MODEL SELECTION IN ITEM RESPONSE THEORY

by

Peter Francis Halpin

B.A. (Honours), University of Calgary, 2002

M.Sc., University of Calgary, 2005

a Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in the Department
of
Psychology

SIMON FRASER UNIVERSITY
Spring 2010

APPROVAL

Name: Peter F. Halpin

Degree: Doctor of Philosophy (Department of Psychology)

Title of Thesis: Model Selection in Item Response Theory

Examining Committee:

Chair:
Dr. Thomas Spalek
Associate Professor

Dr. Michael Maraun
Senior Supervisor
Professor

Dr. Kathleen Slaney
Supervisor
Assistant Professor

Dr. Jack Martin
Supervisor
Professor

Internal Examiner:
Dr. Rachel Altman
Assistant Professor
Statistics and Actuarial Science

External Examiner:
Dr. Victoria Savalei
Assistant Professor
Department of Psychology
University of British Columbia

Date Approved : January 29, 2010

Abstract

The problem of model selection is addressed from a general perspective and solutions are considered within the domain of item response theory (IRT). Selection is conceptualized as including both the evaluation of individual models and the simultaneous comparison of multiple candidates. Traditional tests of goodness of fit can often be regarded as dealing with the former situation, while information criteria can only be applied to the latter. The significance of this last point is pursued in some detail. In terms of optimization, it is shown that information criteria do not provide a means of determining how well their various objective functions are satisfied. This implies that some further criterion is required in order to establish whether the candidates recommended by any information criterion are indeed satisfactory. The need for such a criterion motivates the present work. This approach begins by conceptualizing parametric stochastic models as sets of probability distributions. In any given application the purpose of such a model is to predict the relative frequencies with which an outcome variable takes on its values. This notion of prediction is described in terms of the inclusion of the distribution of the outcome variable in the set of distributions implied by the model: If this is not the case, the model is said to be inaccurate. The concept of accuracy then serves as a basis for selection in IRT. In particular, any

IRT model can be represented as a manifold embedded in Euclidean space, and the proximity of any observed distribution to a point on this manifold can be interpreted in terms of the norm of their difference. Describing the geometric properties of sets of candidates provides a means of selection that is not tied to any particular set of observations; this is an important area of further investigation.

Acknowledgments

I would like thank Dr. M. D. Maraun for sharing with me many insights into psychometrics and also for his support in developing my own perspective. To my parents, thank you for patiently keeping me sane throughout the highs and lows of academic pursuit, and to my brother, for always encouraging me to take credit for what I have accomplished.

Contents

List of Tables

Chapter 1

Introduction

The problem of evaluation is basic to the application of statistical models and has been addressed from a large variety of perspectives. From the perspective taken in this dissertation, the character of the problem may be stated as follows: How do we decide when a model is adequate to its intended purpose? There are many purposes to which a model may be directed, for example representation, prediction, or explanation, and it is important not to premise the statement of the problem on only one or another of these. Regardless of the purpose to which a given model is applied, the question of how well it meets this purpose can always be asked – that is, we can always raise the problem of evaluation. The main presupposition of this questioning is the necessity of justifying the interpretation of statistical models, and this is especially relevant to their applications in scientific research.

Although the problem of evaluation may be phrased in very general terms, its solutions are necessarily more particular. In principle the evaluation of a model must admit its stated purposes, and in practice consideration must be limited to some formalization of the models of interest; we can neither consider all purposes nor all models simultaneously. For this reason, there can be no single solution to the evaluation problem. Nonetheless, it is desirable to articulate purposes and models that are sufficiently broad to find relevance to a wide range of applications. Having established this initial context the goal is to describe methods of evaluation that can be advantageously employed therein.

This dissertation is concerned with methods of model *selection*, which is understood to mean the simultaneous evaluation of one or more models. This approach is premised on a set of models that are considered to be potentially adequate for a stated purpose, and these are referred to as *candidates*. The possibility of multiple candidates reflects an important characteristic of contemporary stochastic modeling, namely a plethora of choice. This is the context that the approach taken in this dissertation finds its relevance. This approach is essentially a method of *elimination*, and it strives for a more stringent evaluation of models than is afforded by current methods based on so-called information criteria (IC)[1]. For reference, the guts of the argument against IC as a sufficient means of model evaluation are given in the following two paragraphs. This motivates the approach taken in this dissertation, and it assumes that the reader is familiar with IC. The argument is elaborated in the following section of this introductory chapter.

In contrast to IC-based approaches, it is argued that a model has been properly evaluated only if the means of evaluation have allowed for the possibility that the model is judged to be inadequate. Without such a criterion of (in)adequacy, selection can only proceed in an *ad hoc* manner and this occurs, for example, when the optimal value of an IC is defined as its minimum over a set of candidates. In particular, this

[1] The notation 'IC' is used to denote the singular and the plural of information criterion

optimal value cannot be defined independently of a set of candidate models. The intuitive strategy behind IC-based selection is to choose the *best* model in a set of candidates, where 'best' is defined by the objective function of the particular IC in question. However, this strategy offers no assurance that the best best is also a good one, that it attains a value of the objective function that should regarded as adequate. IC cannot be used for this latter purpose because they are related to their objective functions by unknown quantities, and hence their objective functions do not provide grounds for direct interpretation of their numerical values.

To see that the method of evaluation offered by IC is not sufficient for the selection of adequate models, consider the following argument. As is well known, the use of IC is not appropriate when the set of candidates consists of a single model. In comparison with conventional methods of testing the goodness of fit of a single model, this would be similar to declaring the level of significance to be the observed probability of the test statistic. But when the set of candidates consists of multiple models, the optimal value of an IC is no less arbitrary. In both cases, the preferred candidate cannot be inadequate since adequacy *qua* optimality is defined with respect to that model.[2] Thus IC do not properly evaluate the models they recommend, even though they may be regarded as evaluating the other candidates with respect to the recommended model. In terms of practical applications the basic point can be stated as follows: IC can lead us to chose an unsatisfactory model by comparing it to models that are even less satisfactory. Therefore IC do not provide a sufficient solution to the problem of model evaluation. Rather, the appropriate use of IC is seen to be contingent on choosing a suitable set of candidate models. Obviously, to *assume* that any set of

[2] The same rationale holds when "ties" occur.

candidates must include an adequate model is just to beg the question. Perhaps less obviously, to assert that an “inadequate” model is better than no model is just to assume that the chosen model meets a standard of adequacy (i.e., that it is better than no model). If this assertion is not taken as an empirical one, it reflects only a dogmatic application of statistical methods. Clearly this is not what we should hope to find at the root of a theory of model evaluation.

The present work addresses these difficulties by pursuing a criterion of model adequacy along the following lines. In chapter two the purpose of statistical models is reduced to that of prediction. If a model predicts observations well (i.e., if it is accurate), then it is to be regarded as adequate (§2.1 and §2.2). There are various notions of prediction found in the selection literature and that employed here is based on, though not identical to, that found in the work of Bamber and van Santen (1985, 2000). Arguments are adduced for why this conception of prediction is a good criterion for model evaluation. For instance it is quite minimal and therefore has broad application. It can also be used to formulate selection procedures that are *fast* in the sense of allowing one to make conclusions about a set of candidates by means of a number of computations that is potentially less than the number of candidates. In particular, a selection algorithm is proposed for the case of disjoint models (i.e., non-nested and non-overlapping families of probability distributions). While this algorithm will not often find application, it demonstrates a *principle* of model elimination (§2.3). This principle can be interpreted as instantiation of Popperian falsification or Platt’s (1964) strong inference. In its current form, the basic idea is that we can always get rid of large numbers of inaccurate models faster than we can find a single “optimal” model.

Although accuracy is here taken as the first or basic purpose of stochastic models, it has often been argued that this is not a sufficient grounds for model preference. For instance, this is seen in the over-fitting or overparameterization of a model, in which case perfect accuracy can be achieved at the expense of triviality. Thus it is important to distinguish two related aspects of the problem of model evaluation. The first of these is termed *data-based* selection, and this refers to the task of finding a model that fits the data. This is the problem to which the concept of accuracy is applied in this manuscript. The second aspect is referred to as *model-based* selection, and this is the problem of defining properties that are desirable for a model to possess. Some usual ideas here include parsimony, non-triviality, and non-equivalence among sets of models. Note that in principle, model-based selection does not require a particular set of observations, nor even a set of more than one candidate. Rather, it can be thought of as an aesthetic consideration of the model itself. Yet, no matter how compelling a set of equations is to one person, it can always be wrong, or more particularly, wrongly applied. For this reason it is important to separate considerations about model accuracy from considerations about how "nice" a model is. These higher purposes are only addressed tangentially in the course of this dissertation.

For clarity, it should be noted that this argument against the sufficiency of accuracy as a criterion of model selection is not like the argument raised above against the sufficiency of IC as methods of model evaluation. In the present case the argument is metaphysical, it pertains to the purpose of models and whether the purpose adopted here is a good one. In the former case the purpose of a model is not in question, no inquiry is made into the validity of the objective functions of IC. Rather it is claimed that, if we accept these purposes, the quantities in question are not sufficient to the problem of evaluation. This argument is epistemological, it pertains to the justification of the interpretation of stochastic models. The position taken in the present work is that accuracy is sufficient to a particular interpretation of stochastic models, namely that they fit the data. That this purpose is not also sufficient to other purposes of stochastic models is recognized by distinguishing data-based and model-based selection.

The third chapter applies the general approach outlined in chapter two. The family of models that serve as the focus are those found in item response theory (IRT). In particular, consideration is restricted to the traditional dichotomous-response models for tests of fixed order and fixed length (§3.1). A measure of predictive accuracy is proposed for these models, although it applies also to other models of finite-valued variables. The proposed quantity is a Euclidean distance between two multinomial distributions (§3.2). This shares some properties with the Kullback-Leibler divergence of information theory when a model is correctly specified. Although this distance is readily estimated in principle, when "empty cells" become problematic the use of marginal distributions or subtests is considered (§3.3). Data-based selection with this measure of accuracy is illustrated by means of an empirical example based on the Self Monitoring Scale (Snyder, 1974; §3.4).

An interesting feature of IRT models is that the model itself, as well as the data, can be viewed as normative. This is importantly dissimilar from, say, applying a statistical model to the description of a specific experimental phenomena predicted by a scientific theory. In this latter case, we would be hard-pressed to justify the omission of data that are not predicted by a model, and the usual course of action is to replace an unsatisfactory candidate with one that is more accurate. On the other hand, the data to which IRT models traditionally have been applied are pen-and-paper tests of educational achievement. In this situation the data generating instrument (i.e., the test) is under just as much scrutiny as the models themselves, if not more so. As such, if certain test items are found to be inconsistent with respect to a given model, this can be taken as grounds for omitting those items rather than changing the model. In the present approach, this is referred to as *test construction* and it is interpreted as data-based selection when the models rather than the data are viewed as normative. The application of the proposed measure of model accuracy to problems of test construction is one of its main strengths (§3.3). Note that the question of whether the "changing of roles" between data and models is appropriate in any given case is not a statistical matter. Of course, this does not suggest that it is not an important matter, only that it will not be considered in the present discussion.

The final chapter briefly summarizes the current work with an eye to its limitations and future directions. At the outset it is worthwhile to contextualize these efforts more generally, and this is the purpose of the remainder of this introductory chapter.

1.1 A Review of Model Selection Methods

Many different approaches to the general problem of model selection have been taken, for instance in statistics (e.g., Linhart & Zucchini, 1986; Rao & Wu, 2001), economics (e.g., Dyrmes et al., 1972; Grasa, 1989; Vuong, 1989), information theory (e.g., Akaike, 1973; Rissanen, 2007; Schwarz, 1978) and elsewhere (e.g., Bozdogan, 1987; Forster, 2000; Myung, Balasubramanian & Pitt, 2000). Although many of these approaches have been developed quite recently, it would require many tomes of technical details to provide a comprehensive review. The intention of the present section is only to provide some contextualization of the contributions of the subsequent chapters. For a more thorough review, the reader is referred to Claeskens and Hjort (2008) or Rao and Wu (2001). For the most part, historical details have yet to surface from the primary sources, although Burnham and Anderson (2002) give a short historical motivation of the modern context and de Leeuw (1992) discusses the influential 1973 paper of Akaike.

The main task of this review is to summarize the characteristics of two information criteria, AIC and BIC, to describe their application to model selection, and to point out some aspects of this application that would be desirable to improve. Although there is a remarkably large number of IC currently in circulation, there are also various reasons for focusing only on these two. For instance they have historical precedence (Akaike, 1973; Schwarz, 1978), and their asymptotic properties are characteristic of many other IC as well as some related selection procedures (Claeskens & Hjort, 2008, chap. 4; Grünwald, 2007, §17.3; Sin & White, 1996, Yang, 2005). The present rationale for focusing on AIC and BIC is as follows. Information criteria are typically interpretable either in terms of Kullback-Leibler divergence / "true models" or in term of posterior probabilities. AIC and BIC are, respectively, the prototypical IC corresponding to these two interpretations. It is these interpretations that are the main concern here, and in particular it is argued that AIC and BIC provide incomplete information about the concepts that are the basis of their interpretation.

This argument can be clearly stated in term of optimization: the quantities estimated by AIC and BIC are not equivalent to the objective functions that motivate those quantities. Rather, the quantities estimated by AIC and BIC can be interpreted as affine transformations of their objective functions, and in both cases those transformations contain unknown constants.[3] As a consequence, the numerical values of the objective functions have an unknown relationship to the numerical values of the IC. Thus AIC and BIC cannot be used to determine if any given model is acceptable in the sense defined by their objective functions. They can however be used for purposes such as ranking models, ranking differences among models, and for interpreting other properties that are preserved by affine transformation (see Suppes & Zinnes, 1963). This point is well known in the technical literature and the present review merely serves to spell-out its significance. The extension of this argument to IC other than AIC and BIC is not made explicitly, because, to the best of my knowledge, all other selection statistics that go under the title of 'information criteria' have been explicitly introduced as different ways of estimating the quantities proposed by Akaike (1973) and Schwarz (1978). However, it should be recognized that the arguments made here are limited to such such statistics.

Before addressing the IC-based approaches, traditional methods based on testing goodness of fit are mentioned . These are summarized only to make the following two points. Firstly, they are properly viewed as applications of the theory of estimation to the problem of model selection, rather than a treatment of the latter as an independent topic. This point is not contentious – tests of goodness of fit are just hypothesis tests and hypothesis tests are at root methods of point estimation. This is a main reason that such tests have not been received as a general solution to the problem of evaluation, and in particular their application is largely limited to correctly specified,

[3] In the case of BIC this is a rather simplified interpretation; its validity depends on the prior probabilities being constant across models. Also, this transformation is linear rather than affine, but this distinction is of no relevance to the arguments made here.

nested models. Since IC-based approaches are not limited to such cases, this first point serves to indicate the advantages offered by the newer methods.

The second point shows that there are also disadvantages. In some circumstances, goodness of fit tests can be interpreted to make decisions about a single model, and so they are not intrinsically limited to ranking multiple candidates. Importantly, this has to do with the quantity that is the basis of such decisions, namely probability. Unlike the values of IC, probabilities have interpretable properties beyond those afforded by the class of affine transformations. In particular, it makes sense to talk about "significance levels" (i.e., values close to zero). Although the stipulation of significance levels can be more or less arbitrary, this arbitrariness is regarding which value to select, not the meaning of the selected value. On the other hand, the numerical values of IC have an unknown relationship to the numerical values of their objective functions. Therefore, we cannot use IC to determine whether any given model meets a pre-established criterion of adequacy – although such a criterion may be defined in terms of the objective functions themselves, the values of the objective functions have an unknown correspondence with those of the IC. In short, there can be no significance levels for IC.

There appears to be little appreciation of this last point in discussions of the newer approaches. Indeed, the general consensus seems to be that tests of goodness of fit are now a bygone approach and that selection should proceed exclusively by means of IC or related methods. Here the interpretability of tests of goodness of fit has too quickly been forsaken for the broader applicability of IC. In particular, it is clear that lore is accumulating about how to interpret the values of the newer statistics. For example, one of the most pervasive of these "heuristics" is the notion that IC measure a trade between model fit and complexity. While these interpretations may have limited pedagogical value, they occlude the fact that IC are estimates of well-defined quantities and that they can be imbued with no more meaning than afforded by those quantities. An auxiliary purpose of this introduction is to dispel some of the mystification surrounding IC-based model selection.

Before proceeding, it should be mentioned that the notation employed in this initial discussion is defined in more detail in the following chapter. For now it is assumed that conventional interpretations of statistical quantities are familiar to the reader. Also note that derivations of the relevant quantities are sketched only in enough detail to explain their basic interpretation, and applications are not considered. For a more pedagogical discussion see Claeskens and Hjort (2008).

1.1.1 Tests of Goodness of Fit

Traditional approaches to model selection are largely based on tests of goodness of fit of two nested models.[4] The model-implied probability distribution of a model M may be written as $p_{_M}$, and let the parameter of this model be denoted by $\theta \in \Theta_M \subseteq \mathbb{R}^K$ where K is an integer greater than zero. When one model, M_*, is nested within another, M^*, this means that $p_{_{M_*}} = p_{_{M^*}}$ but $\Theta_{M_*} \subset \Theta_{M^*}$. This is denoted by $M_* \subset M^*$, where the orientation of the asterices serves to indicate which model is nested.

Note that selection between two nested models requires only the comparison of

[4] A variety of graphical methods for evaluating model fit could also be classed under the title of traditional. These are not considered here.

the possible parameter configurations of a single model M^*. When a certain parameter configuration has a special substantive interpretation, it is termed a 'submodel'. In this context, selecting among parameter arrangements of M^* is interpreted as selecting among models. Tests of such models are usually based on the (asymptotic) distribution of a function of their likelihood ratio. The reference distribution is derived on the premise that both likelihoods are sampled from the same population and rejection of this assumption is interpreted to imply or otherwise motivate the rejection of M_*. Here it may be noted that the preferable model may correspond to either M_* or M^*, depending on the research scenario. The following two examples serve to illustrate this general approach to selection.

At each comparison in a stepwise linear regression, M_* implies that $q \geq 1$ of the regression coefficients in M^* are fixed to zero (e.g., Rao & Wu, 2001). If the parameters of the models are estimated by maximum likelihood then a testing distribution is $F_{(q,\, s)}$, where s is the degrees of freedom of the residual sums of squares of M^* (see Searle, 1971, §3.6 & §3.7). If the two models are statistically different, M_* is rejected and M^*, the model with more free parameters, is viewed as making gains over the simpler model. As another example, consider the likelihood ratio tests applied in structural equation modeling (SEM; e.g., Bollen, 1989). In this case M_* implies an overidentified covariance matrix. M^* can imply another overidentified covariance matrix or, more usually, it is an "unrestricted model" obtained by assuming that the population covariance matrix is any positive definite matrix (i.e., setting the estimated "model-implied" covariance matrix to that of the sample). Let us focus on the latter scenario, in which case the null distribution of negative two multiplied by the likelihood ratio is asymptotically distributed as $\chi^2_{(r-K)}$, where r is the number of unique elements in the sample covariance matrix (Browne, 1984, Corollary 4.1). If the tail probability of the test statistic is small, this is again taken as grounds to reject M_*, but this time we reject the model of interest as placing invalid restrictions on the population covariance matrix.

Especially in sequential procedures, the interpretation of these tests can become rather convoluted but, as outlined here, their general premise is just that of parametric hypothesis testing. As such, they are subject to a large number of criticisms, for instance regarding the truth of the null hypothesis and issues pertaining to having either too little or too much power. An important and long standing complaint about these tests is that they do not generalize nicely to non-nested models (e.g. Leamer, 1978). Applications of conventional likelihood ratio tests to two non-nested models have been proposed, but these are often computationally intractable, do not have a clear interpretation, and in general are not satisfactory (see Dyrmes et al., 1972). Vuong (1989) has made important advances by describing the asymptotic distributions of likelihood ratio statistics for non-nested and / or misspecified models. These results are of interest in their own right but, as discussed by Vuong (1989), his tests are premised on the approach described by Akaike (1973). In short, Vuong's test are of the difference between two AIC-like quantities. These tests are therefore are properly viewed as a modification of the approach described below in connection with AIC. They are also dissimilar with respect to the following characteristic.

One important aspect of goodness of fit tests is that they can be interpreted, in some cases, as evaluating a single model. For example, in the SEM case described above, although an unrestricted model is used to construct the likelihood ratio, it is not possibly false. Of course, certain assumptions are required in order to derive the reference distribution (see Browne, 1984), but these are arguably general enough to be classed under the ever-mysterious rubric of “regularity conditions” rather than taken as constitutive of a model *per se*. And, regardless of how the technical interpretation of the unrestricted model is settled, it is nonetheless clear how it is used in model selection. Its role is to assess the validity of a single M_*. In particular, if a test of M_* versus the unrestricted model yields a significant p-value , one rejects M_*. This is not taken to mean that we should stick with M^*, and this is just because M^* is not interpreted as a model. Rather it is usually taken to imply that we need to modify or otherwise improve M_*.

The essential point about this example is that it is possible to reject M_*. This is done on the basis of some level of significance, and although the significance value itself may be more or less arbitrary, the scale on which it is interpreted is not arbitrary. In measure theoretic terms, a probability measure always assigns unity to the entire sample space and zero to the empty set. Therefore the numerical magnitudes of probabilities can be interpreted relative to their possible range. In particular, it makes sense to talk about how small or large a given value is without explicitly referencing another value, and it is for just this reason that significance levels make sense. As we shall see directly, there can be no significance levels for information criteria, and in general they cannot be used to make conclusions about a single model. Arguably, this is the most significant difference between conventional tests and the newer information theoretic approaches.

1.1.2 Akaike's Information Criterion

The first information criterion was due to Akaike (1973; see de Leeuw, 1992, for historical commentary) and is named after him (AIC). In his paper Akaike motivates KL (Kullback, 1983) as an appropriate loss function for model selection. This can be viewed as the most basic and far-reaching insight of his approach. It provides a general premise for model selection by defining a purpose of models, namely the minimization of KL. Moreover, this purpose is also seen to provide a motivation for maximum likelihood (ML) estimation, a motivation that is arguably less arbitrary than simply declaring it to be a principle. The relevant version of Akaike's loss function may be written as

$$KL(p_O, p_M) = \int \big(ln(p_O(x))p_O(x) - ln(p_M(x;\theta))p_O(x)\big)dx \tag{1.1}$$

where p_O is the probability distribution of the data. This distribution is often referred to as the "true model" and it is discussed in more detail in the following chapter.

KL has various interpretations in statistics, information theory, and computing. In statistical applications the title 'KL divergence' is used most commonly and in this context it describes a pseudo-distance between the distributions p_O and p_M (see Claeskens & Hjort, 2008). On this perspective, Akaike's goal in treating KL as a loss function is typically interpreted as trying to find the "least false" parameter values of *M*, or as searching for the model-implied distribution that is closest to p_O. KL's information theoretic interpretation as relative entropy is based on its relation to Shannon information (Kullback, 1983). In computing it can be interpreted as the number of extra bits required to encode information from p_O when using a code that is optimal for $p_{M}(x;\theta)$ (see Grünwald, 2007, for discussion of these concepts). Generally speaking, KL has been viewed by some, including Akaike (1973), as a truly foundational quantity (e.g., Burnham & Anderson, 2000), whereas others seem to regard it as just another loss function (e.g., Linhart & Zucchini, 1986).

