data collected will suffice for the analytic goal – which, as has been argued, generally has a predictive perspective even if prediction is not recognized as the main explicit goal.

A third benefit of focusing on prediction is ensuring that inferences are testable and hence that any theories they represent are testable. Testability is not the same as interpretability, but a good predictor will typically permit some, perhaps limited, interpretation. For instance, given a predictor that uses explanatory variables one can often determine which of the explanatory variables are most important for good prediction. One would expect these to be the most important for modeling as well. More generally, apart from interpretability, theories for physical phenomena that arise from estimating a model and using hypothesis tests to simplify it must be validated predictively.

It is worth noting that, heuristically, there is almost an 'uncertainty principle' between interpretability and predictive accuracy: it's as if the more interpretability one wants, the more predictive performance one must sacrifice, and conversely. After all, the best predictors are often uninterpretable (e.g., those for the Tour and Fires data in Sec. 1.2.1, the Bacterial NGS data in Sec. 1.2.2, and the Music data in Sec. 1.2.3). Moreover, interpretable predictors (typically based on models) are almost always predictively suboptimal: it is a mathematical fact that, for instance, Bayes model averaging[1] (which is difficult to interpret) is better than

[1] Here and elsewhere Bayesian is abbreviated to Bayes for brevity when convenient.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

using any one of the models in the average (which is usually easy to interpret), at least under squared error. Also, the adaptivity of predictors to data which has little interpretability often outperforms conventional model averages or model selection; see Wong and Clarke (2004), Clarke *et al.* (2013). Thus, interpretability does not lead to good prediction and good prediction does not require interpretability – although sometimes interpretations can be derived from predictors. Indeed, relevance vector machines (RVMs) are mathematically the best predictors in some settings (reproducing kernel Hilbert spaces), but statistically they overfit and can therefore be suboptimal because of excessive variance, meaning some terms have to be dropped for improved predictive error. This does not make RVMs more interpretable – if anything it makes them more complex and hence less interpretable – but it can make them excellent predictors.

Importantly, prediction in and of itself does not require an unseen world of abstract population quantities or measure spaces. Predictors such as 'someone with high coronary artery calcium is likely to benefit from statin treatment', paraphrased from Blaha *et al.* (2011), do not require anything we have not measured or cannot measure. Similarly, 'tomorrow's weather will be the same as today's' is a purely empirical statement. We may wish to invoke the mathematical rigor of measure theory to provide a theoretical evaluation of our prediction methods under various assumptions but this is a separate task from prediction per se. Indeed, in many cases the asymptotic properties of predictors, in terms of sample size or other indices, are of interest but cannot be obtained without making assumptions that bear scant relation to reality. For instance, formally a random variable is a deterministic function on an invisible and unspecified set. Is this a reasonable way to encapsulate the concept of randomness mathematically? The answer is probably no; it's just that a better one has yet to be proposed and accepted.

A fourth reason to focus on prediction is that predictive errors automatically include the effect of uncertainty due to the data and to all the choices used for prediction. That is, when a predictor $\hat{Y}$ of Y is wrong by $|\hat{y} - y|$, the error includes not just the bias and variability of any parameters that had to be estimated to form $\hat{Y}$ but also the bias and variability due to the predictor class (or model class if models are used) itself as well as the variability in the data. This is a blessing and a curse. One of the problems with prediction is that point predictors are more variable than point estimators, so prediction intervals (PIs) are typically wider than confidence or credibility intervals (CIs). Moreover, just like CIs, model-based PIs tend to enlarge when model uncertainty is taken into account. The consequence of this is that predictive inferences tend to be weaker than parametric or other inferences about model classes. It would be natural for investigators to prefer stronger statements – even if the justification for them rests heavily on ignoring model uncertainty. However, even though inferentially weaker, point predictors and PIs have the benefit of direct testability and accurate reflection of uncertainty, which point estimators and CIs usually lack.

One of the earliest explorations of model uncertainty was by Draper (1995), who compared two ways of accounting for model uncertainty in post-model selection inference that include prediction. Draper (1995) argued that model enlargement – basically adding an extra level to a Bayesian hierarchical model – is a better solution than trying to account for the variability of model selection from criteria such as the Akaike or Bayes information criteria in terms of the sample space. He also argued that it is better to tolerate larger prediction intervals than to model uncertainty incorrectly. (As a curious note, Draper (1997) found that

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

there are cases where correctly accounting for modeling uncertainty actually reduces predictive uncertainty.) Of course, if PIs are too large to be useful then the arguments that a modeling approach is valid are more difficult to make, and any other inferences – estimates, hypothesis tests – may be called into question. However, to quote Draper (1995): 'Which is worse – widening the bands now or missing the truth later?'

There are *two criticisms* of the predictive approach that must be answered and dispensed with. First, a criticism of the predictive approach that is used to justify direct modeling approaches is that being able to predict well does not imply that the phenomenon in question is understood. The answer to this criticism is that modeling only implies understanding when the model has been extensively validated, i.e., found to be true, and this validation is primarily predictive. So, announcing a model before doing extensive validation – as is typically done – provides only the illusion of understanding. Prediction is a step toward model building, not the reverse, and predictive evaluation is therefore more honest. While the result of this kind of validation may be a predictor that is not interpretable, it is better than having an interpretable model with poor predictive performance. It may be that the traditional concept of modeling is too restrictive to be useful, especially in complex problems.