Another important aspect of Akaike's treatment of the model selection problem is the meaning of his often overlooked "outer expectation" (de Leeuw, 1992). In the usual practice, the expectation of a loss function with respect to p_O yields the corresponding risk (i.e., average loss) function. It seems that Akaike's intended use of KL divergence presents a problem here, since there is already such an expectation in equation 1.1. Hence he makes a rather interesting modification to the notion of risk. In particular, he obtains his risk function by taking the expectation with regard to a transformation of the data, namely an estimator $\bar{\theta}$ of $\theta \in \Theta$. Furthermore, he requires the observations on which $\bar{\theta}$ are calculated to be independent from those on which other expectations with regard to p_O are calculated (e.g., to be from different i.i.d. samples). Intuitively, Akaike wants to average his loss over the values of an estimator of θ rather than the data itself. Technically, this explains the peculiar behavior of his outer expectation. Following de Leeuw (1992) this behavior will be emphasized by letting X and Z be distributed according to p_O, requiring that $X \perp Z$, and, where relevant, writing $\bar{\theta}(\cdot)$ for an estimator of $\theta \in \Theta$. Konishi (1999) otherwise explains this situation as one where we have observations on X and want to generalize them to some future observations on Z. Using this notation, Akaike sought to minimize

$$R_{KL} = 2 \int \int \Big(\ln(p_O(x)) p_O(x) - \ln\big(p_M(x;\bar{\theta}(z))\big) p_O(x) \Big) p_O(z) dx dz. \tag{1.2}$$

Equations 1.1 and 1.2 are the basic quantities considered by Akaike (1973). Let us now briefly address how these motivate the ML estimator for θ and his famed information criterion. By inspection of equation 1.1, it is clear that the term $E_{\scriptscriptstyle X}\big(\ln(p_{\scriptscriptstyle O})\big)$ does not depend on M. Therefore, for fixed $p_{\scriptscriptstyle O}$, minimization of $KL(p_{\scriptscriptstyle O}, p_{\scriptscriptstyle M})$ depends only on maximizing $E_{\scriptscriptstyle X}\big(\ln(p_{\scriptscriptstyle M})\big)$. Akaike's argument requires the existence of a unique argument that maximizes this expectation, which may be denoted by θ_o. Next note that, for $i = 1, \ldots, n$ i.i.d. observations from $p_{\scriptscriptstyle O}$, the average log-likelihood, $n^{-1}l(\theta) = n^{-1}\sum \ln(p_{\scriptscriptstyle M}(x_i; \theta))$ converges almost surely to $E_{\scriptscriptstyle X}\big(\ln(p_{\scriptscriptstyle M})\big)$ (see van der Vaart, 1998, §5.5). Therefore the ML estimator $\hat{\theta}$ asymptotically approaches θ_o, and the principle of maximum likelihood is seen to be motivated by a principle of minimum KL divergence.

As per the foregoing discussion of equation 1.1, $\bar{\theta}$ in equation 1.2 is replaced by $\hat{\theta}$. The next task is to minimize R_{KL} in $\hat{\theta}(Z)$, and, similarly to the above approach, Akaike's move is to maximize

$$Q = E_{\scriptscriptstyle Z}\Big(E_{\scriptscriptstyle X}\big(\ln(p_{\scriptscriptstyle M}(X; \hat{\theta}(Z)))\big)\Big) = E_{\scriptscriptstyle Z}\big(W(Z)\big). \tag{1.3}$$

The difficulty here is to obtain an estimate of Q when we only have observations on X. Other than $\hat{Q} = n^{-1}l(\hat{\theta}(x_i))$ there is little to work with, but in this case the expectations over X and over Z are estimated from dependent data (e.g., from the same sample). In particular, $\hat{Q}$ can be shown to be positively biased by means of second-order Taylor expansions of $\hat{Q}$ and $W(Z)$ about θ_o. Writing the bias term as $B = \hat{Q} - W(Z)$ it follows that $\hat{Q} - Q = B + W(Z) - Q$ and so $E_{\scriptscriptstyle Z}(\hat{Q} - Q) = E_{\scriptscriptstyle Z}(B)$. Derivations of an asymptotic approximation to $E_{\scriptscriptstyle Z}(B)$ can be found, for example, in Claeskens and Hjort (2008, pp. 30-31), and Konishi (1999, pp. 390-392). Akaike's (1973) derivation is not particularly easy to follow for a variety of reasons discussed by de Leeuw (1992). Describing these details and the long lists of regularity conditions that accompany them would take us too far astray here. The basic result is stated for reference:

$$E_{\scriptscriptstyle Z}(B) = n^{-1} E_{\scriptscriptstyle Z}\big(\sqrt{n} u' J^{-1} u \sqrt{n}\big). \tag{1.4}$$

In equation 1.4, the vector u is the score of $p_{\scriptscriptstyle M}(z;\theta)$ evaluated at θ_o and the asymptotic distribution of $\sqrt{n}u$ is $N_K(\mathbf{0},\, \Sigma)$. The matrix J is the expectation with respect to $p_{\scriptscriptstyle O}$ of the negative of the Hessian of the log likelihood evaluated at θ_o.

There are a variety of ways of obtaining Akaike's estimate of equation 1.4. Importantly all of these require that $p_{\scriptscriptstyle O} \in p_{\scriptscriptstyle M}(\Theta)$, where, as discussed in the following chapter, $p_{\scriptscriptstyle M}(\Theta)$ denotes the family of distributions implied by M. In this case $J = \Sigma$ (under certain regularity conditions) and equation 1.4 can then be read as asking for the expectation of an asymptotically $n^{-1}\chi^2_{(K)}$ distributed quantity, leading to $E_{\scriptscriptstyle Z}(B) = K/n$. Since B is a quadratic form, we could also take $E_{\scriptscriptstyle Z}(B) = \mathrm{tr}(J^{-1}\Sigma) = K$. Both of these lead to Akaike's (1973) criterion. Here this is presented in its "smaller-is-better" version:[5]

$$AIC(M) = -2l(\hat{\theta}) + 2K. \tag{1.5}$$

The 'smaller-is-better' title indicates that minimizing $-2Q$ is equivalent to maximizing $2Q$. It also quite clearly states the use of this statistic in model selection: For a given set of observations, the candidate that implies the smallest value of AIC is declared to be the best model. Otherwise stated, that model is estimated to minimize R_{KL} over the set of candidates. Note also that the sample size, which is constant in this

[5] Some writers leave AIC in its "larger-is-better" form, although Akaike presented his statistic as it is shown here, presumably to emphasize its relation to $-2l(\hat{\theta})$. The present form is also that seems to be more commonly employed in current statistical software.

model selection context, is omitted from AIC. Because Akaike's purpose was only to minimize R_{KL}, it makes no difference to estimate nQ instead of Q.

As outlined here, AIC is a bias-corrected, asymptotic estimate of the maximum of nQ when $p_{O} \in p_{M}(\Theta)$. While there can be no disagreement that this is an original and interesting approach to model selection, it does have certain shortcomings. For example, many writers have addressed how to better estimate equation 1.4, and this has led to a variety of other bias-corrected information crtieria (e.g., Bozdogan, 1987; Konishi, 1999). However, there has been little discussion of the shortcomings of the quantity to be estimated, Q. Indeed the principle of minimizing R_{KL} via maximizing Q seems to have been widely accepted as something of a panacea for the problems of model evaluation. Yet, while KL may be accepted as a suitable description of the purpose of a model, this does not imply that we should also accept Q as the best quantity by which to realize this purpose. Indeed, it is a remarkable historical phenomenon that "half of a divergence" could have so quickly brought about a veritable revolution in the theory of model selection.

One important component of this phenomenon is that the form of AIC, rather than the quantity Q, often serves as the basis of its interpretation. Moreover, the same interpretations are also applied to other IC, even when these estimate different quantities (e.g., Claeskens & Hjort, 2008, §1.1). While such "heuristics" arguably have some pedagogical value when making comparisons among IC, they can also lead to confusion about what these quantities actually are, or more particularly, what they estimate. One familiar heuristic is in terms of penalized likelihoods or penalized goodness of fit. It is worth pondering the meaning of this penalization, since it is both ubiquitous and polymorphous. By inspection of equation 1.5, it is clear that the more likely a set of observations are under M, the smaller (better) the value of the statistic. However, AIC increases in the number of parameters of M, and hence the more parameters M has the worse off it is. This penalty for increasing parameters is quite naturally interpreted in terms of William of Occam's old concern about the unnecessary multiplying of entities (Forster, 2000), and so it is often taken as an instantiation of the principle of parsimony. This also relates to issues of overparameterization, and another frequently employed concept is that of model complexity (e.g., Grünwald, 2007; Myung, 2000). These recasting of AIC also seem to assume that $l(\hat{\theta})$ necessarily doesn't decrease with K, although I do not know of a general proof of this claim for non-nested models. In any case, it seems that IC are also quite universally interpreted in terms of an implied "trade off" between fit and complexity. It is striking that none of these interpretations ostensibly has much to do with the manner in which AIC is derived – they do not relate to the minimization of R_{KL} let alone the quantity Q. Otherwise stated, these interpretations seem to propose various ideas of what a good model should do, but these ideas do not seem to square with minimizing KL divergence. It is therefore worthwhile to spell-out the interpretation of AIC with respect to its derivation, and indicate how this is dissimilar from the types of interpretations outlined here.

The numbers yielded by AIC are estimates of the maximum of nQ and they are thereby interpretable as minimizing R_{KL}. This can serve to order candidate models with respect to their pseudo-distance from p_O, and this is the application that AIC has found in model selection. However, AIC doesn't tell us anything about the magnitudes of these distances, and this would require the other term in R_{KL}, namely $E_X \ln(p_O)$. For example, it could be the case that all the models are much farther from p_O than they are from each other, in which case the differences among their AICs is much less substantial than their divergence from p_O. This would be a case of choosing the best of the worst, and it illustrates that the magnitudes of numerical values of AIC cannot be interpreted in terms of R_{KL}. Alternatively, all the candidates could be very close to p_O, in which case the differences among their AICs may be indicative of substantial differences in relative proximity to p_O. This would be a case of choosing the best of the best, and it illustrates that the magnitude of differences between values of AIC cannot be interpreted with respect to R_{KL} either. Otherwise stated, both the raw values and differences among values of AIC can only be interpreted relative to other such values. For this reason AIC can only be interpreted as ordering a set of > 1 candidate models, and in particular it cannot tell us if any of those models minimizes R_{KL} to an extent that might be regarded as satisfactory. As discussed below, similar shortcomings are found with other information criteria. To facilitate the summary of these shortcomings at the end of this chapter, any fit statistic that is only interpretable as ordering models (or their differences) will be referred to as a *relative fit index*.

It may be noted that the interpretation of AIC as a relative fit statistic is not suggested by the heuristics discussed above. It would be natural to expect that a quantity purporting to measure trade-offs between fit and complexity would tell us whether we are getting a good deal in any particular case or at least whether the difference between two deals is important. On the other hand, the fact that AIC cannot provide this information is obvious when considering what it estimates.

1.1.3 Schwarz's Information Criterion

As mentioned, other information criteria have been developed along lines similar to AIC, their differences mainly being how the bias of the normalized log-likelihood is adjusted for, how outliers are dealt with, special forms for particular models, and so on (see Claeskens & Hjort, 2008, chap. 2). A different type of information criterion is due to Schwarz (1978) and its interpretation is premised on Bayes' factors. As such it usually goes under the title of the *Bayesian information criterion* (BIC), although unlike many Bayesian methods it avoids the use of prior probabilities. This and its computational affinity to AIC are perhaps the two main reasons that this criterion has been a main runner-up to that of Akaike.

The Bayes' factor is obtained by considering the ratio of two posterior probabilities, otherwise called the 'posterior odds' (Kass & Raftery, 1995). In this ratio the denominators of the two posterior probabilities cancel, and the prior probabilities also cancel when they are assumed to be equal. What is left is the ratio of two likelihoods, which is called the Bayes' factor of the numerator in favor of the denominator. In the context of model selection the Bayesian approach is to average likelihoods over θ rather than maximizing them in θ. This requires a prior distribution for θ, and these distributions are naturally viewed as conditional on the M to which θ belongs. Letting x^* denote a fixed realization of X, the notation $p_{M}(x^*\theta) = L_M(x|\theta)$ is introduced to distinguish the joint probability distributions implied by M from its likelihood function. Using this notation and letting θ_r denote the parameter of M_r, the Bayes' factor for M_i over M_j is:

$$B_{ij} = \frac{L_{M_i}(x)}{L_{M_j}(x)} = \frac{\int L_{M_i}(x|\theta_i) f_{\theta_i|M_i}(\theta_i) d\theta_i}{\int L_{M_j}(x|\theta_j) f_{\theta_j|M_j}(\theta_j) d\theta_j} \tag{1.6}$$

Schwarz (1978) was concerned to approximate the logarithms of the integrals in equation 1.6 under the assumption that p_{M} is an exponential family distribution. Raftery (1995) and Claeskens and Hjort (2008, chap. 3) provide approaches that do not specify the form of p_{M}, although these derivations incur so much asymptotic error that it is surprising to find anything left at the end. By all accounts, the main trick with the derivation of BIC is to introduce the sample size into equation 1.6. Of the various approaches I have encountered, none of them accomplish this in a particularly convincing manner.

Begin by writing the average likelihood of a model as:

$$L_M(x) = \int L_M(x|\theta) f_{\theta|M}(\theta) d\theta = \int \exp\{\, g_M(\theta) \,\} d\theta. \tag{1.7}$$

Then consider the second order Taylor expansion of $g_M(\theta)$ about its unique maximizing value, the posterior mode. As with the above discussion of AIC, the optimal argument of the objective function in question will be denoted by θ_o. And, as with AIC, the existence of this value is a basic requirement of the argument discussed here (Tierney & Kadane, 1986). In particular it serves to ensure that the Taylor expansion yields $\dot{g}_M(\theta_o) = 0$.[6] As seen below this approach additionally depends on the large sample approximation $\hat{\theta} \approx \theta_o$, which has an error of $O(n^{-1})$ (Kass & Raftery, 1995, §4.1.2).

The Taylor expansion yields

$$g_M(\theta) \approx g_M(\theta_o) - 1/2[\theta - \theta_o]' \, \ddot{g}_M(\theta_o) \, [\theta - \theta_o] \tag{1.8}$$

[6] The dot-notation is used for the derivatives of functions when context makes clear what variables these are with respect to. This is to avoid confusion with the use of primes for matrix transposition.

where $\ddot{g}(\theta)$ is the negative of the Hessian of $g(\theta)$. Substituting this approximation into equation 1.7, $\exp\{\, g_M(\theta_o)\,\}$ can be brought outside the integral. Kass, Tierny, and Kadane (1990) explain the conditions under which Laplace's method of integrals can be employed to approximate that of the exponentiated quadratic form that remains. The basic idea is to use Aitken's integral (Searle, 1979, chap. 2). After taking logarithms, the resultant quantity is

$$\ln(L_M(x)) \approx \ln(L_M(x|\theta_o)) + \ln(f_{\theta|M}(\theta_o)) + K/2\ln(2\pi) - 1/2\ln(|\ddot{g}(\theta_o)|). \qquad (1.9)$$

The error in equation 1.9 is $O(n^{-1})$ (Kass & Raftery, 1995). Various methods for obtaining a final approximation of $\ln(L_M(x))$ are available from equation 1.9. Raftery (1995) treats $\ddot{g}(\theta_o)$ as the expected Fisher information matrix and then uses the large sample approximation $\ddot{g}(\theta_o) \approx nI(\theta_o)$ where $I(\theta)$ denotes the expected information matrix of a single observation. BIC is then obtained by dropping all terms of $O(1)$ or less.[7] In our case this leaves

$$\ln(L_M(x)) \approx \ln(L_M(x|\theta_o)) - K/2\ln(n). \qquad (1.10)$$

Substituting $\hat{\theta}$ for θ_o and multiplying by -2 yields the usual form of BIC:[8]

$$BIC(M) = -2l(\hat{\theta}) + K\ln(n). \qquad (1.11)$$

[7] Comparing Claeskens and Hjort (2008, p. 80) and Raftery (1995, pp. 131-132) proves interesting here. They both take the approach of omitting terms that are asymptotically of order 1 or less, but then drop different terms from their equations corresponding to equation 1.9

[8] Schwarz (1978) did not multiply his statistic by -2. This addition is apparently due to it's similarity with AIC.

Kass and Raftery (1995) discuss various restrictions on $f_{\theta|M}$ that reduce the order of the asymptotic errors along various steps of the derivation, especially for those of $O(1)$. Whether or not these restrictions are palatable is not of interest here.

Equation 1.11 is a biased, asymptotic estimate of the logarithm of a function of the average likelihood of a model. It is arguably much less exciting than AIC, since Bayes' factors were first popularized forty years earlier by Jeffreys (1939) and, as an estimate of Bayes' factors, BIC leaves much to be desired. Indeed, as noted by Schwarz (1978), BIC is mostly interesting in comparison to AIC, and its application as a method of model selection is identical: The best candidate is the one that implies the smallest value of equation 1.11. It is important to note, however, that 'best' is now interpreted in terms of a different objective function. In particular, Bayes's factor is usually interpreted as being proportional to the ratio of the posterior modes of two candidates. Thus AIC and BIC represent different purposes for models, one framed in terms of minimizing relative entropy and one phrased in terms of maximizing posterior probabilities. It is for this reason that AIC and BIC should be thought of as different types of approaches to the problem of model selection.

Despite their different interpretations, it is readily seen that BIC is also a relative fit index. In particular Jeffreys (1939) suggested ranges of magnitude for the interpretation of Bayes' factors, and these are by definition premised on the comparison of differences of the quantity estimated by BIC. By Jeffreys' interpretation it may be argued that, when the ratio of two model's BIC's is $10 : 1$, this gives stronger support than ratio of $2 : 1$. In general, a candidate with the smallest BIC can be viewed as the model with the most evidence in its favor. However, the modal posterior probability of that model may be small or large, and so BIC does not actually tell us whether a model is satisfactory in the sense defined by its objective function. Thus, similarly to the quantity Q, Bayes' factors can only serve purposes of ordering models or differences between them, and as with AIC, BIC must be regarded as a relative fit index.

Before summarizing these considerations, a final remark can be made concerning literature that focuses on the evaluation of AIC and BIC. On the basis of heuristics such as those discussed in connection with AIC, there seems to be a general consensus that it makes sense to evaluate the performance of different IC by means of the some further criterion. For instance, consider the work of Sin and White (1989). They show that, asymptotically, both AIC and BIC select models that minimize KL. However, when two models have equal KL, only BIC selects that with fewer parameters. Therefore BIC is regarded as 'consistent' but AIC is not. While being mathematically quite impressive, what is most remarkable about this result is that it evaluates the asymptotic properties of two statistics without any concern for what they estimate. Recall that AIC is supposed to find the model that minimizes R_{KL} by estimating the maximum value of nQ for each candidate. Accordingly, if two models reach this maximum, then this is just what AIC should tell us. It should not be expected to "favor" a model with fewer parameters and if it did we should properly conclude that it is not a robust estimate of nQ. On the other hand, if a statistic that minimizes KL and favors models with fewer parameters is desired, then presumably one wants to estimate a quantity like that described by Sin and White, not a quantity like Q. In a similar vein, the idea of evaluating BIC by how well it minimizes KL is ostensibly premised on the idea that a model should minimize KL. But as discussed above, BIC is based on the idea that a candidate should maximize posterior probabilities. It is then an interesting coincidence that BIC has the consistency properties described by Sin and White, but this should not be mistaken for the *purpose* of BIC.

That the objective functions of different IC are mixed and matched in the selection literature has arguably resulted from mystifying these statistics in terms of a variety of heuristic but non-existent quantities. An auxiliary purpose of this section has been to show that these statistics were not designed as multi-purpose, trade-off optimizing, instantiations of the principle of parsimony. As such, it would be rather surprising to find that they achieve this purpose "by accident." On the other hand, development of a set of concepts by which to describe the performance of selection statistics is indeed desirable. Yet, if these concepts are to lead to progress in the theory of model selection, they should be premised on a coherent theory of the purpose of models, not representative of differing purposes. Such an approach is taken in the following chapters.

1.2 Summary

At this point some aspects of traditional goodness of fit approaches have been briefly mentioned and the two prototypical information criteria discussed in sufficient detail to understand their basic differences and similarities. The present section makes some brief conclusions about these quantities. Here it may be noted that, although many IC have been omitted from this review, the reader familiar with these statistics will have little difficulty extending the remarks made in the foregoing sections. In short, such statistics are typically interpretable in terms of "true models" or posterior probabilities, and to the best of my knowledge, all IC fall into the category of relative fit indices. However there also exist other relatively recent approaches to model selection, approaches not based on IC. In particular, the minimum description length (MDL) principle (Grünwald, 2007; Rissanen, 2007) is based on computational theories of complexity and, in its modern applications, is founded on the notion of a so-called "universal code." This approach is both interesting and convoluted. A characterization of this principle within the context of this literature review is not feasible for a variety of reasons including the following.

1. A statement of the MDL principle requires foundational concepts from coding theory that cannot be directly translated into statistical terminology. In terms of model selection, its basic assertion is that a model that allows for a set of observations to be "optimally" encoded should be preferred, so long as the model itself can be "optimally" encoded with regard to the class of universal codes. These two notions of optimality are not identical. The overall result is a principle of shortest code length, which is itself premised on notions of data compression, regularity, and stochastic complexity.

2. The MDL principle yields multiple quantities for model selection based on the type of universal code employed. It can also be the case that multiple criteria satisfy the MDL principle for the same set of candidate models. While some of these criteria are asymptotically equivalent to BIC, this is not always the case.

3. MDL can also be applied to problems of estimation and to pre-quential prediction; the principle itself cannot be treated only in terms of model selection.

These points indicate that a proper summary of MDL is a major undertaking. Rather than falsely characterize this principle within the relatively narrow interests of this literature review, it is more appropriate to note that, although MDL is in some cases subject to the arguments made in this summary, this may not always be the case. The reader is referred to the work of Rissanen (e.g., 2007) and Grünwald (e.g., 2007).

The central limitation of tests of goodness of fit is that they are, like all statistical tests, methods of estimation. As such they have only been successfully applied to a relatively narrow range of selection problems, namely selection among correctly specified, nested models. For this reason such tests have been largely superseded by information criteria. These newer methods are readily applied to selection among relatively large numbers of competing models, regardless of the relations among their families of model-implied distributions; neither the number of candidates or the relations among them have entered into the derivation or discussion of IC. In comparison with these methods, the older tests do seem to clunk about with their heavy assumptions and narrow conclusions. But, in a certain sense, these are *better* conclusions. If their assumptions are accurate, tests of goodness of fit indicate how likely a set of observations are under a given model. This can be used to make a decision about that model, and this decision is only contingent on how likely we require the observations to be.

On the other hand, relative fit indices do not provide any standard by which to judge whether a model is "good enough." This can lead to a variety of undesirable situations. In particular, it is not hard to imagine a list of very poor candidates, where 'poor' is defined in terms of the objective function of some relative fit index. In such a case, selection based on that index will nonetheless recommend one of these candidates rather than identifying them all as inadequate. Such a forced-choice recommendation may be required in some circumstances and so it is not the recommendation *per se* that is objectionable, but the fact that we can never know if we are in this kind of situation when using relative fit indices. In effect, relative fit simply pushes the problem of model selection back onto the choice of a good set of candidates. That is, if we are to employ relative fit with any confidence, we must have some way of selecting candidates that are known to be "not that bad." In such a case, it makes good sense to select among these model on the basis of their ranking. However, if it were the case that a set of reasonable models could be established beforehand, then presumably whatever criterion by which *that* selection was made could be also used to determine the best model. And so it seems that the advocate of relative fit must walk a crooked line – such an index can always determine a best model yet it can never be used to assure us that this same model is not also bad one.

The use of relative fit indices as criteria by which to evaluate the adequacy of models requires a disagreeable compromise. As such, better solutions to the problem of model evaluation are desirable. One problematic aspect of the IC discussed here is that they cannot be used to evaluate a single, individual model. Therefore, it is desirable to consider approaches to selection that can be used for this purpose. In a sense, it would be nice to combine the interpretability of significance levels with the versatility of IC, yielding methods of selection that can be applied to any number of non-nested or misspecified models, but that also allow for the evaluation of each model in isolation. Thus we could rank models by how well they obtain some quantifiable purpose, not dissimilar in principle from KL or posterior probability, and also make absolute decisions about whether or not that purpose is met to some satisfactory degree. This is a basic idea behind the approach taken in the remainder of this dissertation.

Chapter 2

Models, Predictions, and Elimination

The definition and purpose of models in general and stochastic models in particular have long been topics of the philosophy of science (e.g., Suppes, 1960). These issues must play a foundational role in a theory of model selection and so some conceptual groundwork should be done in this area. The central question to be addressed is the purpose, or intended use, of stochastic models (§.2.2). The answer to this question provides a general criterion by which to evaluate models, and hence it allows for the formulation of a general method of selection (§2.3). In order to clearly articulate their purpose, it is helpful to have a definition of models in place, and this is the first task of this chapter (§2.1).

The overall result of this discussion is to *reduce* the purpose of stochastic models to that of prediction, and the precise meaning of the term 'prediction' is given in in definition 2.2. In particular it is claimed that a model should accurately predict those phenomena that one seeks to interpret by means of its application. It is hardly contentious to assert that a model must be accurate, however, it readily argued that this can only be a necessary condition on a model's worth. One oft-cited problem here is overparameterization or model complexity. The typical example is fitting a $(k-1)$-degree polynomial to k bivariate data points (e.g. Myung, 2000; Grünwald, 2007). Another example occurs in the case of saturated log-linear models (e.g., Christensen 1997). The *prima facie* difficulties of treating prediction as a basic criterion of model selection are also addressed in this chapter (§2.2). In short, these difficulties are treated as entirely secondary in importance. The fact that some models are trivially accurate does not imply that judging the accuracy of every model is a trivial matter. Yet this seems to be just the rhetorical strategy to which such examples are employed: By establishing that accuracy is only a necessary condition of a model's worth, we then move on to topics such as parsimony and complexity as though we had somehow gotten prediction in the bag. As discussed in the previous chapter, this mistake is evidenced by "trade-offs" that purport to select the best model from a set of candidates without being able to settle whether any of those models are particularly good. Intuitively, the strategy in anointing prediction as the *primary* or *basic* purpose of a model is to find effective methods for getting rid of models that have no empirical mettle before beginning the relatively less important and otherwise quite meaningless task of determining which to prefer.

Naturally, this strategy is always with respect to a given set of observations and a model's intended interpretation. This approach should not be mistaken as foisting unreasonable evaluation criteria on statistical models, because the question of "how accurate" a model must be is always settled in context. Yet, regardless of how this standard (or any other) is decided, it is a direct consequence that some models might not achieve it. Thus we find that by establishing accuracy as a criterion for model evaluation, we have thereby invoked a principle of elimination (§2.3). This principle summarizes the approach to selection developed in this dissertation.

2.1 A Definition of Parametric Stochastic Models

This section discusses and then provides a definition of parametric stochastic models. An interesting starting point for this discussion can be found in the work of Bamber and van Santen (1985, 2000). They develop a measure theoretic approach to model testability and identification, and their definition of models is very much in line with the approach taken in this dissertation. In addition to formalizing the notion of a model, their approach also emphasizes the fact that models make predictions. However, their definition is not directly applicable to the cases of interest here. Firstly it does not explicitly deal with stochastic models. This is addressed here by reformulating their notion of an outcome space in terms of random variables. Secondly, it is desirable to formulate models that make multiple predictions, whereas their approach treats predictions "one at a time." Despite these modifications, many of the ideas presented in this and the next section are properly viewed as a variation on their approach. This is acknowledged by adapting their terminology wherever this can be done felicitously. A comparison of the two approaches is provided at the end of the following section.

Following Bamber and van Santen (2000), a model can be defined in terms of its components. This "component-wise" definition is familiar from measure theory, where, for example, a probability space is defined in terms of the set of possible outcomes of an observation procedure, a σ-algebra defined on that set, and a probability measure defined on the σ-algebra. This type of definition is typically used to introduce technical mathematical concepts, although in this case a concept with already established meanings, 'model', is given a precise technical interpretation. That is, a particular interpretation of the concept is stipulated. The components of this definition are as follows, and the discussion and elaboration of these components is the purpose of this section.

1. *The outcome or response variable of interest.* This variable is conceptualized as random and as having an unknown probability distribution. This conceptualization serves to explicitly restrict the domain of interest to stochastic models, that is, to the statistical treatment of research variables.

2. *A family of model-implied parametric probability distributions.* This "family" is any probability distribution that is not contradicted by the specification of a model. Restricting consideration to parametric models is necessary for the approach taken in chapter 3.

3. *The possible parameter configurations of a model's probability distributions.* Parameters are not treated as random, but they are treated as variables whose domain is a model's *parameter space.* Thus a model-implied family of probability distributions can be described as the image of a function whose arguments are the model parameters.

These components allow for the problem of evaluation *qua* prediction to be stated explicitly. In this context, the problem is to assess whether or not the distribution of a set of observations is included in the family of distributions implied by a model.

The first component corresponds to Bamber and van Santen's conception of an *outcome space* of a research scenario. Intuitively, this space represents all of the conceivable realizations of the outcome variable(s) of interest in a particular research context. Otherwise stated we are here concerned with the values of the variables to be modeled. As a simple example, percent correct is an outcome variable that has found wide application in experimental psychology. Its possible values are the interval $[0, 100]$. Accordingly, in a research scenario where the outcome of interest is the percentage of correctly recognized stimulus words re-presented after some experimental manipulation, the outcome space would be the real numbers between zero and one hundred inclusive. Similar examples can be constructed for other research scenarios. In Bamber and van Santen's treatment, the outcome space defines the co-domain of a model's prediction function. On the present approach, the outcome space serves to describe the type of outcome variable of interest (e.g., discrete or continuous, bounded or unbounded).

Let a J-dimensional, real-valued outcome variable be denoted by $O = (O_1, \ldots, O_J)$. Its realizations $O = o$ represent the possible values observable in a given research context. Because we are concerned with stochastic models, the basic move is to conceptualize O as random. This is accomplished by treating O as a function or map from some probability space to $\mathbb{R}^J$, with a probability distribution (e.g., a discrete mass function or a density function) defined through its probability measure. A probability measure can more naturally be used to define the cumulative distribution function (c.d.f.) of a random variable, and because there is a one-to-one correspondence between c.d.f.s and probability measures, there is little need to consider probability spaces themselves so long as one is content with a foundation in cumulative probability. However, the use of probability density functions and discrete mass functions of outcome variables is more convenient for the present analysis. Following the measure theoretic terminology these will be collectively referred to as probability distributions, and are denoted by p_{O}. For continuous random variables this approach requires that the densities of interest are the Radon-Nikodym derivative of some c.d.f., although for the discrete case the cumulative distributions are defined through counting the mass points (Billingsley, 1986 provides more discussion of these topics; Cohn, 1980 provides a good introduction to measure theory).

The notion of an outcome variable serves to make explicit that a model is a model *of something*, for example, of percentages of correctly recognized words under a given research context. Treating these outcomes as random introduces a basic framework for their analysis, however, this is not taken to imply that an adequate description of their stochastic behavior is available in any given case. This point is central to the non-analytic or contingent character of models. For this reason it is important to keep in mind that the term 'outcome variable' is shorthand for 'the possible outcomes of a research scenario' and its purpose is just to represent research outcomes and their associated probabilities. Although O is one of the components by which a model is to be defined, it would be misleading to interpret it as 'the possible outcomes of a model,' since there is no assumption that any of the probability distributions specified by a given model correspond to that of O.

It is worth emphasizing a related distinction here, that between stochastic and non-stochastic interpretations of research variables. Traditional approaches to statistical modeling "tack on" the stochastic properties of models in a manner that is prone to misinterpretation. This misinterpretation can also be found in many current treatments of model selection, for example when curve fitting is treated as being of a kind with linear regression (e.g., Bamber & van Santen, 2000; Forster, 2000; Grünwald, 2007; Linhart & Zucchini, 1986; Myung, 2000). The point to be made in the following paragraphs is that the outcomes of stochastic models are not the values of research variables but the probabilities of those values, which is why those variables have here been conceptualized as random.

By way of comparison with the approaches mentioned above, let us consider the example of simple linear regression. There are two mathematically equivalent ways specifying this model: 1) Writing the outcome variable as a linear function of its predictors and residual, with the distributional properties of the terms of this function stated separately; 2) stating the distribution of the outcome variables as a function of the predictors. The first approach, which is the more traditional (e.g., Searle, 1971), has the advantage that it makes clear the motivation of the model in terms of the data. However, on this approach the stochastic properties of the model are treated as assumptions or riders, when in fact they constitute its non-trivial character. For instance, it is a tautology to claim that an outcome variable is one part prediction and one part prediction error. Although the "distributional assumptions" of the regression model may be of little intrinsic interest to the applied researcher, they are nonetheless required for the model to be interesting.

By specifying the model in terms of the distribution of the outcomes, one obtains a more integrated representation. On this approach, the model is specified by an expectation function together with assertions about other parameters and the shape of the conditional distributions. This defines a family of probability distributions, and thus the model is better interpreted as making assertions about the probabilities of the variables of interest rather than the variables themselves. From this perspective the outcomes of a regression model are properly conceptualized in terms of probability distributions, or equivalently, in terms of random variables. The same point holds for those models of interest in this manuscript: In IRT the probabilities of response patterns are modeled rather than response patterns themselves. These probabilities define a multinomial distribution, not a particular set of observations.

The foregoing has sought to clarify that the possible outcomes of research variables are not equivalent to the possible outcomes of stochastic models. In the latter case we are more directly interested in probability distributions. Following Amari (1985), a family of parametric probability distributions can be viewed as a set $S = \{\, p(x; \theta) \,\}$, where x is a realization of a real, possibly vector-valued random variable X, p is the probability (density or discrete mass) function of X, and θ is a K-dimensional real-valued parameter of p. For example S might represent all of the univariate normal distributions, or all bivariate normal distributions with covariances matrices equal to $\sigma^2 I$, or all 2-parameter logistic IRT models. The family of probability distributions that are defined in the specification of a model will be referred to as its *model function* and will be denoted either by S or in terms of its elements $p = p(x; \theta)$. Additional notations for p are introduced below.

Returning to the example of simple linear regression, the model function for a random sample of $i = 1, \ldots, n$ observations can be written in its conventional vector notation as follows: $\mathbf{y} \mid \mathbf{x} \sim N(\boldsymbol{\mu}(\mathbf{x}), \sigma^2 I)$ where $\boldsymbol{\mu}(\mathbf{x}) = X\boldsymbol{\beta}$, $X = [\mathbf{1}\ \mathbf{x}]$, and $\boldsymbol{\beta} = [\beta_0, \beta_1]'$. In IRT, a general form of the model function is given by equation 3.2.

It is important to distinguish the single probability distribution of the outcome variable, p_O, from the family of probability distributions given by a model of O. As noted above, it is not assumed that $p_O \in S$. In the sequel it is useful consider X as a dummy variable. its values hold a place in $p(x;\theta)$ to which a given realization $O = o$ may be assigned during the process of evaluating the adequacy of a model. Thus we may, for example, talk about the likelihood of a set of realizations of O under various models without assuming that any of these is the "true model," a notion which is scrutinized at the end of this chapter.

Thirdly, define the *parameter space* of a model. Letting $\theta = (\theta_1, \ldots, \theta_K)$ denote the real-valued, K-dimensional parameter of a model M, the subset of $\mathbb{R}^K$ for which θ is defined is called its parameter space. Letting u denote an arbitrary point in $\mathbb{R}^K$, this may be written as

$$\Theta = \{\, u \mid u \textit{ is a parameter of } M \,\}.$$

Again considering the example of a simple linear regression model, the parameter space is

$$\Theta_L = \{\, (\beta_0, \beta_1, \sigma^2) \mid -\infty < \beta_j < \infty,\ 0 < \sigma^2;\ j = 0, 1 \,\}.$$

Similarly, the Rasch model from IRT has a real-valued "item parameter", say β_j, for each of $j = 1, \ldots, J$ items, and also a set of parameters defining the distribution of the latent variate. In the case that the distribution of the latent variate is normal with mean μ and variance σ^2, the parameter space of the J-item Rasch model may be written as

$$\Theta_R = \{\, (\beta_1, \ldots \beta_J, \mu, \sigma^2) \mid -\infty < \beta_j, \mu < \infty,\ 0 < \sigma^2;\ j = 1, \ldots, J \,\}.$$

For each $\theta \in \Theta$ the model function assigns a value $p \in S$. This is made explicit by writing $p(\theta)$, and when it is important to distinguish the probability distributions of M from some other distributions the notation $p_{\mathrm{M}}(\theta)$ is employed. The values of $p(\theta)$ are the various parameterizations of a model's probability distributions and $p(\Theta) = S$. That is, a model's family of probability distributions is equivalently represented as the image of its model function in θ.

The foregoing discussion is summarized in Definition 2.1.

Definition 2.1 A model M is an ordered triple (O, p, Θ) consisting of an *outcome variable*, O, the *model function*, p, belonging to M, and the *parameter space* of p, $\Theta \subseteq \mathbb{R}^K$

This definition can be related to other interpretations of models. As presented above, a model is defined over a range of possible parameter arrangements. In some literatures this is referred to as a family of models, with the term 'model' being reserved for its particular parameterizations (e.g., Linhart & Zucchini, 1986). When this distinction proves relevant to the current work, it is made by referring to definition 2.1 as a *general model* and a subset of its parameterizations as a *specific model.* The present work is concerned almost exclusively with general models, the strategy being to consider, insofar as possible, *all* of the implications of a model with reference to the potential observations on a research variable. It should be noted that this distinction is not the same as that made by Bamber and van Santen (2000) under similar terminology. Two further interpretations of definition 2.1 depend on whether we are concerned with O or one of its realizations. In the case that $O = o$ is fixed and θ is variable, this definition yields the likelihood function of a model. In cases where both o and θ are fixed, this will be called a model realization. A realization

Table 2.1: Four Interpretations of Definition 2.1

	Fixed θ	Variable θ
Fixed o	model realization	likelihood function
Variable o	specific model	general model

of a model can be represented, for example, in terms of the numerical values of the parameters corresponding to a given set of observations or in terms of the likelihood function evaluated for a given value of θ. These considerations are summarized in Table 2.1.

2.2 Prediction as the Purpose of Models

As noted at the outset of this chapter, it is not intended that we discover the purpose of models by an examination of definition 2.1, but that this definition be of service in clearly articulating their purpose. In this section this purpose is conceptualized in terms of a formal definition of prediction. This definition is interpreted and some remarks are made concerning its role in this approach to model selection.

A model has been conceptualized in terms of its model function, the possible parameter arrangements of that function, and a random variable representing the outcomes of a research scenario. In this section the first concern is to conceptualize when a model can be said to accurately predict an outcome variable O. The basic challenge is to obtain a mechanism equivalent to Bamber and van Santen's (1985, 2000) *prediction function*, a function that relates the outcomes of a research variable to the parameter space of a model. Although it is tempting to regard $p(\theta)$ as fulfilling this role, $p\colon \Theta \to S$ is function from the parameter space of M into the space of its possible probability distributions. By contrast, O is a function from its probability space to its possible realizations in $\mathbb{R}^J$. It is only this latter space that grounds a model to empirical observation. Otherwise stated, a model function defines a class of random variables, and the present concern is to evaluate a model in terms of the possible realizations of a random variable that may or may not belong to that class.

To this purpose it is useful to consider functions of $p_{_M}$ that can be written in terms of θ but which are not explicitly given by the components of θ. Obvious candidates here are the moments of $p_{_M}$ or of probability distributions that can be derived from $p_{_M}$. A trivial example is given by the case of simple linear regression discussed in previous section. For any fixed value $\mathbf{x} = \mathbf{x}^*$, $E(\mathbf{y} \mid \mathbf{x}^*)$ is implied by the regression parameters, but the conditional expectation itself is not contained in θ. Treating $E(\mathbf{y} \mid \mathbf{x}^*)$ as a function of θ, say $\mu(\theta)$, it can be seen that $\mu(\Theta) \subseteq \mathbb{R}$. We may then ask whether $E(O \mid \mathbf{x}^*) \in \mu(\Theta)$. If this is the case, it will be said that the regression model accurately predicts the outcome $E(O \mid \mathbf{x}^*)$ or that the outcome is consistent with the model. This is the basic idea to be pursued in this section.

As stated, the linear regression example given above is trivial since in this case $\mu(\Theta) = \mathbb{R}$. This is gotten around easily enough, for example by considering conditional expectations for multiple fixed values of $\mathbf{x}$ as a vector-valued function of θ. Nonetheless, the example raises the problem of assessing the quality of predictions, a topic that is discussed in the following section. Another concern here is the feasibility of implementing such predictions with sample-based methods. Of course any selection procedure can be judged by how convenient it is to apply, and problems of application are considered in chapter three. The present focus is merely to conceptualize what a prediction is and consider its role as the basic purpose of stochastic models.

The idea behind the above example is to evaluate the "same" function in two different ways, once through the parameters of the model (i.e., to derive it), and once through the random variable O (i.e., to compute it on p_O or, in practice, to estimate it from realizations on O). The moments of a random variable are good examples of such functions, but there is no reason to restrict consideration to moments in the conventional sense. In particular, a model can have a lot of implications, many of which may not be anticipated in a given application, but which may nonetheless be useful for determining the adequacy of the model. These implications can be conceptualized in terms of functions that depend on a probability distribution. In particular, two classes of functions are currently of interest, those with either random variables or parameters as their arguments. As described in the following two paragraphs, "sameness" is conditional functional equality, where the condition or proviso is functional equality of distribution.

Any two functions f and g computed on the random variables X and Y respectively are called the *same* if and only if $f = g$ when $X = Y$. The expression $X = Y$ means the random variables are identical maps from their shared probability space to Euclidean space. Any two functions f and g satisfying this requirement are called a *data function* of a random variable. They are not called 'data function*s*' because they are the *same* function. For example, the data functions that are of primary interest in the present approach are operator functions on random variables, such as the expectation or covariance of X or of random variables computed on X. Clearly these operator functions satisfy the requirement of sameness. Also, any random variable computed on X trivially satisfies the condition of sameness since the functional equivalence of random variables is not dependent on their probability distribution. For example if

$f = g = (\cdot)^2$, this does not depend on the argument $(\cdot)$. For the present level of analysis there is little to be lost in slurring over the distinction between the expectation of a random variable X and expectations of random variables computed on X, and this is one of the conveniences of the notion of a data function. As explained below in definition 2.2, the type of data functions of interest in the present approach are "population level" functions.[1] Also note that two data functions $f(X)$ and $g(Y)$ can have equal values for all or any $X = x$ and $Y = y$ without implying that $X = Y$. For instance two random variables can have the same mean without being the same variable – but this is the *same* function computed on those variables.

Next recall that a model function has been defined as family of probability distributions with a parameter θ defined over $\Theta \subseteq \mathbb{R}^K$. Consider the parameters θ and θ' of two families of probability distributions $p(\Theta)$ and $p'(\Theta')$ respectively, and also the functions f and g taking arguments $\theta \in \Theta$ and $\theta' \in \Theta'$ respectively. These functions are called the *same* just in case $f = g$ when $p(\Theta) = p'(\Theta')$ and in such cases these will be refered to as a *prediction function*. Note that if $p(\Theta) = p'(\Theta')$ and f ad g are the same, then $f(\theta) = g(\theta')$ when $\theta = \theta'$. For example we may write the moments of two families of probability distributions in terms of their parameters using the moment generating function (m.g.f.) of each family. Then if we are considering the *same* moments, f and g, those moments are functionally equal (i.e., $f = g$) when we consider identical families of distributions, and these functions numerically equal (i.e., $f(\theta) = g(\theta)$) when we consider the same member of that family. The principal difference between data functions and predictions functions is just the arguments they take – the former take random variables, and the latter take parameters.

[1] Arguably this is reason to use the name 'data constants' instead of 'data functions,' however, these constants vary depending on the transformations applied to X.

Definition 2.2 formalizes the concept of prediction in terms of data functions and prediction functions and introduces some associated terminology. This definition is based on a third application of the notion of sameness.

Definition 2.2 A *prediction* P of a model (O, p_{M}, Θ) is an ordered triple (g, X, f) such that $g\colon \mathbb{R}^J \to X$, $f\colon \Theta \to X$, and, for each $\theta \in \Theta$, $g(O) = f(\theta)$ when $p_O = p_M(\theta)$. The set X is called a *prediction space* and $f(\Theta) \subseteq X$ is called a *prediction range*, or later, a *prediction manifold*. If $g(O) = f(\theta)$ for some $\theta \in \Theta$, then O is said to be *consistent* with the prediction, or, equivalently, the prediction is said to be *accurate* with respect to O. M is accurate with respect to O if all of its predictions are.

The requirement that $g(O) = f(\theta)$ when $p_O = p_M(\theta)$ ensures that the "same" function is considered on $\mathbb{R}^J$ and Θ. For example, if the prediction function is the first moment of p_M the data function must be the first moment of p_O. Aside from this obvious restriction, a prediction is just any function of a model's parameters that can be related to a corresponding function of the data. More specifically, the terminology of definition 2.2 suggests the comparison of $g(O)$ and $f(\Theta)$ and hence it may be equally well thought of as describing a *test* of a prediction. In particular, if one of the possible values of a prediction function is equal to the value of the corresponding data function, the outcome variable is consistent with the prediction and the prediction is accurate with respect to the outcome variable. If an outcome is not consistent with a prediction, it is contradicted by that prediction. That is, if $g(O) \neq f(\theta)$ for all $\theta \in \Theta$ then a consequence of O, namely $g(O)$, is contradicted by an implication of M, namely $f(\Theta)$. Thus we say that the prediction, and hence M, is inaccurate with respect to that outcome. If a model is inaccurate this means, for all $\theta \in \Theta$,

$p_O \neq p_M(\theta)$. Otherwise stated $p_O \notin S$, and as described below, M is to be regarded as inadequate to its purpose on these grounds.