Second, it must be admitted that in practice the predictive approach is frequently harder than modeling. It's usually easier to find a not-implausible model (based on 'modeling assumptions' that boil down to the hope that they are not too far wrong), estimate a few parameters, verify that the fit is not too bad and then use the model to make statements about the population as a whole than it is to find a model that is not just plausible but actually close enough to being correct to give good predictions for new individual members of the population. Here, 'close enough' means that the errors from model misspecification or model uncertainty are small enough, compared with those from other sources of error, that they can be ignored. The problem, however, is that there are so many plausible models that finite data sets often cannot discriminate effectively amongst them. That is, as a generality, the plausibility of a model is insufficient for good prediction because one is quite likely to have found an incorrect model that the data have not been able to discredit yet. Since models that do not give sufficiently good prediction have to be disqualified, their suitability for other inferential goals must be justified by some argument other than goodness of fit. Thus, on the one hand, the task of finding a good predictor is usually harder than the task of finding a plausible model.

On the other hand, in reality a predictive approach is easier than implementing a true, accurate, modeling approach. Truly implementing a modeling approach requires that the model be correct or at least indistinguishable from correct. Given that the true model (when it exists) is rarely knowable this is an extremely difficult task. However, finding a serviceable predictor is easier, because it asks for less: giving good predictions is an easier task than uncovering a true model because bad predictions from a model invalidate the model while failure to provide good modeling inferences does not per se invalidate a good predictor. For example, if one predicts tomorrow's weather to be the same as today's weather this predictions may be reasonably accurate even though there is no underlying model from which to make inferences. Indeed, a good predictor may correspond to a dramatic simplification of a true model such that the prediction is good but the specific modeling inferences are poor.

Taking a predictive approach also requires another shift of perspective, namely that the data to be collected in the future are extremely important. This flies in the face of modeling

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-02828-9 — Predictive Statistics
Bertrand S. Clarke , Jennifer L. Clarke
Excerpt
More Information

which focuses on a specific data set and what it says about a hypothetical population rather than what it says about future outcomes. It also flies in the face of standard scientific practice, which underweights confirmatory studies. As a gedanken (thought) experiment, imagine how scientific practice, funding decisions, and scientific publishing would change if the confirmation of studies (by different experimental teams) were weighted as highly as initial findings. It's not that prediction is against rapid scientific advancement; rather, it's that prediction is a check that the advancement is real (not based on errors, luck, or malfeasance) so that scarce resources don't get squandered on spurious results.

Despite the considerations so far, which are fairly well known, the main approach taken by statisticians has been to look at *model classes* and use them to generate predictors in cases where prediction was an acknowledged goal. Here, however, the key point is that much of traditional statistics has been done precisely backward: instead of modeling, or more generally choosing a model, and then predicting, one should propose a predictor class and find a member that performs well. Then, if model identification is desirable for some reason, in principle one can convert the predictor to a model within a class of models that are believed to be plausible. For instance, in some settings Bayes model averages yield good predictors. One can form a single model from a Bayes-model-average predictor by looking at the most important terms in the models that go into the average. In the special case of averaging linear models, one can regard this as a way to find coefficients on variables using a criterion different from least squares. (The difference is that Bayes model averaging combines models after determining coefficients rather than determining coefficients for a combined model.) As another example, one can use a kernelized method such as an RVM, take a Taylor expansion of the kernel in each term of the RVM, and take the leading terms as a model. In this way one might obtain a model that is interpretable and gives good predictions. If the predictions are not quite as good as those from the original predictor, at least one can see the cost paid to obtain interpretability.

Forthrightly, the point of this book is that the *paradigm problem of statistics is prediction*, not estimation or other forms of inference, and problems should be formulated in such a way that they can be answered by producing a predictor and examining its properties. The tendency that analysts and methodologists have toward model formulation and therefrom to estimation, decision making, and so forth is misguided and leads to unreproducible results. This often happens because model uncertainty is usually the biggest source of error in analyses, especially with complex data. Predictor uncertainty may be just as big a problem, but it is visible in the predictive error while model uncertainty is very hard to represent accurately.

Conventional statistical modeling recognizes the problem of model uncertainty in a variety of ways. Most recently, model uncertainty has been recognized in the desire for sparse models that satisfy the oracle property (they can correctly select the nonzero coefficients with high probability; see Sec. 10.5). Clever as all this is, it is merely a way to find a model that is not implausible, i.e., cannot obviously be ruled out. Indeed, shrinkage methods generally perform better predictively when they shrink less, i.e., are less sparse and distort the data less. Moreover, as they shrink less, shrinkage methods tend to improve and become competitive with the better model-averaging methods; see Clarke and Severinski (2010). As a generality, shrinkage methods frequently are a source of model misspecification since sparse models are rarely true outside narrow contexts. Indeed, the desire for sparsity is a variant on the desire for models (and small variance), since models are a way to summarize

© in this web service Cambridge University Press www.cambridge.org