It is also important to note that this definition implies that a model can make an uncountable number of predictions, many of which will be trivial or otherwise uninteresting. For example let $f = c$ for some constant c. Then $\mathcal{F} = \{\, c{+}r \mid r \in \mathbb{R} \,\}$ is an uncountable set prediction functions yielding an uncountable number of predictions for any model M, all of which are meaningless. In the following section predictions that are useful for model selection are defined. For now it may simply be observed that an outcome variable is only said to be consistent with M if all of M's predictions are accurate.

Although it is "too broad," this definition of prediction does adequately describe intuitive cases. For example, when one estimates the higher-order moments of a set of observations in order to assess the assumption of i.i.d. normality, this can be phrased exactly in terms of definition 2.2. The skewness of a normal distribution is known to be zero from its m.g.f., and its m.g.f. can be treated as a function of its parameters. On the basis of this knowledge of *all* normal distributions, we can make the *prediction* that any set of observations that is normally distributed will have zero skewness "in the population". If we then estimate the skewness of the data, for example using a confidence interval, and it is inferred that the population skewness is not different from zero, it is natural to say that the observations are consistent with the prediction of zero skewness. If not, the model should not be regarded as accurate. The example serves to illustrate that definition 2.2 is in fact very intuitive, and the generality of the definition allows us to apply these intuitions to situations that are otherwise unintuitive. In particular, how can we tell if an IRT model accurately predicts the stochastic properties of a set of response patterns? Regarding his work on differential item functioning, Dr. P. W. Holland humorously described this situation along the following lines:[2] We have a good idea of what normally distributed data are "supposed to look like," but what about "IRT distributed" data — do we have any idea what to look for here? Definition 2.2 tells us what to look for in the general sense of what a prediction is "supposed to look like."

The intention of conceptualizing prediction has been to describe the purpose of models, thereby establishing a general criterion for model selection. Whence it is claimed that the basic purpose of a model is to make accurate predictions, and if a model fails to make accurate predictions it is inadequate to its purpose. Stipulating the purpose of models in this manner has the advantage of clarity and motivates the course to be taken in sequel. At the same time it is recognized that many objections have been raised against the use of prediction as a criterion by which to judge the worth of models (e.g., Myung, 2000), scientific theories (e.g., Grasa, 1989), and knowledge in general (e.g., Goodman, 1954). This stipulation therefore requires some initial justification, although its ultimate defense is to be found in the theory of selection of which it is the foundation. The remainder of this section also serves to further elaborate various aspects of definition 2.2.

It may be noted at the outset that there has not been, to the best of my knowledge, any serious objection to the idea that accuracy is a necessary condition for model adequacy. As it has been conceptualized here, prediction is just a certain type of mathematical consequence of a model, and if a consequence is wrong then, by pain of contradiction, so is whatever implied it. However, at least since the heyday of

[2] Personal communication, July 2008

Popperian falsification, the idea that prediction is a *sufficient* criterion for scientific worth has been deemed woefully inadequate. Goodman's paradox is perhaps the most famous example of this, where one is forced to defend the induction that all emeralds are green against the apparently ridiculous but equally accurate claim that they are all "grue." The general consensus is well put by Grasa:

> Conformity with facts is a necessary but not sufficient condition when preferring one theory over others. (Grasa 1989, p. 13).

What then is the justification for treating prediction as the basic purpose of stochastic models? Are we not flying in the face of 50 years of rigorous thinking on this matter? To clarify this situation it is important to note that it has not been argued that accuracy is the *only* purpose of a model, but that it is the basic or first purpose. We may otherwise state Grasa's meaning by saying that a model is not adequate if it is not accurate, and this is just what is meant by 'basic' or 'first'. Thus one advantage of a theory of selection based on prediction is that it is widely applicable – it is reasonably applied to all models that claim to be "in keeping with the data." With this breadth comes the limitation that this criterion is not sufficient to evaluate every purpose to which a model may be applied. For example, the accuracy of a model does not imply that its parameters have a meaningful interpretation with respect to the research scenario in which the observations were generated, or that it is useful to explain or represent those outcomes in terms of the model, or even that "repeat applications" of the model (i.e., models with different outcome variables) will yield accurate predictions. Moreover, this doesn't even imply that its predictions are of any importance. The point stands, however, that there is little reason to discuss any of these matters if a model is not accurate. The basic idea here is to identify an appropriate domain for these more refined purposes, namely accurate models.

Taking another perspective, it may seem that this requirement of accurate predictions is far too stringent. Here we refer to that old dogma, touted whenever somebody seems to get too serious about the worth of a stochastic model: “All models are false but some are useful.” Shall we concede that a model can be useful even though it is inaccurate? Is such a position compatible with that discussed in the previous paragraphs? The answer to both of these questions is in the affirmative. The concession is made in two ways, firstly by recalling that a model makes *lots* of predictions. In order for it to be accurate, all of these must be accurate. But if the intended use of a model only involves a few of its predictions, there is no reason to require that all the other ones are accurate also. This would be like firing a meteorologist for bad business advice. We should only properly be asking him for advice about the weather, and so long as he is doing alright with those predictions, we should regard him as an adequate meteorologist. So with our “useful” models.

To clarify this point, it has been stated that a model’s purpose is to make accurate predictions. However, a model makes a lot of predictions, and only a subset of these may be of direct relevance to a given application of the model. In such cases, the purpose of a model is to be accurate over a user-defined subset of predictions. Otherwise stated, a model is not useful if it does not accurately predict those functions of the data which are of relevance to a given application. This is a utilitarian, rather than epistemological, interpretation of the criterion of accuracy.

Secondly, although accuracy has been described as a qualitative aspect of predictions, a quantitative reformulation is given in chapter three. Thus we may speak of models being more or less accurate, and consider how much accuracy we require in a given situation, for example, on the basis of loss functions describing that particular situation. Perhaps the most intuitive way to do this is just by counting accurate predictions. So long as some proportion of a model's predictions are accurate, then, as with our weatherman, this may be viewed as good enough. The method that is pursued in chapter three is to consider the distance of $g(O)$ to $f(\Theta)$. In the case that $g(O)$ is vector-valued this is the norm of its projection onto $f(\Theta)$, and the meaningful standardization of this metric presents an interesting problem in each case. The general point is that accuracy can be thought of in degrees and then some degree which is sufficient to one's purposes can be stipulated. With regard to the method of model selection described in the next section, how this stipulation is made is irrelevant.

From these considerations it seems that even the most pragmatic of modelers should admit the criterion of accuracy as reasonable. At the same time, a judicious interpretation of this criterion should not lead to mistaken conclusions about the worth of a model for purposes "higher" than prediction. To my mind, it would be fruitless to pursue the interpretation of prediction much further than this. This discussion has been intended to facilitate a coherent and therefore limited theory of model selection. As argued in this section, these limitations are not so much the breadth but the depth of its application: It covers the shallow waters of pragmatic modeling and keeps "sinkers" out of the deep end.

The main difference between this approach and that of Bamber and van Santen (1985, 2000) is that the models presented here make lots (i.e., an infinity) of predictions. On their approach a model is always specific to an experimental design, and it is therefore constituted by a single prediction. Also, their approach is not explicitly stochastic. At root, however, the prediction functions defined here are similar to the prediction functions described in their work. For this reason much of the terminology employed so far has been adapted from their approach, and this is only proper. In the sequel, the divergence between these approaches becomes much sharper. As noted, they deal with topics of model identification and a measure theoretic interpretation of the testability of a single model. The present work turns to consider how the notion of prediction can be utilized for selection among a set of ≥ 1 candidate models, thus bringing us closer to the main currents of contemporary model selection.

2.3 The Principle of Elimination

Having defined models and their predictions the task now is to consider how these elements can be combined into a method of model selection. It is useful to begin by adapting the notation of the previous sections to explicitly incorporate multiple models and multiple predictions. This allows for a statement of a "second-order" selection problem, the solution of which is described in terms of *efficiency* in the computational sense of time-to-completion. This solution leads to a general algorithm for model selection in the case of ≥ 2 candidates. Consideration of this algorithm provides two related rationales for the framework developed in this chapter. 1) It demonstrates a particularly effective method of prediction-based selection. This motivates finding the kinds of predictions described below in definition 2.3, namely those that cannot be accurate for all candidates. This is an *ad hoc* kind of motivation since it depends on getting the results. 2) More fundamentally, this section demonstrates that thinking about selection in terms of prediction can lead to good solutions to the problem of model evaluation. Rather than concerning any particular type of prediction, this motivates the framework itself. Otherwise stated, this motivation does not depend on "getting good solutions" but on defining problems that would be fruitful to solve.

The distinction between these rationales is important for the interpretation of the selection criteria defined in this section, because, as explained below, their existence requires assumptions that will not hold in many circumstances. The selection algorithm on which they are premised is therefore better interpreted as an *idealized* selection procedure. It serves as a standard for evaluating selection procedures in general. That is, if the assumptions of the selection algorithm hold, we can ask how well alternative selection methods approximate its performance. This role is not unlike that of a "true model" when this model is assumed to be contained in a set of candidates and selection statistics are evaluated by whether or not they identify that model (e.g., Claeskens & Hjort, 2008, chap. 4; Gr üwald, 2007, §17.1). However, the algorithm presented in this section does not depend on the assumption of a true model and therefore it can be thought of as embodying a "new" principle of selection. This is here termed the *principle of elimination*, and it can be interpreted as an instantiation of Platt's (1964) notion of strong inference or as otherwise reworking Popper's theory of falsification. The chapter ends by describing how this principle can guide prediction-based selection in cases where the assumptions of the selection algorithm do not hold. In particular, it is discussed how the principle applies to the case of a single model.

2.3.1 A Rudimentary Selection Algorithm

A method of model selection is premised on a set of candidate models $\mathcal{M} = \{\, M_\gamma \mid \gamma \in \Gamma \,\}$ where Γ is a set of model indices (e.g., integers). Each $M_\gamma = (O, p_\gamma, \Theta_\gamma)$ has the same outcome variable O, and p_{M_γ} is abbreviated to p_γ. The predictions of a model are denoted by $\mathcal{P}_\gamma = \{\, P_\gamma^\alpha \mid \alpha \in A_\gamma \,\}$ where A_γ is a set of indices (e.g., reals) over the predictions of each M_γ and $P_\gamma^\alpha = (g^\alpha, X^\alpha, f_\gamma^\alpha)$. The data functions g^α are not subscripted because O is the same for each M_γ and g^α is a function on O that only depends on p_O. Since $g^\alpha \colon \mathbb{R}^J \to X^\alpha$ this implies that X^α is also constant over models. Thus P_γ^α only varies over γ through its prediction function f_γ^α.

Next let us introduce the following naming convention for the prediction indices $\alpha \in A_\gamma$ over a set of candidate models. This convention is analogous to the restriction placed on prediction functions and data functions. It states that if two predictions have the same superscript then they are the "same," which means that the predictions are equal whenever the models are. This allows us to refer to predictions over models and requires the axiom of choice. For a set of candidate models $\mathcal{M} = \{\, M_\gamma \mid \gamma \in \Gamma \,\}$ such that $\gamma, \gamma' \in \Gamma$, $\alpha \in A_\gamma$, and $\alpha' \in A_{\gamma'}$, write $\alpha = \alpha'$ if and only if $P_\gamma^\alpha = P_{\gamma'}^{\alpha'}$ when $M_\gamma = M_{\gamma'}$. Thus, by convention, $P_\gamma^\alpha = P_{\gamma'}^\alpha$ when $M_\gamma = M_{\gamma'}$ and if $P_\gamma^\alpha = P_\gamma^{\alpha'}$ then $\alpha = \alpha'$. Similarly to the above applications of "sameness", the basic idea is that the predictions of two different models have identical superscripts when those predictions are about the same quantity (e.g., a regression function, a covariance matrix, marginal probabilities, and so on). Using this naming convention, consider the set $\mathcal{P}^\alpha = \{\, P_\gamma^\alpha \mid \gamma \in \Gamma \,\}$. It contains $|\Gamma|$ predictions that are all equal whenever their models are equal. Because each prediction in $\mathcal{P}^\alpha$ has the same superscript, this can be dropped for parsimony of notation when referring to the predictions or their components, so long as the reference to $\mathcal{P}^\alpha$ is clear. As an extension of the terminology used in previous section, $\mathcal{P}^\alpha$ will be called a 'prediction' of $\mathcal{M}$, and when $|\Gamma| = 1$ this is just how the term was used above.

As another extension of the terminology of the previous chapter, a candidate M_γ

is said to be ‘accurate’ with respect to the outcome variable O if for all $P_\gamma^\alpha \in \mathcal{P}_\gamma$ and for some $\theta_\gamma \in \Theta_\gamma,\ g^\alpha(O) = f_\gamma^\alpha(\theta_\gamma)$. In many circumstances we will be content with M_γ that contain a distribution sufficiently similar to p_O, which amounts to choosing a subset of predictions $A'_\gamma \subset A_\gamma$. In both cases we are faced with the following selection problem: For a given O with unknown probability distribution and a given $\mathcal{M}$, how to best go about determining whether the predictions $\mathcal{P}_\gamma$ are accurate with respect to O for at least one $\gamma \in \Gamma$. In the general case this task is non-trivial because $|\Gamma|$ may be large, and for each $\gamma \in \Gamma$, A_γ is uncountable. Otherwise stated, we might have a lot of candidates and they each make a lot of predictions. Consequently it is not obvious which predictions should be of interest for the purposes of model selection.

It is useful to conceptualize this question of “which predictions” in terms of a second-order selection problem. If the first-order problem is selection among the M_γ, the second-order problem is selection among the $\mathcal{P}^\alpha$. In the remainder of this section a solution to this second-order selection problem is posed, which in turn yields a method of model selection. As with our first-order problem, this solution is premised on an appropriate purpose. The basic purpose in this case is rather more straightforward than for the models themselves: A selection procedure must *end*. Using a finite number of predictions, either some or none of the $M_\gamma \in \mathcal{M}$ are to be identified as sufficiently accurate with respect to O. A natural extension of this rationale leads to a preference for predictions that end the selection procedure most quickly. This preference applies equally well to cases where interest is restricted to a subset $A'_\gamma \subset A_\gamma$. For example, if we could decide among the candidates on the basis of computing one or two $g^\alpha(O)$, this would be nice.

As with the definition of a model’s predictions, this second-order purpose is not novel – it is simply that of *efficiency*, in the computational sense of time-to-completion. In this case we are concerned with the time-to-completion of the selection problem stated above. In the following I formulate a selection algorithm (in very humanistic pseudo-code) for $2 \leq |\Gamma| < \infty$ that is $O(\log_2(|\Gamma|))$ efficient. This allows for a consideration of how predictions can be useful for model selection. In particular, the basis for solving the second-order selection problem is to find predictions that are accurate for some models and inaccurate for others. A type of prediction that always has this property is described in the following definition.

Definition 2.3 For a set of candidate models $\mathcal{M} = \{\, M_\gamma \mid \gamma \in \Gamma \,\}$, a prediction $\mathcal{P}^\alpha = \{\, P^\alpha_\gamma \mid \gamma \in \Gamma \,\}$ is called a *selection criterion* if there exists a subset $C^\alpha \subset X^\alpha$ and $\gamma, \gamma' \in \Gamma$ such that $f^\alpha_\gamma(\Theta_\gamma) \subseteq C^\alpha$ and $f^\alpha_{\gamma'}(\Theta_{\gamma'}) \subseteq X^\alpha - C^\alpha$. In the usual sense of the term, C^α is called a 'partition' of X^α. Denote the set of all selection criteria of $\mathcal{M}$ as $\mathcal{C}$. If $\mathcal{C} \neq \emptyset$ then the $M_\gamma \in \mathcal{M}$ are said to be *distinct*.

Roughly speaking, a selection criterion is a prediction that cannot be accurate for all candidates. For such $\mathcal{P}^\alpha$, either $g(O) \in C$ or $g(O) \in X - C$. If $g(O) \in C$ then any M_γ such that $f_\gamma(\Theta_\gamma) \subseteq X - C$ is inaccurate with respect to O and by construction of C there is at least one such γ. A similar circumstance holds for the case where $g(O) \in X - C$. Intuitively, C can be thought of as a "bubble" that encompasses the image of some prediction functions and excludes that of some others. Those within C may be nested, overlapping, or disjoint, and similarly for those without. Thus the only restriction placed on the $f_\gamma(\Theta_\gamma)$ is that some must be in C and some not. If such a C cannot be found then it is trivial to show that none of the $f_\gamma(\Theta_\gamma)$ are disjoint.[3]

[3] Assume that two $f_\gamma(\Theta_\gamma)$ are disjoint and let C equal either one of them. This is the simplest example of a selection criterion.

In such a case it would be possible that $g(O) \in f_\gamma^\alpha(\Theta_\gamma)$ for all γ. Such a prediction is not a selection criterion.

As an illustration consider again the sense in which skewness is a prediction of a distribution. Let O be a real-valued (i.e., univariate) random variable and let $\mathcal{M}$ denote a set of candidate models of O. By a didactic abuse of notation let $\mathcal{P}^{skew}$ denote the skewness predicted by the set of candidates. In this case $X = \mathbb{R}$ and let $C = \{0\}$. For any γ such that the family of probability distributions of $M_\gamma \in \mathcal{M}$ has the property of symmetry, $f_\gamma(\Theta_\gamma) = C$. For example, this is the case if p_γ is a family of normal distributions. Similarly, for any γ whose family of probability distributions is asymmetrical, for example a family of Poisson distributions, $f_\gamma(\Theta_\gamma) \subseteq X - C$. For any candidate whose family of distributions includes both symmetrical and asymmetrical cases, for example the non-central t-distributions, which are necessarily symmetrical when the non-centrality parameter is equal to zero but not otherwise, these models have prediction ranges that are subsets of neither C nor its complement. If there are not two candidates $M_\gamma \in \mathcal{M}$ such that one of their prediction ranges is a subset of C and one of them a subset of $X - C$, then $C = \{0\}$ does not satisfy the requirement of definition 2.3. If there is no subset of $\mathbb{R}$ that is disjoint with some but not all of the $f_\gamma(\Theta_\gamma)$ then $\mathcal{P}^{skew}$ is not a selection criterion of $\mathcal{M}$.

Another interesting example of selection criteria can be found in Halpin and Maraun (under review). Here the notion of conditional association (Holland & Rosenbaum, 1986) is applied to the problem of selecting between linear factor models (LF) and latent profile models (LP; Bartholomew & Knott, 1999, pp. 153 - 156;). In particular, when a subset of outcome variables, say Y, are conditioned (i.e., regressed) on a positive function of the remaining variables, say X, LF implies that the covariance matrix of Y is constant over X, whereas LP implies that this function is nonlinear. This result can be used to formulate a variety of selection criteria between LK and LP, for instance tests of equality of covariances of Y across ranges of X. Note that this is just the application of differential item functioning (DIF) to continuous rather than discrete outcome variables. As such, the Mantel-Haenszel test originally suggested by Holland and Rosenbaum (1986) can also be re-cast as a selection criterion. This may be unobvious for two reasons: 1) DIF is used to select *data* rather then models, and 2) no alternative candidates are specified in DIF analysis. Regarding the former, the difference between model selection and data selection is contingent on whether the model or the data is viewed as normative. This was discussed in the introductory chapter and in both cases we are essentially concerned with comparing model implications to properties of the data. Regarding the latter, the work of Halpin and Maraun (under review) can be extended to the case of discrete outcome variables to show that finite mixtures of IRT models are not conditionally associated (see Maraun, Slaney, & Goodyn, 2003, for the case of a two class mixture with dichotomous outcome variables). Thus, although DIF is used as a technique for selecting test items with a single class rather than multiple classes of respondents, it can also be interpreted as a selection criterion between "homogeneous" IRT models and finite mixtures of IRT models. Some of Mokken's (1971) results on non-parametric IRT can be similarly recast when a set of candidates is made explicit and the relevant quantities are derived under the alternative models. Many of the classic derivations in factor analysis, for example Spearman's results on triads and tetrads of correlations, can also be interpreted as selection criteria that have not been explicitly phrased with regard to a set of candidates. Thus, as with the definition of prediction in the previous section, the definition of selection criteria should not be regarded as a "new" idea so much as a novel formulation of a familiar idea.

The following points about $\mathcal{C}$, the set of selection criteria of a set of candidate models, are important to note before proceeding. Firstly, if $\mathcal{C}$ contains all of the selection criteria of $\mathcal{M}$ it also contains all of the selection criteria of any subset $\mathcal{M}' \subseteq \mathcal{M}$. More obviously, if P^α is a selection criterion for any two $M_\gamma, M_{\gamma'} \in \mathcal{M}$, then $P^\alpha \in \mathcal{C}$.

Secondly there are familiar cases where no selection can exist, for example when the $M_\gamma \in \mathcal{M}$ are nested. More generally, the existence of selection criteria for any two candidates requires that $p_\gamma(\Theta_\gamma) \cap p_{\gamma'}(\Theta_{\gamma'}) = \emptyset$. Otherwise there exist $\theta \in \Theta_\gamma$ and $\theta' \in \Theta_{\gamma'}$ such that $p_\gamma(\theta) = p_{\gamma'}(\theta')$. This implies that, for all α, $f^\alpha_\gamma(\theta) = f^\alpha_{\gamma'}(\theta')$, since these are just the same functions computed on identical distributions. Otherwise stated, selection criteria can only exist for families of probability distributions that are disjoint. This is a severe restriction on the domain of application of definition 2.3, but, as we shall see, this restriction allows for a nice solution to the selection problem.

Lastly, the C^α are not unique for a given selection criterion $\mathcal{P}^\alpha \in \mathcal{C}$, but they are finite when $|\Gamma|$ is. This can be seend as follows. Denote by C^α_i, $i \in I \subseteq \mathbb{N}$, the possible partitions of X^α that satisfy definition 2.3. Then in the trivial case $C^\alpha_2 = X^\alpha - C^\alpha_1$. It is readily seen that the maximum value of $|I|$ occurs when all the $f^\alpha_\gamma(\Theta_\gamma)$ are disjoint, in which case $|I| = 2^{|\Gamma|} - 1$. The question of how to best choose C^α_i, and hence $\mathcal{P}^\alpha$ is discussed directly.

Selection criteria can be used to formulate a variety of selection procedures. The following is a simple example of such a procedure, phrased as an algorithm (i.e., a determinate set of instructions). The purpose of considering this first algorithm is to address how it can be improved, thereby giving a clearer idea of how to use selection criteria in model selection. The focus of the algorithm is its first step; in its present phrasing it is unhappily vague.

The input is an outcome variable O, a set of candidate models $\mathcal{M}$, and the set of selection criteria of those candidates. The stopping rule for this algorithm assumes that for each subset $\mathcal{M}' \subseteq \mathcal{M}$ the $M_\gamma \in \mathcal{M}'$ are distinct and that for some $\gamma^* \in \Gamma$ and some $\theta \in \Theta_{\gamma^*},\ p_O = p_{\gamma^*}(\theta)$. That is, the algorithm is formulated to require that selection criteria exist for each subset of candidates, and that one of candidates contains the "true model." After consideration of the algorithm, the consequences of dropping these two assumptions are compared. The result of dropping the assumption of a true model yields the principle of elimination, and this principle provides a mainstay for present approach to selection. The consequence of dropping the assumption of disjoint models includes as a special case any set of candidates with only a single model. This allows for discussion of model selection in situations where the principle of elimination does not lead to determinate solutions.

First Rudimentary Selection Algorithm

Step 1: Choose a selection criterion $\mathcal{P}^\alpha \in \mathcal{C}$ and a partition C_i

Step 2: Compute $g(O)$.

Step 3: If $g(O) \in C_i$ assign all M_γ such that $f_\gamma(\Theta_\gamma) \subseteq C_i$ to $\mathcal{M}'$ Otherwise assign all M_γ such that $f_\gamma(\Theta_\gamma) \subseteq X - C_i$ to $\mathcal{M}'$

Step 4: Set $\mathcal{M}$ to $\mathcal{M}'$

Step 5: If $|\mathcal{M}| > 1$ repeat from Step 1. Otherwise halt.

As noted, this procedure is not determinate unless its first step can be written unambiguously. Additionally, its can be seen that the efficiency of this algorithm is not very optimal. In the best case scenario, the algorithm first happens upon a $P^\alpha \in \mathcal{C}$ such that $f_{\gamma^*}(\Theta_{\gamma^*})$ is disjoint from all other $f_\gamma(\Theta_\gamma)$ and it selects $C_i = f_{\gamma^*}(\Theta_{\gamma^*})$. In this case the algorithm halts after the first iteration. In the worst case scenario, however, this selection algorithm eliminates one M_γ at each iteration, requiring a total of $|\Gamma| - 1$ repetitions before it halts. The worst case efficiency of the selection algorithm is therefore $O(|\Gamma|)$.

This procedure can be made determinate and its worst case efficiency improved by finding a good way to select the $\mathcal{P}^\alpha$ and C_i^α. To this purpose define the sets

$$A_i^\alpha = \{\, \gamma \mid f_\gamma^\alpha(\Theta_\gamma) \subseteq C_i^\alpha \,\}; \quad B_i^\alpha = \{\, \gamma \mid f_\gamma^\alpha(\Theta_\gamma) \subseteq X^\alpha - C_i^\alpha \,\}; \quad R_i^\alpha = \Gamma - A_i^\alpha \cup B_i^\alpha.$$

For any selection criterion $\mathcal{P}^\alpha \in \mathcal{C}$, these sets are disjoint and their cardinalities sum to $|\Gamma|$. For fixed α define an "optimized" partition $C^{\bar{\alpha}}$ as any C_i^α whose index satisfies

$$\bar{i} = \arg\min_i \, ||A_i^\alpha| - |B_i^\alpha|| + |R_i^\alpha|. \tag{2.1}$$

The minimum of this sum, call it an efficiency loss function, is $\lfloor 0 \rfloor$. This occurs when $|A_i^\alpha| = |B_i^\alpha|$ and the remainder term is equal to zero. Then, assuming such an i exists, an "optimized" partition is one for which $\lceil |\Gamma| /2 \rceil$ of the candidate's prediction ranges are in C_i^α and the rest are in $X^\alpha - C_i^\alpha$. Clearly, the use of such a partition assures us of selecting about 1/2 of the candidates. Whether or not the true minimum of the quantity in equation 2.1 is obtainable for a given α, the argument that minimizes it is not unique. Therefore, arbitrarily denote by $C^{\bar{\alpha}}$ that partition whose index has the smallest value. With regard to definition 2.3, henceforth it will be assumed that a selection criterion is always used with its optimized partition. That is, for each $\mathcal{P}^{\alpha} \in \mathcal{C}$ we now consider only the single partition, $C^{\bar{\alpha}}$.

Using optimized partitions it is then possible to defined an "optimized" selection criterion $\mathcal{P}^{\bar{\alpha}} \in \mathcal{C}$ as any $\mathcal{P}^{\alpha}$ whose superscript satisfies

$$\bar{\alpha} = \arg\min_{\alpha} ||A^{\alpha}| - |B^{\alpha}|| + |R^{\alpha}|. \tag{2.2}$$

Here A^{α}, B^{α}, and R^{α} are defined using $C^{\bar{\alpha}}$ in place of C_i^{α}. Equation 2.2 states that an optimized selection criterion is one whose optimized partition is most optimal. Again assuming that such an argument exists, the minimum of the loss function is $\lfloor 0 \rfloor$. If this minimum is realized, the selection criterion is chosen such that about 1/2 of candidates are selected. The argument that satisfies equation 2.2 may not be unique, in which case we arbitrarily select that prediction whose index has the smallest value. Any other choice would be fine also, and this can depend on the properties of the indices.

The first selection algorithm may now be re-written using optimized selection criteria.

Second Rudimentary Selection Algorithm

Step 1: Find $\mathcal{P}^{\bar{\alpha}}$.

Step 2: Compute $g(O)$.

Step 3: If $g(O) \in C^{\bar{\alpha}}$ assign all M_{γ} such that $f_{\gamma}(\Theta_{\gamma}) \subseteq C^{\bar{\alpha}}$ to $\mathcal{M}'$. Otherwise assign all M_{γ} such that $f_{\gamma}(\Theta_{\gamma}) \subseteq X - C^{\bar{\alpha}}$ to $\mathcal{M}'$.

Step 4: Set $\mathcal{M}$ to $\mathcal{M}'$

Step 5: If $|\mathcal{M}| > 1$ repeat from Step 1. Otherwise halt.

Although this second algorithm is still rudimentary in many ways, it is a clear improvement over the first. As per the above discussion of optimized selection criteria, this is a properly determined instruction. Further, if it is assumed that optimized criteria can always be found such that the loss function in equation 2.2 reaches its true minimum, then at each iteration about 1/2 of the remaining models are selected. In this case the total number of iterations is always $\lceil \log_2(|\Gamma|) \rceil$. This is a non-trivial improvement in worst-case efficiency from the first algorithm. A natural comparison here is between linear (sequential) and binary search. Although selection criteria do not order the input, they serve to partition the candidates into halves, and then we "follow the true model" in a manner similar to a binary search. As noted, this efficiency depends on how optimal the optimized criteria are, and, of course, on actually obtaining and cataloguing such criteria for the relevant models. Nonetheless, this improvement in efficiency demonstrates that prediction-based selection is *potentially* better than a linear search of the model space could *possibly* be. In comparison the IC-based approaches discussed in the first chapter, this means that we should want to do more than slowly order candidates with respect to various incommensurate metrics. Rather, we should want to quickly find accurate models.

There are many questions that can be asked about selection algorithms and selection criteria, and most of these have to do with pulling off this approach or some semblance thereof. This is the purpose of chapter three. The present chapter closes by summarizing this prediction-based approach to selection under a principle of model elimination and by providing an outlook on selection "in the trenches."

2.3.2 Selection Without True Models: The Principle of Elimination

The foregoing discussion of selection algorithms has assumed that one of the candidate models contained the "true model," meaning that $p_O \in p_\gamma(\Theta_\gamma)$ for some $\gamma \in \Gamma$. This allowed it to be shown that prediction-based selection finds this model in a manner analogous to binary search when suitable selection criteria are available. As with search algorithms, it is easy to modify the instructions to deal with cases where the target is not necessarily included in the input. In fact, the selection algorithm does not need to be modified at all, only its interpretation. In its present formulation, the result of the selection procedure is either to eliminate all models or stop when only one remains. For succinctness, let us consider these as the representative outcomes of a selection procedure and ignore cases where more than one model is retained; the latter requires only a straightforward generalization of the following remarks.

In the case that one model remains, all that is incumbent upon dropping the assumption of a true model is that this last be viewed as preferable in a Popperian sense rather than true in a naive sense. There is no uncertainty in the manner in which such a model is chosen – the instructions are no less determined without a true model, and the selection procedure would not have turned out otherwise unless the instructions were otherwise. Naturally, such revisions could be undertaken. For instance, the choice of selection criteria could take into account the ease with which they can be implemented, their substantive interpretations, or aesthetic considerations, and any of these may fairly trump cold efficiency. Yet, as long as selection is based on predictions satisfying definition 2.3, none of these revisions will change the nature of the procedure or its inevitable outcome. The last model, if there is one, will be preferable in the sense that it is accurate with respect the predictions used to select it. Thus, rather than "following a true model" we are exacting a standard of "truth" and in doing so we are led to the best candidate.

In summary, the selection algorithm described above does not require that a model's adequacy be judged relative to a true model or to indeed to any other model, but only to the means by which it is selected. In other words, when a candidate is preferable in the sense described here we can point to list of predictions that "came off." As described in this previous chapter, this is not similar to reporting the value of a relative fit index, in which case we can only point to the other candidates.

Let us turn now to the case in which none of the candidates are retained. This can hardly be reason for surprise, because it is precisely this result that has been made possible by dropping the assumption of a true model. Returning to the analogy of search algorithms, if the target is not in the input, this is just what the algorithm is supposed to tell us. There may be cases in which it is desirable to replace the assumption of a true model with a clause that one or more models must always be retained. To my mind, however, such a clause is incompatible with the interpretation of models as scientific theories or consequences thereof, because accuracy is not optional in such contexts. Although it is perhaps unavoidable, *this should not be interpreted as absolute or dogmatic accuracy.* As mentioned in the previous section and outlined in more detail in the following chapter, one can always state a level of accuracy that is sufficient to one's purposes. The types of problems suggested by such a task have the following flavor: For a given set of candidates and a user-defined set of predictions, find the "accuracy score" that minimizes some stated loss function. Clearly this is just another version of what has been done in this section, a version particular to a given context rather than designed to demonstrate general principle of selection.

By whatever means an acceptable level of accuracy is determined, it must be acknowledged that this is different from saying that the best candidate is the most accurate one, and thereby stating one's criterion level of accuracy in an *ad hoc* manner. From the perspective taken here, if a set of models fall short of their intended use, then so be it. This is a direct consequence of having a purpose in one's mind when using those models, a purpose defined independently of "modeling." As such, the general approach to model selection in this dissertation may be referred to under a principle of elimination. The *possibility* of eliminating models gives value to any that are retained.

2.3.3 Selection Without Selection Criteria

As noted above, selection criteria will not exist for all sets of candidates, let alone "optimized" selection criteria, and there is also no reason to imagine that these will be easy to come by. If a forced-choice between models cannot be arranged, there is little recourse than to employ the "garden variety" predictions of definition 2.2. For any prediction such that $f_\gamma(\Theta_\gamma) \subset X$ for at least one $\gamma \in \Gamma$, it is possible that $g(O) \notin f_\gamma(\Theta_\gamma)$. That is to say, so long as a prediction is not trivial, it may prove useful for selection. Note that unlike definition 2.3, this formulation of non-trivial predictions does not require multiple candidates. Therefore selection via non-trivial predictions can be applied to the case of a single candidate. However, it is important to note that whether or not a non-trivial prediction is useful for selection is entirely contingent on O. As a result, the efficiency of selection algorithms based on only non-trivial predictions rather than selection criteria can be very poor. For example, if selection is with respect to the full (i.e., uncountable) set of predictions, it may be the case the selection algorithm never ends. Thus the worst case efficiency is infinite, even if we only require to distinguish between two candidates. On the other hand if the number of predictions has been limited to a user-defined or otherwise finite set, the worst case efficiency is then equal to the number of predictions, but the algorithm may not select among any of the candidates. Clearly these situations are far from "optimized." Indeed, a natural question to ask is whether this kind of selection is better than selection with trivial predictions, in which case these same results are not contingent on the data.[4]

In short, selection without selection criteria brings us into the messy domain of deciding among indistinct models, where, of course, a single model is indistinct from itself. Although this is a much more realistic selection scenario, not much can be said about it from a strategic vantage point. Rather, the advances must be made on the ground, and this is what the following chapter of this thesis accomplishes. In particular, a measure of accuracy is provided for IRT models, and this can be used to judge how close a model-implied family of probability distributions is to an observed distribution. Unlike AIC and its derivatives, this measure has a minimum value of zero corresponding to the zero point of KL, and unlike KL, its maximum is unity. It is therefore meaningful to stipulate a "cut-off" value of accuracy that can be employed for the purpose of elimination. This approach therefore addresses the concerns of the previous chapter. However, it does not share the computational efficiency discussed in

[4] Such questions could be considered along following lines. Treating the prediction space as a probability space, the compliment of the probability of a model's prediction range could be interpreted as the chance of selecting against that model. Extending this idea to multiple candidates and multiple predictions, this could be interpreted in terms of the probability that a selection procedure ends.

connection with selection criteria, because many IRT models are nested. Furthermore, in application the proposed quantity must be computed for each model, rather than derived in advance for the models of interest and computed only once for the given outcome variable. In this sense its application is more similar to IC than to selection criteria. In short, although the method of selection proposed in the following chapter embodies the principle of elimination, there is still much that can be done to improve this approach.

In summary, it is the case that selection criteria will not usually be available, but the principle of elimination can nonetheless guide selection in these less tractable cases. Many difficult problems are to be faced in this domain, and the purpose of this chapter has been to offer a general strategy for finding such problems.

Chapter 3

Item Response Theory: Data-based Selection

The topic to be addressed in this chapter is data-based selection of IRT models. As described above, this connotes selection with respect to an outcome variable and this is to be distinguished from selection based solely on the properties of models. In particular, this chapter suggests a measure of Euclidean distance between two multinomial distributions. When one of these is the distribution of the data and the other is a model-implied distribution that minimizes this distance, it is interpreted as a measure of accuracy. Although this approach has much in common with least squares estimation, it is not proposed as a method of estimation, but as a method of selection. In practice, this amounts to changing the focus from θ to $p_M(\theta)$. This result is a method of evaluating whether $p_O \in p_M(\Theta)$, and moreover it allows the minimal distance between p_O and $p_M(\Theta)$ to be quantified in terms of the interval $[0, 1]$.

The chapter begins by specifying a general class of IRT models and considering their predictions (§3.1). Because the outcome variable is finite valued, lower case notation for random variables will be adopted in this chapter, with the realizations of these variables indicated by subscripts. This is consistent with the IRT literature on which this work is based. As a consequence of the finiteness of the outcome variable o, p_O and $p_M(\theta)$ are always representable as points on the $(2^J - 1)$-dimensional unit simplex in $\mathbb{R}^{2^J}$ (Holland 1990). This serves as a "fundamental" prediction space, since any other predictions are constituted by functions contingent on p_O and $p_M(\theta)$.

The question of a model's accuracy with respect to an outcome variable is considered (§3.2). Since the unit simplex is in $\mathbb{R}^{2^J}$, it is "natural" to apply Euclidean distance to this task. In §3.3 this distance is shown to have an obvious maximum, which allows for the high-dimensional norm to be represented in the interval $[0, 1]$. The relation of this measure to KL divergence and AIC is considered, and it is readily seen that minimizing KL is equivalent to maximizing accuracy when a model is correctly specified (§3.2). However, unlike AIC, the proposed measure is bounded and its unit is readily interpreted, thus allowing for "cut off" values to be employed for model elimination. It is also considered how to choose a value of $p_M(\theta) \in p_M(\Theta)$ on which to compute this distance when the model is not correctly specified. The ML estimator is suggested, although the results are not definitive. Thus the "model-implied component" of accuracy is reduced to evaluating $p_M(\hat{\theta})$.

Another issue relevant to the application of this procedure is the estimation of p_O with sparse data (§3.3). It is unrealistic to anticipate that all 2^J response patterns will have sufficient observations for estimation of these probabilities when J is large. Ways of making J smaller are therefore considered, the standard solution of "collapsing across cells" being sufficient to this task. That is, rather than considering all J

items, subtests of length $J' < J$ are considered instead. J' can be chosen such that estimation of the marginal probabilities of the response patterns is feasible. This amounts to computing the accuracy of various marginal distributions of p_{O} and $p_{M}(\theta)$ and it is obvious that each of these will have perfect accuracy when $p_{M}(\theta)$ does. When accuracy is not perfect, subtests can be used to identify problematic items as those which reduce model accuracy. Importantly, once $p_{M}(\theta)$ has been estimated, these considerations are computationally inexpensive. In short, although in many cases it may be infeasible to estimate the accuracy of a model with respect to an entire test, the present approach can be applied to subtests. For this reason it is also useful test construction purposes.

The chapter ends with consideration of a numerical example for two models of the general class considered. This example demonstrates the usefulness of the current approach, but also that there is still much work to be done.

3.1 The Prediction Manifold of IRT Models

Item response theory concerns a wide variety of research scenarios. A conventional example is a pen-and-paper achievement test, containing fixed items administered in fixed order under standardized testing conditions. Let the set $T = \{1, \ldots, J\}$ represent such a test, each of its integer components corresponding to one test item. For each of the $j = 1, \ldots, J$ items define the indicator variable

$$o_j = \begin{cases} 1, & \text{if item } j \text{ is answered correctly} \\ 0, & \text{if item } j \text{ is answered incorrectly.} \end{cases}$$

In this context the outcome variable $o = (o_1, \ldots, o_J)$ represents the possible *response patterns* to the test. There are $i = 1, \ldots, 2^J$ possible realizations $o_i = (o_{i1}, \ldots, o_{iJ})$ and to each of these a probability $p(o = o_i) = p_i$ is associated. Using the indicator

$$\delta_i(o) = \begin{cases} 1, & \text{if } o = o_i \\ 0, & \text{if } o \neq o_i \end{cases}$$

the discrete mass function of o is then

$$p_O(o) = \prod_i p_i^{\delta_i(o)} \ . \tag{3.1}$$

This outcome variable is the starting point for the considerations of this chapter. The research scenario that has motivated it is very modest. The domain of IRT has been extended to other research contexts in educational testing, for example to tests with polytomously scored variables (Bock, 1972; Samejima, 1972), and to tests with items that are not fixed but conditional on responses to previous items within the same test (i.e., adaptive testing, see Van der Linden & Glas, 2000). Many applications outside of the domain of educational testing have also been formulated (e.g., De Boeck & Wilson 2004). Moreover, this research scenario has given no concern to the population of respondents over which o is to be defined. For instance, it may be asked whether this population is properly viewed as homogeneous with respect to the items (Holland & Wainer, 1993). These are only a few of the important advances made since the early work of Birnbaum and also of Rasch. Although we shall have plenty to talk about within the basic IRT framework, it will be clear to the reader that general approach taken in this chapter can be extended to any model of a finite-valued outcome variable whenever the parameters of that model can be estimated.

A general form of IRT model functions for o is given in equation 3.2. As described in Table 3.1, the notation of this equation is not entirely consistent with that usually found in IRT literature.

$$p_{\scriptscriptstyle M}(x_i, \theta) = \int \prod_j P_j(y)^{x_{ij}} \big(1 - P_j(y)\big)^{1-x_{ij}} dF(y) \tag{3.2}$$

Note that equation 3.2 has the general form $p(x_i) = \int p(x_i \mid y) dF(y)$ where, for each y, $p(x_i \mid y)$ is a joint distribution of J independent Bernoulli variates. This provides a probability distribution of x under the restriction that $\sum_i p_{\scriptscriptstyle M}(x_i, \theta) = 1$. The probability distribution implied by equation 3.2 is obtained by replacing p_i with $p_{\scriptscriptstyle M}(x_i, \theta)$ in equation 3.1 and letting $o = x$. This distribution is denoted by $p_{\scriptscriptstyle M}(x)$ and the notation for its parameters is abbreviated by $p_{\scriptscriptstyle M}(x_i, \theta) = p_i(\theta)$ when the reference to M is clear.

Also note that θ is not explicitly represented in the right hand side of equation 3.2. Rather, by choosing the form of the P_j and F one obtains the different model function falling under this class of models. Usual examples for the P_j include logistic or normal c.d.f.s, with the same form being assumed over all j. Recent work has considered more general options (e.g., Miyazaki & Hoshino, 2009). Various means of dealing with the distribution of the latent variate are employed depending on the context, although when estimating IRT models F is typically normal (e.g., Baker & Kim, 2004). So, in the usual practice, θ is just a vector of location and dispersion parameters, perhaps also including additional quantities (e.g., guessing parameters). Letting a denote the number of parameters of each P_j and b the number of parameters

Table 3.1: Summary of IRT Notation

Notation	Title	Comments
$x_{ij} \in \{0,1\}$, $j = 1, \ldots, J$	Item response	A variable representing a response to the j^{th} item of a test of fixed order and fixed length. The subscript i denotes the response pattern to which x_{ij} belongs.
$x_i = (x_{i1}, \ldots x_{iJ})$, $i = 1, \ldots 2^J$	Response pattern	J-dimensional vector of item responses; the random variate considered in equation 3.2
$p(x_i, \theta)$	Model function	Model-implied probability of the i^{th} response pattern.
y	Latent variate	In the present treatment, $y \in \mathbb{R}$.
P_j	Item response function	A non-decreasing, $[0,1]$ function of y representing the probability of a correct response to the j^{th} item.
F	The c.d.f. of y	Typically assumed to be normal.

of F, the length of θ is $K = aJ + b$ and $\Theta \subseteq \mathbb{R}^K$.

Other approaches to equation 3.2 have also been taken. For example, in deriving results about the general class of models it is often desirable to treat the P_j, F, or both non-parametrically, with the non-parametric restrictions applying to all of the usual parametric forms. For instance Cressie and Holland (1983) considered a variation of equation 3.2 in terms of expectations over y and their analysis therefore only required the existence of these expectations rather than a particular form of F. Holland and Rosenbaum (1986) required only that P_j the are non-decreasing in y (see also Rosenbaum, 1983). In the following analysis, the form of the model-implied marginal probabilities only becomes essential when considering the numerical examples in §3.4.

As outlined above, the models of interest are of the form by $M = (o,\ p_{IRT},\ \Theta)$. A very convenient property of these models is that equation 3.1 can be represented as a 2^J-vector of probabilities (Holland, 1990). This can be accomplished by letting $\mathbf{o} = (o_1, \ldots, o_{2^J})$ represent the possible response patterns for a set of J fixed items and evaluating $p_O(o_i)$ at each of these. Then $\mathbf{p}_O = (p_O(o_1), \ldots, p_O(o_{2^J}))$ is vector representation of each of the probabilities given by equation 3.1. Note that $\mathbf{p}_O$ is also a data function since for any two finite-valued outcome variables o and o', $\mathbf{p}_O = \mathbf{p}_{O'}$ whenever $p_O = p_{O'}$. Clearly the converse implication is also true. For this reason any other data function computed on o is functionally dependent on $\mathbf{p}_O$ (because by definition it is functionally dependent on p_O; see §2.2). Therefore $\mathbf{p}_O$ is viewed as *the* data function of p_O. Indeed, the one-to-one correspondence between equation 3.1 and $\mathbf{p}_O$ has led other writers to identify the two (e.g., Amari, 1985; Chafai & Concordet, 2009). In the present treatment these two representations are distinguished because, in general, a probability distribution cannot be represented as a finite-dimensional vector of probabilities.

In a similar manner, $\mathbf{p}_M(\theta) = (p_M(x_1, \theta), \ldots, p_M(x_{2^J}, \theta))$ is interpreted as *the* prediction function of M. As with $\mathbf{p}_O$, its co-domain is readily seen to be

$$X_J = \{(t_1, \ldots, t_{2^J}) \mid t_j \geq 0,\ \Sigma t_j = 1\}.$$

X_J is just the $(2^J - 1)$-dimensional unit simplex in $\mathbb{R}^{2^J}$ (e.g., Munkres, 1984, §1.1), although in the present context it can also be thought of as the space of all 2^J-valued multinomial distributions, or what Holland (1990) referred to as a 'probability simplex.' The prediction $P = (\mathbf{p}_O, X_J, \mathbf{p}_M)$ is then an alternative representation of the model M.

This approach is convenient because $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$ is amenable to standard real analysis as a parametric representation of a manifold embedded $\mathbb{R}^{2^J}$. Arguably this is not obvious from looking at equation 3.1. Such an analysis requires the following regularity conditions (cf. Kreyszig, 1968, §16; Munkres, 1991, §23).

(R1) $\mathbf{p}_{\scriptscriptstyle M} \colon \Theta \to \mathbf{p}_{\scriptscriptstyle M}(\Theta)$ is continuous and injective.

(R2) $\mathbf{p}_{\scriptscriptstyle M}$ has continuous derivatives in θ up to the necessary order.

(R3) $\partial \mathbf{p}_{\scriptscriptstyle M} / \partial \theta_k,\ \theta_k = 1, \ldots, K$ are linearly independent functions in θ_k.

The interpretation of these conditions is as follows. The first is that of model identification – each $\theta \in \Theta$ can yield only one $\mathbf{p}_{\scriptscriptstyle M}(\theta) \in \mathbf{p}_{\scriptscriptstyle M}(\Theta)$. This implies the existence of the coordinate function ${\mathbf{p}_{\scriptscriptstyle M}}^{-1} \colon \mathbf{p}_{\scriptscriptstyle M}(\Theta) \to \Theta$, which allows the $\mathbf{p}_{\scriptscriptstyle M}(\theta)$ to be indexed by their parameter vector $\theta \in \Theta$. In usual treatments it is required that ${\mathbf{p}_{\scriptscriptstyle M}}^{-1}$ is also continuous in which case $\mathbf{p}_{\scriptscriptstyle M}(\theta)$ is a homeomorphism and $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$ can be defined as K-dimensional by means its correspondence with Θ. These considerations are important for defining $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$ as a manifold, but in the present case ${\mathbf{p}_{\scriptscriptstyle M}}^{-1}$ is not available, so this is of less relevance in practice. Rather, we will simply "work on" $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$, and avoid use of topological properties that explicitly require ${\mathbf{p}_{\scriptscriptstyle M}}^{-1}$ (e.g., change of coordinates).

(R2) ensures that relevant derivatives exist. The present analysis is restricted to second-order derivatives. (R3) requires that the partial derivatives span the tangent plane at any point on $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$. An important consideration here involves taking derivatives with respect to θ under the integral sign in equation 3.2. Nothing in following section requires doing this explicitly. Yet this is required in obtaining (marginal) ML estimates of θ on which the data analyses of the present work are ultimately based

(Baker & Kim, 2004, chap. 6). It is also the case that these derivatives need to be evaluated in order to ensure that the second and third regularity conditions hold. While the second requirement is easily checked for a given model, I know of no general analytic strategy by which to determine the linear independence of a set of K nonlinear, vector-valued equations. However, based on the computations involved in E-M estimation of θ (Baker & Kim, 2004, chap. 6), the matrix of partial derivatives could be evaluated at the ML estimates and its rank inspected numerically. This would ensure that the third condition holds in practice.

The following sections apply this representation of IRT models to their data-based selection. As described in the present section, the basic idea is to treat the set $\mathbf{p}_M(\Theta)$ as a geometrical object or manifold embedded in Euclidean space. This "likelihood manifold" represents all of the probability distributions implied by a model. The parameters of a model are a set of coordinates on the manifold, they serve to "name" each of its probability distributions. In this context, estimation procedures such as ML can be interpreted as algorithms for finding coordinates that satisfy a specified optimization criterion with respect to a given value of $\mathbf{p}_O$. The general goal of data-based selection can then be interpreted as determining whether one of the coordinates of the likelihood surface identifies the probability distribution of an outcome variable.

3.2 Accuracy: The Problem of Choosing θ

The basic problem in data-based selection is to decide if a given outcome variable is consistent with a given model. In the present context this problem can be formulated by letting $\mathbf{p}_O$ be any fixed point in X_J and considering its relationship to $\mathbf{p}_M(\Theta)$. In particular we require some means of deciding when $\mathbf{p}_O \in \mathbf{p}_M(\Theta)$. Moreover, as discussed at the end of the first chapter, if $\mathbf{p}_O \notin \mathbf{p}_M(\Theta)$ it is desirable to have a measure of the discrepancy between the two, so that this information can be used to judge whether or not a model is "close enough" to be useful. Because $X_J \subset \mathbb{R}^{2^J}$, an obvious measure of discrepancy between an observed distribution $\mathbf{p}_O$ and a point $\mathbf{p}_M(\theta) \in \mathbf{p}_M(\Theta)$ is the Euclidean distance $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$. In using the norm as a measure of model accuracy, consideration must be given to how to choose $\mathbf{p}_M(\theta) \in \mathbf{p}_M(\Theta)$, or equivalently (by R1), how to choose $\theta \in \Theta$. This consideration is the focus of the present section.

In the first place it may be noted that the distance between a given $\mathbf{p}_O$ and any *specific* model $\mathbf{p}_M(\theta)$ can be always computed. However, the problem set out in the previous paragraph does not concern the distance of $\mathbf{p}_O$ to a specific model but to a *general* model. For this reason a rationale is required in choosing a point or points in $\mathbf{p}_M(\Theta)$ for which to evaluate the norm. The natural choice is the orthogonal projection of $\mathbf{p}_O$ onto $\mathbf{p}_M(\Theta)$, which amounts to treating $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$ as a loss function in θ. This distance is minimized if $\mathbf{p}_O = \mathbf{p}_M(\theta)$, in which case it can be concluded that $\mathbf{p}_O \in \mathbf{p}_M(\Theta)$. If $\mathbf{p}_O \notin \mathbf{p}_M(\Theta)$ then there may exist multiple values of θ that minimize the Euclidean distance, but there is a unique solution if $\mathbf{p}_M(\Theta)$ is convex in the neighbourhood of $\mathbf{p}_O$.[1] In either case, the discrepancy between $\mathbf{p}_O$ and $\mathbf{p}_M(\Theta)$ can be quantified in terms of the norm of the orthogonal projection(s) of the former onto the latter.

As described above, the matter of choosing θ is clearly related to estimation of

[1] This is easy to see by assuming for the purpose of contradiction that $\mathbf{p}_M(\Theta)$ is convex and that there are two unique orthogonal projections of $\mathbf{p}_O$ onto $\mathbf{p}_M(\Theta)$. These three points form an isosceles triangle, and the points of the face opposite $\mathbf{p}_O$ are in $\mathbf{p}_M(\Theta)$ by the definition of a convexity (see e.g., Munkres, 1984, §1.1). Then the orthogonal projection of $\mathbf{p}_O$ onto its opposing face must be in $\mathbf{p}_M(\Theta)$.

model parameters. The foregoing rationale for choosing $\mathbf{p}_{M}(\theta) \in \mathbf{p}_{M}(\Theta)$ motivates a nonlinear least squares (NLS) approach to parameter estimation in IRT. This is rather inconvenient, since estimation in IRT traditionally does not employ NLS but ML. This tradition is firmly grounded in the classic asymptotic results for the ML estimators of correctly specified models. It has also been discussed how KL divergence motivates the ML estimator in more general contexts (§1.1.2). While the present approach has led to a different objective function, it too is concerned with the distance between two distributions, and so it might be expected, if only on intuitive grounds, that it should lead to results that are consistent with KL. Moreover, covariance-based generalized least squares (GLS) estimation for IRT has been worked out (Christofferson, 1975; Muthen, 1978) and this has been shown to be asymptotically equivalent to ML for correctly specified models (Browne, 1984). These results are again suggestive. In short, it would be nice to show that the ML estimator also minimizes the Euclidean distance between $\mathbf{p}_{O}$ and $\mathbf{p}_{M}(\Theta)$, thereby aligning the proposed approach to model selection with the established practice in IRT. This is the goal of the present section.

The following proposition shows that, under certain conditions, the maximum of the log-likelihood coincides with a stationary point of the norm function. Further considerations about whether the maximum likelihood is the minimum of the Euclidean distance are suggestive but inconclusive. Therefore the arguments for regarding the ML estimate as a good choice of $\theta \in \Theta$ for which to evaluate $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||$ must be regarded as incomplete.

Proposition 3.1 Assume that for some IRT model M the likelihood function of $p_{M}(x)$ has a unique maximum θ_o. Assume further that the likelihood is computed over a set of i.i.d. outcome variables whose realizations include all response patterns

$x_i, \ i = 1, \dots 2^J$, and that $p_M(x_i, \theta_o) > 0$ for all i. Then

$$\frac{\partial}{\partial\theta}||\mathbf{p}_{\mathrm{O}} - \mathbf{p}_{\mathrm{M}}(\theta)||\bigg|_{\theta=\theta_o} = \mathbf{0}. \tag{3.3}$$

Proof: The strategy is to compare the first-order derivatives of $||\mathbf{p}_O - \mathbf{p}_M(\theta)||^2$ and of the logarithm of $p_M(x) = \prod_i p_M(x_i, \theta)^{\delta_i(x)}$. This latter is the model-implied log-likelihood of a single response pattern when it is treated as a function of θ for fixed x_i. Generalization to the joint distribution of multiple i.i.d. x_i is straightforward and the details are omitted for succinctness.

Begin by writing $p_M(x_i, \theta) = p_i(\theta)$ and letting

$$u_i(\theta) = \frac{\partial}{\partial\theta} p_i(\theta) = \left[\frac{\partial}{\partial\theta^1} p_i(\theta) \cdots \frac{\partial}{\partial\theta^K} p_i(\theta)\right]$$

denote the gradient (row) vector of the i^{th} component of $\mathbf{p}_M(\theta)$. Then the derivatives of interest are

$$\frac{\partial}{\partial\theta}||\mathbf{p}_{\mathrm{O}} - \mathbf{p}_{\mathrm{M}}(\theta)||^2 = 2\sum_i (p_i - p_i(\theta))u_i(\theta) \tag{3.4}$$

and

$$\frac{\partial}{\partial\theta}\ln(p_{\mathrm{M}}(x)) = \frac{\partial}{\partial\theta}\sum_i \delta_i(x)\ln(p_i(\theta)) = \sum_i \frac{\delta_i(x)}{p_i(\theta)} u_i(\theta). \tag{3.5}$$

Inspection of equations 3.4 and 3.5 shows these to be linear combinations of the gradient vectors $u_i(\theta)$. The $u_i(\theta)$ exist by condition (R2) and the assumption that $p_i(\theta_o) > 0$ ensures that equation 3.5 is defined at θ_o for all i. By definition θ_o is an extremum of $p_{\scriptscriptstyle M}(x)$, and hence $u_i(\theta_o) = \mathbf{0}$ for each i. This can also be seen by setting equation 3.5 to the null vector and noting that $\delta_i(x)/p_i(\theta_o) > 0$ whenever $x = x_i$. Computing the log-likelihood over a set of all possible response patterns, equation 3.5 implies that $u_i(\theta_o) = \mathbf{0}$ for all i. Inspection of equation 3.4 shows this to coincide with a stationary point of $||\mathbf{p}_{\scriptscriptstyle O} - \mathbf{p}_{\scriptscriptstyle M}(\theta)||^2$, and the proposition follows.

Proposition 3.1 should be interpreted with due caution. Multiple values of θ can be stationary points of the log-likelihood function and by the argument given here, if equation 3.5 exists at these points they must also correspond to stationary points of $||\mathbf{p}_{\scriptscriptstyle O} - \mathbf{p}_{\scriptscriptstyle M}(\theta)||$. This is perhaps unsurprising since both quantities are functions of $p_{\scriptscriptstyle M}(\theta)$. What is stated by the proposition is that, when the log-likelihood is computed on a set of observations containing all possible response patterns, its maximum is a stationary point of $||\mathbf{p}_{\scriptscriptstyle O} - \mathbf{p}_{\scriptscriptstyle M}(\theta)||$. In the following paragraphs, some further considerations are brought to bear on the question of whether θ_o, the maximum argument of the likelihood function, is the minimizing argument of $||\mathbf{p}_{\scriptscriptstyle O} - \mathbf{p}_{\scriptscriptstyle M}(\theta)||$.

Some insight is provided by consideration the properties related to the second-order derivatives of the functions in proposition 3.1. Letting $H_i = H_i(\theta)$ denote the Hessian of $p_i(\theta)$ and writing $u_i = u_i(\theta)$ these are:

$$\frac{\partial^2}{\partial\theta'\partial\theta}||\mathbf{p}_{\scriptscriptstyle O} - \mathbf{p}_{\scriptscriptstyle M}(\theta)||^2 = 2\sum_i \big((p_i - p_i(\theta))H_i - u_i'u_i\big) \tag{3.6}$$

and

$$\frac{\partial^2}{\partial\theta'\partial\theta} ln(p_{\scriptscriptstyle M}(x)) = \sum_i \left(\frac{\delta_i(x)}{p_i(\theta)}H_i - \frac{\delta_i(x)}{p_i(\theta)^2}u_i'u_i\right). \tag{3.7}$$

Note that for both equation 3.6 and equation 3.7, the outer product of the gradient row vectors disappears at θ_o, and so we are primarily concerned with linear combinations of the H_i corresponding to those in equations 3.4 and 3.5. Let us first consider equation 3.6. Much can be said about this quantity if $\mathbf{p}_M(\Theta)$ is a convex set. For instance, this implies that the norm is also convex and hence that equation 3.6 is positive semi-definite for all $\mathbf{p}_M(\theta)$ (Boyd & Vandenberghe, 2004, §3.1.5). In such cases $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$ has a single stationary point and this is a global minimum (Borwein & Lewis, 2000, proposition 2.1.2). Under the conditions stated in proposition 3.1, this stationary point must be θ_o.[2] Thus it is necessary that the maximum likelihood is also the minimum distance is when $\mathbf{p}_M(\Theta)$ is convex. Now, the unit simplex X_J is itself convex (Munkres, 1984, §1.1) and it is also the case that Θ is convex for the usual IRT models discussed in §3.1. Also, since p_M is continuous in θ (by R1), $\mathbf{p}_M(\Theta)$ must be a connected subspace of X_J, and this is a necessary condition for convexity. However, to claim that $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$ is convex would seem to requires a model-by-model analysis; I know of no general strategy by which to accomplish this. It may also be noted that with or without the assumption of convexity, inspection of equation 3.6 reveals that it is equal to the null matrix when $\mathbf{p}_O = \mathbf{p}_M(\theta)$.

On the other hand, equation 3.7 must be locally negative semi-definite (i.e., the log-likelihood must be concave) in the neighborhood of θ_o. Since the product of a negative semi-definite matrix and a positive scalar is also negative semi-definite (Harville, 1997, lemma 14.2.3), $H_i(\theta_o)$ is negative semi-definite for all i. However, since the values of $(p_i - p_i(\theta_o))$ are not known when $\mathbf{p}_O \neq \mathbf{p}_M(\theta_o)$, this does not imply anything about equation 3.6. Thus considerations related to the second-order derivatives are not conclusive about whether $\mathbf{p}_M(\theta_o)$ is, in general, the minimizing argument of $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$.

[2] This also implies that the ML estimator described in proposition 3.1 must have a single extremum.

A different approach to this problem can be taken by noting that $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||$ reaches its global minimum if $\mathbf{p}_{O} = \mathbf{p}_{M}(\theta)$. It well known that the KL divergence from equation 1.1 has its minimum when this same condition is satisfied (e.g., Kullback, 1983). Then, assuming that the minimum of KL is obtained for some value of $\theta = \theta^{*}$ (i.e., assuming that $\mathbf{p}_{O} = \mathbf{p}_{M}(\theta^{*})$):[3]

$$\theta^{*} = \arg \min_{\theta} KL(p_{O}, p_{M}(\theta)) = \arg \min_{\theta} ||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||. \tag{3.8}$$

Since the ML estimate $\hat{\theta}$ asymptotically converges to θ^{*} (van der Vaart, 1998, §5.5; also §1.1.1 above), it follows that $\hat{\theta}$ asymptotically minimizes $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||$ when $\mathbf{p}_{O} \in \mathbf{p}_{M}(\Theta)$. Otherwise stated, when M is correctly specified, minimizing KL divergence is equivalent to minimizing the Euclidean distance. This point is addressed again in §3.3.

At this juncture several arguments have been brought to bear on the issue of whether maximum likelihood minimizes the Euclidean distance between a fixed but arbitrary $\mathbf{p}_{O}$ and a given model-implied prediction manifold $\mathbf{p}_{M}(\Theta)$. The following conclusions can be made: 1) Under the conditions of proposition 3.1, the maximum value of the log-likelihood function corresponds to a stationary point, possibly a minimum, of $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||$; 2) If the conditions of proposition 3.1 hold and additionally $\mathbf{p}_{M}(\Theta)$ is convex, then $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta_{o})||$ is the unique global minimum of the norm; 3) The ML estimator asymptotically minimizes $KL(p_{O}, p_{M}(\theta))$, and when $\mathbf{p}_{O} \in \mathbf{p}_{M}(\Theta)$, it is trivial to observe that the minimum of $KL(p_{O}, p_{M}(\theta))$ coincides with that of

[3] Note that θ_{o} is not the same quantity as θ^{*}. The former is the maximum of the likelihood function; the latter is the maximum of the expectation with respect to o of the log-likelihood; see §1.1.1.

$||\mathbf{p}_O - \mathbf{p}_M(\theta)||$. Although these arguments are suggestive, it is unclear whether the ML estimate is a suitable choice of $\theta \in \Theta$ for evaluating model accuracy in general. Further lines of argument are not pursued in this dissertation, and the topic is left as an open problem. Therefore, while the accuracy of any specific model obtained by ML estimation can be evaluated, whether this is equivalent to the accuracy of the general model that has been estimated is not yet settled.

The purpose of this section has been to align the proposal of evaluating model accuracy by means of Euclidean distance with the current estimation procedures in IRT. However, it has been noted that treating $||\mathbf{p}_O - \mathbf{p}_M(\theta)||$ as a loss function motivates a NLS solution to choosing θ, and so it may be asked why NLS should not be pursued instead of ML. While this could present an interesting line of research for robust estimation in IRT, it would presumably also require motivation in terms of being an improvement over ML. That such an improvement is available is not yet clear. At this point it seems that the only motivation for NLS is in terms of favoring the distance loss function proposed in this section over the KL loss function. Here it is important to recall that the distance function has been proposed as a method of evaluating model accuracy, not as a "new" approach to estimation in IRT. Nonetheless, this section shows that the problems of estimation and data-based selection are foundationally intertwined.

3.3 Accuracy: A Summary Measure and Potential Applications

The foregoing section has proposed the Euclidean distance $||\mathbf{p}_{O} - \mathbf{p}_{M}(\theta)||$ as a means of evaluating model accuracy and has considered the problem of choosing a parameter $\theta \in \Theta$ for which to compute it. Yet there remains the matter of whether this quantity is particularly worthwhile for model selection purposes. In applied settings it is common for response pattern residuals to be provided with the standard output of IRT software, and it is perhaps not clear that their sum of squares is an especially useful quantity. More generally, the magnitude of a high-dimensional Euclidean distance is not straightforward to interpret when this is different from zero. The present section addresses such concerns. Firstly it is shown that two points in X_J have a distance of most $\sqrt{2}$. As such it is easy, although unnecessary, to "standardize" this distance. It is discussed how the use of this quantity overcomes many difficulties associated with the interpretation of information criteria (cf. §1.2). In the remainder of this section it is shown how the Euclidean distance can be used as a measure of the accuracy of model not only with respect to a full test, but also with respect to subtests and, in particular, single items. The problem of estimating this quantity in sample-based applications with sparse data is also discussed. The arguments in favor of the usefulness of this relatively simple approach to model evaluation are summarized at the end of this section.

In the following proposition it is observed that the distance between any two points in X_J has a finite upper bound. As a preliminary step it is also shown that the norm of any vector in X_J is at most equal to one.

Proposition 3.2 Let $\mathbf{a}, \mathbf{b} \in X_J$. Then

1. *MAX* $||\mathbf{a}|| = 1$

2. *MAX* $||\mathbf{a} - \mathbf{b}|| = \sqrt{2}$

Proof (Part 1): Let $a_i,\ i = 1, \ldots, 2^J$ denote the i^{th} component of $\mathbf{a}$. Assume $a_i = 1$ for some value of i. Then since $\sum a_i = 1,\ a_k = 0$ for all $i \neq k$ and $||\mathbf{a}|| = 1$. To see that this is the maximum value of $||\mathbf{a}||$ assume that $a_i < 1$ for all i. Then $a_i^2 \leq a_i$, with equality holding *iff* $a_i = 0$. Therefore $\sum a_i^2 < 1$ and $||\mathbf{a}|| < \sqrt{1} < 1$.

Proof (Part 2): Let $b_i,\ i = 1, \ldots, 2^J$ denote the i^{th} component of $\mathbf{b}$. Since $a_i \geq 0$ and $b_i \geq 0$ for all i, $(a_i - b_i)^2 \leq a_i^2 + b_i^2$. Therefore

$$\begin{aligned} ||\mathbf{a} - \mathbf{b}||^2 &\leq \sum_i (a_i^2 + b_i^2) \\ &= ||\mathbf{a}||^2 + ||\mathbf{b}||^2. \end{aligned}$$

It is easy to check that equality holds for $a_i = 1$ and $b_k = 1$ with $i \neq k$, in which case the equation is just an instance of Pythagoras' theorem.

Proposition 3.2 suggests using the quantity $z(\theta) = ||\mathbf{p}_O - \mathbf{p}_M(\theta)||/\sqrt{2}$ as a summary measure of model accuracy. This quantity has several nice properties that are made explicit in the following:

1. $z(\theta) \in [0, 1]$;

2. $z(\theta) = 1$ implies $\mathbf{p}_O \perp \mathbf{p}_M(\theta)$;

3. $z(\theta) = 0$ *iff* $p_O = p_M(\theta)$;

4. $z(\theta) = 0$ *iff* $KL(p_O, p_M(\theta)) = 0$.

These properties serve to address the shortcomings associated with the use of information criteria discussed at the end of chapter one. The first three state that z has a unit range with interpretable end points. Thus, unlike the case for relative fit indices, it is *meaningful* to posit criterion values for model acceptability. How these values are to be chosen in any particular case is not necessarily a simple task and could be thought of as an area of further research. Yet one obvious and non-arbitrary choice is to require that $z(\theta) = 0$. The most significant and perhaps also the most obvious property of z is that it allows for a direct and unambiguous answer to the question of when a model is accurate with respect to a given outcome. In the terminology of the first chapter of this dissertation, it answers the question of whether or not a model achieves its intended purpose.

As noted by the fourth point above, the proposed quantity agrees with a well-established measure of model divergence when these are equal to zero (see §3.2). Thus, if we are happy to minimize KL, then in the more limited context of IRT modeling, we should also be happy to to minimize $z(\theta)$. An important difference between relative fit indices that minimize KL and $z(\theta)$ is that the latter indicates when the objective function in question reaches its true minimum. Moreover, if $z(\theta) \neq 0$ it is possible to say, quite literally, how far away the model is from p_O. Thus we can also use $z(\theta)$ for purposes such as determining whether one model is farther away than another, whether more than one model achieves some criterion value, and so on. In short, $z(\theta)$ can be used to determine the accuracy of any given model and therefore to rank models accordingly. Considering these points, it would seem that the relatively simple approach developed here has much to commend it.

There are also shortcomings. One issue is that the maximum of $z(\theta)$ occurs only in circumstances that are, to say the least, implausible in practice. Not only must the vector representations of two points in X_J be orthogonal, they each must give zero probability to all but one of the o_i (i.,e., the must be Cartesian coordinate vectors in $\mathbb{R}^{2^J}$). So $z(\theta)$ could benefit from a more restrictive upper bound. For instance, only non-orthogonal vectors or vectors with length less than one might be considered. A stricter range may be more readily obtainable by considering a different quantity, for instance the angle between $\mathbf{p}_O$ and $\mathbf{p}_M(\theta)$. For any $\mathbf{a}, \mathbf{b} \in X_J$, it is straightforward to show that the angle between their vector representations in $\mathbb{R}^{2^J}$ is in the interval $[0, \pi/2]$ and that it equals zero *iff* $\mathbf{a} = \mathbf{b}$. In general, however, this angle does not have a simple relation to the distance between $\mathbf{a}$ and $\mathbf{b}$, because the length of the vectors in X_J is variable. Finding more restrictive upper bounds for z or related quantities is a topic of further study.

Another objection that can be raised here is that $z(\theta)$ doesn't take into account model complexity, and therefore that it unwittingly favors overparameterized models. Although this objection has already been acknowledged by the distinction between data-based and model-based selection, it is one that can be fruitfully addressed again here. In the first place it is worthwhile to note that $\mathbf{p}_O$ is not a random sample of observations from a population but the parameters of that population. So long as we have a reliable estimate of $\mathbf{p}_O$, we should want to select a model-implied set of probabilities $\mathbf{p}_M(\theta)$ that is as close to that estimate as possible. If the specific model is completely accurate, then further samples taken from the population p_O will be predicted just as well by using $p_M(\theta)$, because these are the same distribution. On the other hand, if it turns out that an estimated value of $\mathbf{p}_O$ is not reliable, this has nothing to do with $z(\theta)$ or indeed with M. Thus it seems that the overparameterization objection conflates two related issues, both of which are quite important. Firstly, if $\mathbf{p}_{\scriptscriptstyle M}(\Theta)$ is "very large" then we can always find a $\mathbf{p}_{\scriptscriptstyle M}(\theta) \in \mathbf{p}_{\scriptscriptstyle M}(\Theta)$ that is close to any $\mathbf{a} \in X_J$, when in fact we are only interested in $\mathbf{p}_{\scriptscriptstyle M}(\theta)$ that are close to $\mathbf{p}_O$. This type of issue is a topic in model-based selection. The second issue has to do with small sample sizes, which is a problem that leads to the topics addressed in the remainder of this section.

In practice, evaluating $z(\theta)$ requires estimating $\mathbf{p}_O$ from a set of observations and computing $\mathbf{p}_{\scriptscriptstyle M}(\theta)$ for given θ. As discussed in the previous section, the choice of θ determines whether $z(\theta)$ is properly regarded as a measure of the accuracy of the general model to which $p_{\scriptscriptstyle M}(\theta)$ belongs, and this raises issues relating to the estimation of model parameters. On the other hand, requiring that $\mathbf{p}_O$ be estimated independently of its model-implied probabilities places large demands on the number of observations per response pattern. Even in relatively modest applications, 2^J is a big number. The least probable response pattern(s) under p_O cannot have a probability larger than 2^{-J}, and this only occurs when $p_i = p$ for all i. If any one response pattern is more likely, then the rest of them must be less likely on average and in particular at least one must have probability less than 2^{-J}. These considerations indicate that reliably estimating $z(\theta)$ requires very large sample sizes, say on the order of thirty thousand for a test of 10 items. Therefore, a practical concern immediately arises about the use of $z(\theta)$ in smaller applications, and this concern is here discussed in terms of sparse data.

The problem of sparse data can be addressed in a variety of ways. The approach taken here is to reduce J by considering subtests consisting of $2 \leq J' \leq J$ items. The general idea is to choose J' such that, for a fixed sample size, estimation of the response pattern probabilities for subtests of length J' is based on sufficient observations. An important feature of the present approach is that subtests are not obtained by means of "dropping items" and then separately estimating various lower dimensional models from the same sample data. In this case it seems more appropriate to speak of statistically dependent estimates of multiple models. Rather, the idea here is to derive consequences for subtests from a model of the full test.

Importantly, subtests need not be used only for dealing with sparse data. In many cases the accuracy of subtests may be of intrinsic interest, and in particular the J unique subtests consisting of $J - 1$ items are discussed in some detail below. It is described how these may be compared amongst each other or to the overall test in order to evaluate the changes in accuracy that are due to the omission of a single item. Items that lead to substantially better accuracy can be beneficially left out. These considerations provide an example of how accuracy can be used not only to select models but also for test construction. Because subtests pertain to the problem of sparse data as well as to issues of test construction, both of these will be discussed in connection with the results of the following subsection.

3.3.1 Subtests and Subtest Accuracy

In order to address subtest applications of $z(\theta)$, some further notation must be introduced. In §3.1 the set $T = \{1, \dots, J\}$ was used to represent a test consisting of J items, each of which is denoted by an element of T. A subtest is a "user-defined" subset $T' \subseteq T$. For instance, the subtest consisting of all but the first item of T is $T' = \{j \mid j \in T \text{ and } j \neq 1\}$. Let $J' = |T'|$ denote the number of items in T'.

In the first place we require a means of indexing the response patterns for T'. To this purpose define the set

$$R = \{i \mid o_{ij} = 1 \text{ for all } j \in T - T' \text{ and } i = 1, \ldots, 2^J\}.$$

R contains the indices of all response patterns with a "perfect score" on the items not included in T'. For example, R contains two indices if the subtest contains one item, and those indices correspond to the response patterns whose components are all equal to one, with the exception of that component which represents the response to the item on the subtest. In general there are $2^{J'}$ possible responses to the J' items in T', and to each of these corresponds exactly one response pattern o_i that has $o_{ij} = 1$ for all $j \notin T'$. As noted, the purpose of constructing R is to obtain a set of indices $r \in R$ that correspond to the possible response patterns of the items in T'; any other configuration of responses to the items in $T - T'$ could have been used in place of the "perfect score" and this would serve the same purpose.

For each $r \in R$, consider the set

$$I_r = \{i \mid o_{ij} = o_{rj} \text{ for all } j \in T' \text{ and } i = 1, \ldots, 2^J\}.$$

I_r contains the indices of the response patterns o_i that have the same responses to each item in T' as o_r. Note that in particular $r \in I_r$. Intuitively, each index in I_r corresponds to an identical response pattern on T'. There are $2^{J-J'}$ indices that have the same responses to each item on the subtest as o_r, since this is how may different response patterns there are for the items in $T - T'$. Hence the I_r serve to partition the 2^J values of i into $2^{J'}$ sets that each contain $2^{J-J'}$ indices corresponding to identical subtest response patterns. It is easy to check that $\cap I_r = \emptyset$ and $\cup I_r = \{1, \ldots, 2^J\}$

for any given T'.

Lastly consider the sets $O_r = \{o_i \mid i \in I_r\}$. These play the role of response patterns for a subtest of items $T' \subseteq T$. The variable $o' \in \{O_r \mid r \in R\}$ will be used to denote the $2^{J'}$ possible outcomes of a subtest T'. With reference to equation 3.1, the probabilities of each realization $o' = o_r'$ are given by

$$p_r = \sum_{i \in I_r} p_O(o) = \sum_{i \in I_r} \prod_k p_k^{\delta_k(o)} \,. \tag{3.9}$$

The notation $\mathbf{p}_O' = (p_1, \ldots, p_{2^{J'}})$ will be used to denote the $2^{J'}$ vector of subtest probabilities, as this is clearer to read than $\mathbf{p}_{O'}$. As with o, the model-implied probabilities of o' are obtained from equation 3.2 by letting $p_i = p_i(\theta)$ and $x = o$. Their vector representation is denoted $\mathbf{p}_M'(\theta)$.

Equation 3.9 shows that the subtest probabilities are implied by those of the full test, and so they do not need to be re-estimated for each subtest. While it is clear that estimating the p_r either directly or through the p_i is equivalent for a fixed sample, it is perhaps less obvious that θ does not need to be re-estimated for subtests. Specifying separate models for lower dimensional tests is just to consider *different* models of the same data. Equation 3.9 lets us consider subtests in terms of the implications of a single model.

Having established a notation for dealing with subtests, the task now is to consider how this facilitates the application of $z(\theta)$. Let the accuracy of a subtest be denoted by $z'(\theta)$. The computation of $z'(\theta)$ is straightforward, the only difference being to consider the $2^{J'}$ probabilities p_r given by equation 3.9 in place of the 2^J probabilities p_i given by equation 3.1. While the computation of subtest accuracy is obvious, its interpretation requires rather more work. It is natural to begin with the question of how the accuracy of a model with respect to a subtest relates to that of the full test. The answer to this question is the content of the following proposition, which is the main result of this section. Before dealing with this result it is worthwhile to develop some initial insight into the question through consideration of the components of the vector $\mathbf{p}_O - \mathbf{p}_M(\theta)$.

If $z(\theta) = 0$, then $p_i = p_i(\theta)$ for all i and it is easy to see that $z'(\theta) = 0$ also. That is, if a test is perfectly accurate then so are each of its subtests. However, if $z(\theta) > 0$ then $p_i \neq p_i(\theta)$ for some of the $i = 1, \ldots, 2^J$. Without loss of generality, assume that $p_i > p_i(\theta)$ for a particular value of i. Then $\sum_{k \neq i} p_i < \sum_{k \neq i} p_i(\theta)$ and so $p_k < p_k(\theta)$ for at least one $k \neq i$. In general, all of the positive and negative components of $\mathbf{p}_O - \mathbf{p}_M(\theta)$ must "balance out" since $\sum(p_i - p_i(\theta)) = 0$. Roughly, this means that $z'(\theta)$ can be less than $z(\theta)$ when subtest items involve summing over response pattern deviations with opposite signs. On the other hand, if subtest items sum over deviations with the same sign, accuracy can be worsened. These issues are addressed more formally by the following proposition.

Proposition 3.3 For a given model M, let $z(\theta)$ be computed for a test T with J items and let $z'(\theta)$ be computed on a subtest $T' \subseteq T$ with $J' \leq J$ items. Then

$$z_d = \big(z'(\theta)\big)^2 - \big(z(\theta)\big)^2 = \sum_{r \in R} \sum_{\substack{i,k \in I_r \\ i<k}} (p_i - p_i(\theta))(p_k - p_k(\theta)) \tag{3.10}$$

Proof : The derivation is basic.

$$\begin{aligned}
2\big(z'(\theta)\big)^2 &= \sum_{r\in R}(p_r - p_r(\theta))^2\\
&= \sum_{r\in R}\left(\sum_{i\in I_r} p_i - \sum_{i\in I_r} p_i(\theta)\right)^2\\
&= \sum_{r\in R}\left(\left(\sum_{i\in I_r} p_i\right)^2 + \left(\sum_{i\in I_r} p_i(\theta)\right)^2 - 2\left(\sum_{i\in I_r} p_i \sum_{i\in I_r} p_i(\theta)\right)\right)\\
&= \sum_{r\in R}\left(\sum_{i\in I_r} p_i^2 + \sum_{i\in I_r} p_i(\theta)^2 - 2\sum_{i\in I_r} p_i p_i(\theta)\right.\\
&\qquad\left. +2\sum_{\substack{i,k\in I_r\\ i<k}} p_i p_k + 2\sum_{\substack{i,k\in I_r\\ i<k}} p_i(\theta) p_k(\theta) - 4\sum_{\substack{i,k\in I_r\\ i<k}} p_i p_k(\theta)\right)\\
&= \sum_{r\in R}\left(\sum_{i\in I_r}(p_i - p_i(\theta))^2 + 2\sum_{\substack{i,k\in I_r\\ i<k}}(p_i - p_i(\theta))(p_k - p_k(\theta))\right)\\
&= 2\big(z(\theta)\big)^2 + 2\sum_{r\in R}\sum_{\substack{i,k\in I_r\\ i<k}}(p_i - p_i(\theta))(p_k - p_k(\theta)). \ \Box
\end{aligned}$$

Inspection of equation 3.10 shows that the difference between $z(\theta)$ and $z'(\theta)$ depends on the arrangements among the $(p_i - p_i(\theta))$ given by the I_r. But it is not yet obvious how z_d changes with T'. The next few paragraphs explain how the number of items per subtest relates to the value of z_d for a single subtest of length J'. These remarks are then related to the problem of sparse data and issues of test construction.

From equation 3.10 it can be seen that the sets of indices R and I_r govern the total number of terms appearing in z_d. Here it is necessary to recall that for each of the $2^{J'}$ elements $r \in R$, $|I_r| = 2^{J-J'}$. To each I_r there corresponds a $2^{J-J'}$-vector,

D_r, with components $(p_i - p_i(\theta))$, $i \in I_r$. The components below the diagonal of the outer product matrix $D_r D_r'$ are the "crossproducts" $(p_i - p_i(\theta))(p_k - p_k(\theta))$ appearing in equation 3.10. There are $(2^{\,2(J-J')} - 2^{J-J'})/2$ of these crossproducts corresponding to each of the I_r. Summing over the $2^{J'}$ values of r, this yields a total of

$$2^{J'}(2^{\,2(J-J')} - 2^{J-J'})/2 = 2^{J-1}(2^{J-J'} - 1) \tag{3.11}$$

terms appearing in equation 3.10. For instance, if $J' = J$ there are zero crossproducts. If $J' = J - 1$ there are 2^{J-1} crossproducts, one per subtest response pattern. If $J' = J - 2$ there are $2^{J-1} \cdot 3$ crossproducts and six per subtest response pattern. And if $J' = 1$ there are $(2^{\,2J-1} - 2^J)/2$ total terms in z_d, with half of these corresponding to the two response patterns of T'. In general, equation 3.11 shows that the number of terms in z_d decreases in J', and that any subtest of length J' corresponds to a proper subset of the $(2^{\,2J} - 2^J)/2$ unique crossproducts given by the J items on the full test. Also note that for each different subtest of length J', the I_r are different partitions of the indices $i = 1, \ldots, 2^J$. Thus different crossproduct terms appear in different subtests of length J', and so each different subtest of length J' will in general yield a different value of z_d.

Further considerations about the value of z_d are less straightforward. If $z(\theta) \neq 0$, some of the $(2^{2J} - 2^J)/2$ crossproducts of the full test must be negative and some must be positive. Moreover, it can be seen that the sum of these crossproducts must be negative. In particular

$$\sum_i^{2^J} \sum_k^{2^J} (p_i - p_i(\theta))(p_k - p_k(\theta)) = \sum_i^{2^J} \left((p_i - p_i(\theta)) \sum_k^{2^J} (p_k - p_k(\theta)) \right) = 0$$

implies that

$$\sum_{i \neq k}^{2^J} \sum^{2^J} (p_i - p_i(\theta))(p_k - p_k(\theta)) = -\sum_{i}^{2^J} (p_i - p_i(\theta))^2 = -2\big(z(\theta)\big)^2. \tag{3.12}$$

Equation 3.12 shows that the sum of the $(2^{2J} - 2^J)/2$ crossproducts is necessarily negative when $z(\theta) \neq 0$. As noted above, however, when $J' \geq 1$ only a subset of the terms in equation 3.12 appear in z_d. The strongest statement that can be made here is that as J' decreases, it is expected though not certain that z_d will be negative. Because equation 3.12 tells us nothing about the values of the individual crossproducts, it is not possible to say anything definite about a sum of a subset of these values on the basis of that equation. For instance it is possible that only 2^J of them are negative, or that many of them are null. One point of interest here is that when $J' = 1$, its is a well known phenomena that $z'(\theta) = 0$ for the usual IRT models (e.g., Holland, 1990). Equation 3.10 implies that $z_d = -\big(z(\theta)\big)^2$ in this case also, although it is not clear why this occurs. For the case of $J' > 1$, it may only be concluded that z_d can be either positive or negative, depending on the subset of crossproducts that are appear in its calculation.

One possibility here is to try to "build up" the complete set of crossproducts in equation 3.12 from multiple subtests of length J'. In this way, the term z_d would be known and the accuracy of the full test would obtainable from that of the subtests. However, this approach quickly becomes quite complicated. In particular, summing z_d over all possible subtests of length J' will not always include all possible crossproducts and it will often include the same product multiple times. For example in the case where $J' = J - 1$ it is not possible that the sum of z_d over all subtests includes all of the terms in equation 3.12. More generally, using equation 3.11 it can be seen that

$$(2^{2J} - 2^J)/2 \leq \binom{J}{J'} 2^{J-1}(2^{J-J'} - 1)$$

implies

$$2^J \leq \binom{J}{J'}(2^{J-J'} - 1) + 1. \tag{3.13}$$

Equation 3.13 gives a general condition for when the sum of z_d over all possible subtests of length J' will include at least as many crossproduct terms as equation 3.12 – whenever the inequality holds it is the case that at least as many terms appear in the sum of the z_d as in equation 3.12. If the condition does not hold, then it is not possible that summing z_d over all possible subtests of length J' will yield a "full set" of crossproducts, since there are not enough of them. In particular, equation 3.13 requires that $J' > J - 1$. However, for a fixed value of $J' > J - 1$, it is not generally the case that the different subtests yield unique crossproducts. This can be seen by counter example (e.g., $J = 3$ and $J' = 2$).

In general, there does not appear to be a straightforward solution for the value of z_d. This leaves open the question of the relationship between subtest accuracy and full test accuracy when the latter is not known. To address this issue, the following proposition provides a weak lower bound for $z(\theta)$ in terms of $z'(\theta)$.

Proposition 3.4 Let $z(\theta)$ and $z'(\theta)$ be computed as in proposition 3.3. Then

$$z(\theta) \geq \frac{z'(\theta)}{\sqrt{2^{J-J'}}}. \tag{3.14}$$

Proof: Similar to proposition 3.3, we have

$$2z\,'(\theta)^2 = \sum_{r\in R}\left(\sum_{i\in I_r}(p_i - p_i(\theta))\right)^2$$

Consider the sum over I_r. Let $A_i = p_i - p_i(\theta)$ and $B_i = B = 1$. Then applying the Cauchy-Schwarz inequality to $\left(\sum A_iB_i\right)^2 = \left(\sum(p_i - p_i(\theta))\right)^2$ yields

$$\begin{aligned} 2z\,'(\theta)^2 &\leq \sum_{r\in R}\left(\sum_{i\in I_r}(p_i - p_i(\theta))^2(2^{J-J'})\right) \\ &= (2^{J-J'})\,2z(\theta)^2. \end{aligned} \tag{3.15}$$

The result follows by rearranging terms.

Proposition 3.4 states that the accuracy of the full test can be no better than that of a subtest divided by the square root of the number of response patterns per subtest item. The latter term can be thought of as an average squared deviation. For example, in the case of a subtest with $J-1$ items, $z(\theta)$ can be no less than $z\,'(\theta)/\sqrt{2}$. Clearly the lower bound is quite weak, since rearranging equation 3.14 to provide an upper bound for $z\,'(\theta)$ shows that this quickly exceed the theoretical maximum of 2. However, it may be of limited use in application. It may also be noted here that upper bounds on accuracy, either of subtests or the full test, would not be particularly meaningful, since, as discussed under proposition 3.2, the upper range of $z(\theta)$ exceeds that which is plausible in practice.

At this point several results concerning subtest accuracy have been presented, and their interpretations with respect to applications of $z'(\theta)$ are now discussed. In general, if $z_d < 0$ then the observed and model-implied response pattern probabilities of T' are more proximate than those of T. Thinking of the subtest probabilities as a lower dimensional representation of those of the full test, then this representation serves to "conceal" discrepancies between $\mathbf{p}_O$ and $\mathbf{p}_M(\theta)$. On the other hand, if $z_d > 0$ the corresponding result is to simultaneously decrease the dimension of, and increase the distance between, the response pattern probabilities. Perhaps most interestingly, if $z_d = 0$ the lower dimensional test contains all of the prediction error of the full test. While this is already a complicated phenomenon, its present interpretation in terms of subtests must concern items rather than response patterns themselves. The relationship between items and response patterns is not particularly straightforward at the best of times, and this is the case here also. The central complication is that each item appears in each response pattern, so that dimension reduction via item omission simultaneously affects all response patterns. However, the following three points can be observed.

Firstly, it is possible for omission of test items to improve model accuracy (i.e., it is possible that $z_d < 0$). A well known instance of this phenomenon occurs when considering the marginal probabilities of individual items, in which case the residuals are always zero. A less trivial case occurs when a single item is omitted from a test. If this results in $z'(\theta) < z(\theta)$, then "collapsing across" that variable has the effect of reducing the overall error in response pattern predictions. In terms of test construction this is clearly a desirable outcome – it means that leaving out such items yields a test for which we have a more accurate model. As such, when both $z'(\theta)$ and $z(\theta)$ can be computed, their comparison (e.g., their ratio) can be used to screen particular items or subtests. In general, however, omission of any single item need not improve accuracy. As noted above, there are many more possible crossproducts than there are ways to select subtests of length $J' = J-1$. Thus it would not be impossible for omission of each item to increase $z(\theta)$, and hence it may be the case that there is no single item that can be omitted to improve unsatisfactory model accuracy.

This leads to the second point: It is possible for omission of items to worsen model accuracy. When using subtests to compute model accuracy for sparse data, it is important to know how much better accuracy of the model for the full test can be. This has been addressed above by proposition 3.4. Note that the lower bound is a decreasing function of J'. On the other hand, the number of p_r to be estimated is increasing in J', meaning that subtests with fewer items require less response patterns to be estimated. Thus the use of subtests to address the problem of sparse data must balance the need for larger sample sizes per response pattern with the need for meaningful approximations to $z(\theta)$. In particular, the largest subtests possible should always be employed – the maximum number of response patterns that can be reliably estimated should be included when computing $z'(\theta)$.

Let J^* denote the largest number of items whose response pattern proportions can be reasonably estimated for fixed sample size. Then there also exists the problem of deciding which of the $\binom{J}{J^*}$ possible subtests to use to compute $z'(\theta)$. Each of these subtests will, in general, imply a different value of $z'(\theta)$, and by proposition 3.3 these values are related by the equation $z'(\theta) = \sqrt{\left(z(\theta)\right)^2 + z_d}$. Since z_d is unknown when $J' > 1$ there is little here that can be motivated by the above results. Nonetheless, the following suggestion is made.

It would not be unreasonable to consider the median of $z'(\theta)$ over all possible subtests of length J^*. This is because $z'(\theta)$ only varies through z_d, and some of values of z_d will imply that $z'(\theta)$ is closer to $z(\theta)$ than other values of z_d. The median of the $z'(\theta)$ serves to get rid of this variability by picking out a value of $z'(\theta)$ that will be closer to $z(\theta)$ than at least half of those computed. In the best case scenario, some values of z_d will be positive and some negative, and so $z(\theta)$ will fall somewhere in between the minimum and maximum values of $z'(\theta)$. However, there is no assurance that this will be the case, and all of the values of z_d could be positive or negative. In this scenario the mean of the $z'(\theta)$ could be very mislead, yet the median of $z'(\theta)$ is still closer to $z(\theta)$ than half the of the computed values. Therefore it is suggested that the problem of sparse data can be dealt with by taking the median of $z'(\theta)$ over all possible subtests of length J^*. Clearly this situation is less than ideal; a better solution would be to have a large enough sample to reliably estimate $z(\theta)$.

The third point to be made is that response patterns that are perfectly predicted do not affect a model's subtest accuracy. That is, $z'(\theta) = z(\theta)$ when $p_i - p_i(\theta) = 0$ for each $i \in I_r$ and each $r \in R$. It is not exactly clear how to interpret this at the item level, although in terms of test construction there would be no reason to omit such items. In terms of lower dimensional representations of the relation between $\mathbf{p}_O$ and $\mathbf{p}_M(\theta)$ in X_J, subtests that have the same accuracy as the full test could be thought of as error free projections of the full dimensional space. At this point is not clear how this can be employed for data-based selection, although it is an interesting avenue of further research.

Summary

The overall purpose of this section has been to show that a quantity useful for evaluating the accuracy of IRT models is readily obtainable. This quantity is a standardizable

Euclidean distance between an observed multinomial distribution and a model-implied multinomial distribution. While this is a rather simple approach to model evaluation, it nonetheless has several advantages over the use of relative fit indices. In particular it allows for an individual model to be evaluated without reference to other models. This is because the distance function has a clearly interpretable optimal value, namely zero. Other values can, in principle, be chosen as being "optimal enough" for a given application. Regardless of how a criterion value is chosen, any individual model can be eliminated when it fails to meet this criterion. Thus a central problem associated with relative fit, the unwitting selection of a poor model, is avoided. This serves to place model selection on firmer grounds than relative fit, while at the same time allowing for selection among candidates on the basis of how well they minimize a specified loss function. In the case that at least one candidate must be retained, the present approach nonetheless has the advantage of indicating whether that model should be regarded as satisfactory.

While it is reasonably straightforward to see how $z(\theta)$ can be used for model selection with multiple models, it also leads to some less obvious considerations regarding model evaluation by means of subtests. Subtests have been addressed in order to deal with the practical problem of sparse data. To this purpose, proposition 3.4 provides a lower bound on the accuracy of a full model when $z(\theta)$ must be estimated from a subset of test items. As noted, it would be desirable to tighten this bound. Consideration of subtests has also led to consideration of issues of test construction. Although the relation of individual items to the full set of response patterns is inherently complicated, some reasonably clear results have been presented. In particular, if one or more items are omitted, it is possible that the accuracy of the model can be improved.

If this is the case, grounds are provided for constructing tests for which the model is more accurate. In the following section the results presented in the foregoing are illustrated by means of a numerical examples.

3.4 A Numerical Example

In this section data from the Self Monitoring Scale (SMS; Snyder, 1974) are used to illustrate how model accuracy can be used to address model selection, the problem of sparse data, and for test construction purposes. The sample consists of $N = 903$ observations to 25 dichotomously (true / false) scored items concerning the respondents' social behavior. The first six items of the scale were employed in the following analysis. IRT models were estimated using MULTILOG 7. The MML estimation routine was employed with response pattern frequencies as input. Post processing was conducted using SPSS 15, as only basic data manipulation procedures are required for subtest partitioning of response patterns and to estimate z from the MULTILOG output.

The questions of interest are as follows.

1. For the first six items of the SMS, is either the 1-parameter logistic model (1PL) or the 2-parameter logistic model (2PL) accurate with respect to the observed response patterns? The criterion value to be employed for judging accuracy is $z(\theta) = 0$, since this is value is readily interpretable and there is no *a priori* reason to consider any other value. Naturally it is not expected that $\hat{z}(\theta) = 0$, and some discussion of inferential error is given.

2. Is 2PL more accurate than 1PL? Because the models are nested, any observation that is well predicted by 1PL must also be well predicted by 2PL. Using subscripts 1 and 2 for 1PL and 2PL respectively, this means that, if $\mathbf{p}_O$ is close to $\mathbf{p}_{\mathit{1}}(\Theta_1)$, it must also be close to $\mathbf{p}_{\mathit{2}}(\Theta_2)$. As such, finding that 1PL is accurate will bring us face to face with the problem of deciding between equally accurate models. This is a clear case of when accuracy is not a sufficient condition for model preference. Also note that comparing the two models does not imply that either will be ultimately accepted.

3. Can model accuracy be improved through omission of any single item? As a related question, it is also considered whether model accuracy can be worsened through the omission of an item. This makes it clear that it is the particular items removed that lead to improvements in model accuracy, rather than simply the removal of items.

These three questions are addressed by Table 3.2. For each model, the estimates of $z\sqrt{2}$ for the full test and a number of subtests are reported. The lower bound from equation 3.14 is computed on the reported median value for each set of subtests.

Unsurprisingly, neither of the models are estimated to be perfectly accurate on the full test or any of the subtests. The problem of inference can be addressed in a variety of ways, for example by computing confidence intervals on the $\hat{z}$, either analytically or by bootstrap methods. The present approach to inferring the accuracy of a model with respect to the full test is as follows: Compute a single binomial confidence interval on the p_i such that i satisfies $MAX_i\,\{(p_i - p_i(\theta))^2\}$ and consider whether the corresponding value $p_i(\theta)$ falls in the interval. In order to consider the logic of this approach, assume for the moment that the intervals on the p_i have equal

Table 3.2: Accuracy of 1PL and 2PL for the First Six Items of the SMS

Items	1PL		2PL	
Omitted	$\sqrt{2}\hat{z}(\theta)$		$\sqrt{2}\hat{z}(\theta)$	
None	0.06210	$-2\ln(L) = 175.6$	0.06027	$-2\ln(L) = 161.7$
1	0.05822		0.05875	
2	0.07562		0.07426	
3	0.07927	Med.= 0.06330	0.07594	Med.= 0.06385
4	0.06835	$\sqrt{2}z^{*} = 0.04475$	0.06896	$\sqrt{2}z^{*} = 0.04515$
5	0.05825		0.05258	
6	0.05137		0.04700	
1,2	0.07007		0.07207	
1,3	0.07357		0.07303	
1,4	0.06083		0.06151	
1,5	0.04535		0.04423	
1,6	0.04765		0.04783	
2,3	0.09504		0.09211	
2,4	0.08604		0.08770	
2,5	0.07403	Med. = 0.06919	0.06743	Med. = 0.06169
2,6	0.06584	$\sqrt{2}z^{*} = 0.03459$	0.06145	$\sqrt{2}z^{*} = 0.03084$
3,4	0.08422		0.08467	
3,5	0.06981		0.06169	
3,6	0.06508		0.05856	
4,5	0.06919		0.06662	
4,6	0.04496		0.04345	
5,6	0.03686		0.01532	

Note: z^{*}denotes the lower bound from equation 3.14 computed on the median value. L denotes the likelihood statistic.

width for all $i = 1, \dots 2^J$. Then if the computed interval includes $p_i(\theta)$, this must also be the case for all smaller deviations. On the other hand, if the interval does not include $p_i(\theta)$, then we can infer that $\mathbf{p}_O \neq \mathbf{p}_M(\theta)$, and hence that $p_O \notin p_M(\theta)$. This treatment of the inferential problem can be interpreted as recasting the selection problem in terms of the *sup norm* (Munkres, 1991, chap. 1) instead of the Euclidean norm. This approach is chosen for its simplicity.

One complication here is that the width of a confidence interval on a binomial proportion varies not only through the sample size, but also through the magnitude of the proportion. In the present application, this can be addressed by using $MAX_i\,\{\hat{\sigma_{p_i}}\}$ to compute the interval described above. This implies that interval may be wider than that corresponding to the nominal coverage, and hence that we may be too liberal with the evaluation of a model. In practice, however, it can be reasonably expected that all of the p_i will be very small and hence that the $\hat{\sigma}_{p_i}$ will be dominated by the sample size. In such cases there will be negligible variation among the σ_{p_i}

Using this approach, the 95% confidence intervals on the accuracy of 1PL and 2PL with respect to the first six items of the SMS were computed using the Agresti-Coull interval (Brown, Cai & DasGutpa, 2001). In this case $MAX_i\,\{(p_i - p_i(\theta))^2\}$ was satisifed by the i for both models, and so the intervals are identical for both models. These are reported below along with the model implied values.

1PL 95% CI: $[0.05974, 0.09947]$; $p_i(\theta) = 0.04341$

2PL 95% CI: $[0.05974, 0.09947]$; $p_i(\theta) = 0.04529$

From the above intervals it may be concluded that neither 1PL nor 2PL is accurate with respect to the first 6 items of the SMS. Inference on subtest accuracy may be computed in a similar manner.

Because it is also possible to consider sampling error via $\hat{\theta}$, it is worthwhile to explain why this approach was not taken here. In short, this is quite antithetical to the current approach. Recall that each $\mathbf{p}_M(\Theta)$ is an object in X_J and each of its $\mathbf{p}_M(\theta)$ are points on that object given by the coordinate vector θ. There is nothing about this scenario that varies over samples, and rather, given analytically tractable equations, the entire situation could be surveyed in advance of any data collection. On the other hand, $\mathbf{p}_O$ is a quantity that represents the data and is unknown in applications. On the basis of $\mathbf{p}_O$ a point on $\mathbf{p}_M(\Theta)$ can be computed via $\theta \in \Theta$, but it is only through a sample from $\mathbf{p}_O$ that this point is random. Therefore, $\mathbf{p}_O$ rather than θ is unknown, and so estimation error is properly applied to former quantity. However, if an estimated value of p_i is not closer to $\mathbf{p}_M(\hat{\theta})$ than some other point on $\mathbf{p}_M(\Theta)$, this could be interpreted with regard to the choice of θ in computing z (cf. §3.2).

In order to better judge the reported values of $\hat{z}$, the likelihoods of the models are also reported. These are quite poor, as readers familiar with the SMS will have anticipated. By way of comparison, 2PL is known to fit Thissen's LSAT example quite well; in this case $\sqrt{2}\hat{z} < 0.014$ and the 95% confidence interval leads to the conclusion that the model is accurate at the criterion value of $z(\theta) = 0$.

It should also be noted that sparse data is a concern in the interpretation of the full model in Table 3.2. Of the 64 response patterns, 21 (34.4%) had a sample size of $n_i \leq 5$ and only 8 (11.1%) had $n_i \geq 30$. The subtest values of $\hat{z}$ can be more reliably interpreted. For the 5-item subtests, the worst case was 4 (12.5%) response patterns with $n_i \leq 5$, and about 30%-40% of response patterns had for $n_i \geq 30$ on each test. For the 4-item subtests, one reponse pattern had a sample size less than 5, most had

$n_i \geq 30$, and many had sample sizes in the hundreds. When interpreting the subtest values it should be kept in mind that they deviate from the fulltest accuracy by the term z_d in eqation 3.10, and that this quantity is unknown. Thus, although serving to address sparse data, subtest accuracy cannot be taken as directly reflecting the models' accuracy with respect to the full test. As discussed at the end of the previous section, the median of these values can be taken as a better approximation of fulltest accuracy than that of any of the individual subtests. The lower bounds on $\hat{z}$ are also of some use here, although visibly less so for the 4-item subtests. As noted, these lower bounds have been computed on the median values. From Table 3.2, the general consensus is that 2PL is more accurate than 1PL for this sample of SMS data.

Although subtest accuracy does not directly reflect that of the full test, it does reflect that of the subtest itself. In particular, inspection of Table 3.2 for the 5-item subtests shows that omission of the sixth item leads to a lower estimate of z. This can be interpreted to mean that the specific models estimated for this data provide a better description of the 5-item test obtained by omitting the sixth item than for the full test. That is to say, we currently have a better model of *this* subtest than of the full test. It is natural to wonder whether re-estimating the 1PL and 2PL models for this subtest also leads to better accuracy. This is considered in Table 3.3. For both models it is shown that re-estimating the first five items of the SMS leads to a lower value of $\sqrt{2}\hat{z}$ relative to the 6-item test. Indeed, the estimated values of z correspond quite closely to those from Table 3.2, which are again presented in Table 3.3 for the purpose of comparison. Table 3.3 also shows that omitting the third item leads to worse accuracy than for the full six items, and again the values of $\sqrt{2}\hat{z}$ are noticeably similar to those found in Table 3.2. This indicates that it is not just fewer items that

Table 3.3: Re-estimation of Two 5-item Subtests

Item		1PL		2PL	
Omitted		$\sqrt{2}\hat{z}(\theta)$	$-2\ln(L)$	$\sqrt{2}\hat{z}(\theta)$	$-2\ln(L)$
6	Original Model	0.05137	175.6	0.04700	161.7
	Re-estimated	0.05152	68.1	0.04652	53.4
3	Original Model	0.07927	175.6	0.07594	161.7
	Re-estimated	0.07875	134.6	0.07552	126.8

leads to better accuracy, but omission of particular items. The values of $-2\ln(L)$ are also reported to aid in interpreting the values of $\hat{z}$. In general, Table 3.3 shows that the answer to the third question above is clearly in the affirmative – omission of items can lead to improved accuracy, and such items can be detected by means of subtest accuracy.

This example has many limitations, although it serves the intended purpose of illustrating some aspects of the interpretation of model accuracy in an applied context. These illustrations could be pursued much further, for example by considering alternative methods of taking into account the inferential error; by considering simulation studies in order to get an idea of the "empirical" distribution of $\hat{z}$ as a function the model considered, their parameterizations, and sample size; by considering alternative approaches to establishing criterion values of z. With regard to the latter consideration, a particularly interesting line of further study would be to establish the criterion values relative to the independence model (Holland, 1990). In this manner, we may consider whether a given IRT is more accurate than a model that does not involve a latent variate, and in this sense make a conclusion about whether the IRT model is better than "no model." There is still much work to be done to elaborate the application of accuracy to model selection, and also to test construction, and the present section has merely served to illustrate that such efforts can be fruitful. The summary chapter discusses these issues from a more general perspective.

Chapter 4

Summary and Outlook

This research has considered the general problem of model selection. This has been motivated by the observation that current methods, both traditional tests of goodness of fit and newer information theoretic approaches, have left room improvement. While this has long been recognized in the case of testing goodness of fit, the general argument made in the first chapter of this dissertation is that information criteria also have their shortcomings. In particular, relative fit indices cannot address the question of whether any single model is optimal in the sense defined by the objective functions of those indices. The main objective of the current work has been to develop a means of addressing this shortcoming, while retaining the more realistic perspective that models can be, and generally are, misspecified.

The second chapter was quite ambitious in its formulation of the problem of data-based model evaluation of parametric stochastic models. Many further objections to current methods can be read from this chapter, although it has not been my intention to dwell on these any longer than required to motivate a general consideration of the problem of evaluation. Therefore the significance of the ideas presented in this second chapter will be best appreciated by readers who have considered the problem in the course of their own research. In particular, by defining a model as the explicit juxtaposition of model-implied and data-implied quantities, the distinction between the two becomes unavoidable. This separation is fundamental to the idea that a model can be wrongly applied. This notion can be easily overlooked when models are separated from the process of scientific theorizing and treated as an ends in themselves. The idea that "modeling" is a scientific endeavor requires that the application of models be subject to appropriate epistemological standards. The goal of the second chapter was to instantiate such standards. For readers with limited interest in the topic, it may seem that this was a rather long excursion into a domain without much ostensible "pay off." Here it can only be argued that a clear statement of the problem is worth a thousand misguided solutions, and that the contribution of the second chapter can be properly understood from this perspective.

The solution presented in the third chapter was incomplete with respect to establishing a value of θ on which to evaluate the proposed measure of model accuracy. Nonetheless, the proposal of quantifying accuracy by means of the Euclidean distance between an "data-implied" multinomial distribution and the closest model-implied distribution is in principle a sound answer to the question of when a general model can be said to imply a given outcome. This considerations of this chapter are sufficient, however, for the data-based selection of *specific* IRT models. For any given parametrization of an IRT model, the proposed quantity can be computed on sample data, (asymptotic) inferential error can taken into account, and its use in the identification of "problematic" items presents a novel approach to test construction. While much work remains to be done in the application of this quantity, an initial indication of the value of this work has been given.

There are many further topics in model evaluation that have not been addressed in this manuscript. In particular, it is my intention to extend the approach developed in the foregoing pages to issues of model-based selection. However, as argued in the present work, data-based selection must be primary in the application of stochastic models. That the focus of this dissertation has been restricted to this problem is a fair indication of its severity.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F. (Eds.) *Second International Symposium on Information Theory,* pp. 267-281. Budapest: Akademiai Kiado.

Amari, S. (1985). *Differential Geometrical Methods in Statistics.* New York: Springer.

Baker, F. B. & Kim, S. (2004). *Item Response Theory: Parameter Estimation Techniques(2nd ed.).* New York: Marcel Dekker Inc.

Bamber, D. & van Santen, J. P. H. (19- 85). How many. parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29,* 443-473.

Bamber, D. & van Santen, J. P. H. (2000). How to assess a model's testatbility and identifiability. *Journal of Mathematical Psychology, 44,* 20-40.

Bartholowew, D. J. & Hand, D. J. (1999). *Latent Variable Models and Factor Analysis.* London: Arnold.

Billingsley, P. (1986). *Probability and Measure.* New York: Wiley.

Birnbaum A. (1968). Some latent trait models and their use in estimating an examinee's ability/ In F. M. Lord & M. R¿ Novick (Eds.) *Statistical Theories of Mental Test Scores,* pp. 397-479. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51/

Bollen, K. A. (1989). *Structural Equations With Latent Variables.* New York: John Wiley and Sons.

Borwein, J. M. & Lewis, A. S. (2000). *Convex Analysis and Nonlinear Optimization: Theory and Examples.* New York: Springer.

Boyd, S. & Vandenberge (2004). *Convex Optimization.* New York: Cambridge University Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion: The general theory and analytic extensions. *Psychometrika, 52,* 345-370.

Brown, L. D., Cai, T. T. & DasGutpa, A. (2001). Interval estimation for binomial proportions. *Statistical Science, 16,* 101-133.

Browne, M W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *The British Journl of Mathematical Psychology, 37,* 62-83.

Burnham, K, P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach* (2nd ed.). New York: Springer-Verlag.

Chafi, D. & Concordet D. (2009). Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association, 104,* 1071-1079.

Christofferson, R. (1997). *Log-linear Models and Logistic Regression.* New York: Springer Verlag.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychomterika, 40,* 5-32.

Claeskens G. & Hjort N. L. (2008). *Model Selection and Model Averaging.* Cambridge, UK: Cambridge University Press.

Cohn, D. L. (1980). *Measure Theory.* Boston, MA: Birkhauser

Cressie, N. & Holland, P. W. (1983). Characterizing the Manifest Probabilities of Latent Trait Models. *Psychomterika, 48,* 129-141.

De Boeck, P. & Wilson, M. (Eds.) (2005). *Explanatory Items Response Models: A Generalized Linear and Nonlinear Approach.* New York: Springer.

de Leeuw, J. (1992). Introduction by J. de Leeuw. In Kotz, S. and Johson, N.

(Eds). *Breakthroughs in Statisitics (Vol 1),* pp. 599-610. New York: Springer.

Dyrmes, P. J., Howrey, E. P., Ymans, S. H., Kmenta, J., Leamer, E. E. Quandt, R. E., Ramsey, J. B. Shapiro, H. T., & Zarnowitx, V. (1972). Criteria for evaluation of econometric models. *Annals of Economic and Social Measurement, 1,* 291 - 323.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology, 44,* 205-231.

Goodman, N. (1965). *Fact, Fiction, and Forecast.* Indianapolis, IN: Bobbs-Merrill.

Grasa, A. A. (1989). *Economic Model Selection: A New Approach.* Boston, MA: Kluwer.

Grüwald, P. D. (2007). The Minimum Description Length Principle. Canbridge, MA: The MIT Press.

Halpin, P. F. & Maraun, M. D. (under review, MS#0902). Selection Between Linear Factor and Latent Profile Structures. *Multivariate Behavioral Research.*

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577-601.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

Holland, P. W. & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523-1543.

Holland, P. W. & Wainer H. (Eds.) (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Jeffreys, H. (1939). *Theory of Probability.* Oxford: Clarendon Press.

Kass, R. E. & Raftery, A. E¿ (1995). Bayes Factors. *Journal of the American Statistical Association, 90,* 773-795.

Kass, R. E. Tierney, L., & Kadane, J. B. (1990). The validity of posterior asymptotic expansions based on Laplace's method. In (Eds) S. Geisser, J. S. Hodges, S.J. Press and A. Zellner, *Bayesian and likelihood methods in statistics and econometrics,* pp. 473-488. New York: North-Holland.

Konishi, S. (1999). Statistical model evaluation and information criteria. In Ghosh, S. (Ed). *Multivariate Analysis, Design of Experiments, and Survey Sampling.* pp. 369-401. New York: Marcel Dekker Inc.

Kreyszig, E. (1975)*Introduction to Differential Geometry and Riemannian Geometry.* Toronto, Canada: University of Toronto Press.

Kullback, S. (1983). Kullback Information. In S. Kotz & N. Johnson (Eds.) *Encyclopedia of Statistical Sciences (Vol. 4),* pp. 421-425. New York: Wiley

Leamer, E. E (1978). *Specification Searches.* New York: John Wiley and Sons.

Linhart, H. & Zucchini, W. (1986). *Model Selection.* New York: John Wiley and Sons.

Maraun, M. D., Slaney, K. & Goddyn, L. (2003). An analysis of Meehl's MAXCOV-HITMAX procedure for the case of dichotomous items. *Multivariate Behavioral Research, 38,* 81-112.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis.* The Hague: Mouton & Co.

Miyazaki, K. & Hoshino, T. (2009). A Bayesian semiparametric item response model withe Dirichlet process priors. *Psychometrika, 74,* 375-394.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551-560.

Myung, J. (2000). The importance of complexity in model selection. *The Journal of Mathematical Psychology, 44,* 190-204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA, 97,* 11170-11175.

Munkres, J., R. (1991). *Analysis on Manifolds.* Boston, MA: Addison-Wesley Publishing Co.

Munkres, J., R. (1984).*Elements of Algebraic Topology.* Boston, MA: Addison-Wesley Publishing Co.

Platt, J. R. (1964). Strong inference. *Science, 146,* 354-353.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-163.

Rao, C. R. & Wu, Y. (2001). On model selection. In Greenhouse, J. (Ed.) *Model Selection,* pp. 1-57. Institute of Mathematical Statistics Monographs Series.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling.* New York: Springer.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of the item response theory. *Psychometrika, 49,* 425-435.

Samejima, F. (1972). A General Model for Free Response Data. *Psychometric Monographs Supplement*, No. 17.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461.464.

Searle, S. R. (1971). *Linear Models.* New York. Jorhn Wiley and Sons.

Sin, C & White, H. (1996). Information criteria for selecting possibly misspecified models. *Journal of Econometrics, 71,* 207-225.

Snyder, M. (1974). Self-monitoring of expressive behaviour. *Journal of Personality and Social Psychology, 30,* 526-537.

Suppes, P. (1960). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese, 24,* 287-300.

Suppes, P., & Zinnes, J. L. (1963) In R. D. Luce, R. R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology (Vol. 1)*, pp. 1-76. New York: Wiley.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association,* *81,* 82-86.

Van der Linden, W. J., & Glas, C.A.W. (Eds.). (2000). *Computerized Adaptive Testing: Theory and Practice.* Boston, MA: Kluwer.

Van der Vaart, A. W. (1988). *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57,* 307-333.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika, 92,* 937-950.