Estimation with Consistency Analysis

February 2007
Timothy C. Haas
Lubar School of Business Administration
University of Wisconsin-Milwaukee
haas@uwm.edu

Most of this report appears in the Appendix of Haas (2001).

1 Consistency Analysis Algorithm

Let $g_S(\beta)$ be a goodness-of-fit statistic that measures the agreement of a model’s joint distribution identified by the values of $\beta$ and the (possibly) incomplete sample, $S$. Let $g_H(\beta)$ be the agreement between the model’s hypothesis distribution (identified by the values of $\beta_H$) and the model’s distribution identified by the values of $\beta$. Let $g_{S\text{max}}$ be the unconstrained maximum value of $g_S(\beta)$ over all $\beta$ and let $\beta_S$ be the parameter values that correspond to this maximum. Let $g_{H\text{max}}$ be the unconstrained maximum value of $g_H(\beta)$ over all $\beta$. Up to sampling variability in the estimation of $g_H(\beta)$, this value is $g_H(\beta_H)$. The consistency analysis estimator maximizes $g_{CA}(\beta) \equiv (1-c_H)g_S(\beta)/|g_{S\text{max}}| + c_Hg_H(\beta)/|g_{H\text{max}}|$ where, as discussed above, $c_H \in (0, 1)$ is the analyst’s priority of having the consistent distribution agree with the hypothesis distribution as opposed to agreeing with the empirical distribution. Let $\hat{\beta} \equiv \arg\max_{\beta}(g_{CA}(\beta))$ be the consistency analysis estimate of $\beta$. If the model consists of only qualitative random variables and the sample is incomplete, a value for $\beta$ can be found such that the consistent model exactly reproduces the empirical distribution of the observed variables. This is the case studied in Haas (1991).

The idea of consistency analysis is to find parameter estimates such that, under the priorities dictated by $c_H$, the fitted model is as close as possible to both the model defined by the substantive theory and to the model suggested by the sample. Consistency analysis is best used as a sequential learning procedure: $\hat{\beta}$ from one sample is used as $\beta_H$ in the consistency analysis conducted with the next sample.
1.1 Probability Density-Probability Function

Let $\mathbf{U}$ be the model’s $r$-dimensional vector of random variables. Let $\mathbf{U}^{(d)}$ be the vector of the model’s discrete random variables, and $\mathbf{U}^{(ac)}$ the vector of the model’s absolutely continuous random variables. The joint cumulative distribution function of $\mathbf{U}$ can be decomposed as $F_{\mathbf{U}}(\mathbf{u}) = F_{\mathbf{U}^{(d)}}(\mathbf{u}) + F_{\mathbf{U}^{(ac)}}(\mathbf{u})$ where $F_{\mathbf{U}^{(d)}}(\mathbf{u})$ is the pure discrete component – completely determined by the probability mass function (PMF), $P(\mathbf{U} = \mathbf{u})$, and $F_{\mathbf{U}^{(ac)}}(\mathbf{u})$, the pure absolutely continuous component – completely determined by the probability density function (PDF) $f_{\mathbf{U}^{(ac)}}(\mathbf{u}) = \frac{\partial F_{\mathbf{U}^{(ac)}}(\mathbf{u})}{\partial \mathbf{U}}$ (Koopmans 1969). Koopmans gives a hybrid of the PMF and PDF called the probability density-probability function (PDPF) that is convenient for computing joint probabilities of $\mathbf{U}$. The PDPF is defined as:

$$p_{f_{\mathbf{U}}}(\mathbf{u}) \equiv \frac{\partial}{\partial \mathbf{U}^{(ac)}} P(\mathbf{U}^{(d)} = \mathbf{u}^{(d)}, \mathbf{U}^{(ac)} \leq \mathbf{u}^{(ac)}) \quad (1)$$

1.2 CA Components

The function $g_S(\beta)$ measures how close a model with parameter values $\beta$ agrees with the sample. When some of the model’s variables have not been observed, the incomplete sample can only be compared to the PDPF of these remaining or observed random variablees, called here the marginal PDPF. Finding parameter estimates that result in the unconstrained maximum of $g_S(\beta)$ is called Minimum Distance (MD) parameter estimation and was first studied analytically by Wolfowitz (1957). MD estimation is now known to be robust to model misspecification (see Bickel et al. (1993) and Lindsay (1994)).

Often, the model’s conditioning variables (independent variables) effectively define independent models given each combination of values on these variables. Say that there are $d$ unique combinations of these variables. If the sample is a time series, the model’s PDPF should capture the temporal dependence structure. Let $\mathbf{U}^{(i)}_t$ be the vector of observed random variables at time $t$ under conditioning variable combination $i$. Say that these variables are observed at $m$ times, $t_1, \ldots, t_m$. Then, a PDPF is needed for the random vector $\mathbf{U}^{(i)} = (\mathbf{U}^{(i)}_1, \ldots, \mathbf{U}^{(i)}_m)'$, $i = 1, \ldots, d$. Because $\mathbf{U}^{(i)}$ and $\mathbf{U}^{(j)}$ are independent for $i \neq j$, the PDPF of $\mathbf{U} = (\mathbf{U}^{(1)}_1, \ldots, \mathbf{U}^{(d)}_m)'$ is the product of the constituent joint PDPF’s: $p_{f_{\mathbf{U}}}(\beta)(\mathbf{u}) = \prod_{i=1}^{d} p_{f_{\mathbf{U}^{(i)}}}(\beta)(\mathbf{u}^{(i)})$.

Using $n_{MC}$ simulated realizations from the model of $\mathbf{U}^{(i)}$, the PDPF of $\mathbf{U}$ can be
estimated with the local gaussian kernel estimator (Thompson and Tapia 1990, p. 180). Say that \( S \) consists of observations \( u_{\text{obs}} \) on the variables \( U \). For this particular sample, the value of the marginal likelihood function can be estimated with this simulation-based density estimator, i.e., \( \hat{L}(\beta|u_{\text{obs}}) = pf_U(\beta)(u_{\text{obs}}) \). Further, if \( c_H = 0 \) and \( g_S(\beta) = \hat{L}(\beta|u_{\text{obs}}) \), then \( \hat{\beta} \) is exactly the MML estimate of \( \beta \).

Cressie and Read (1984) give a class of objective functions for MD parameter estimation: for \( \lambda \in (-2, 1) \) and \( \delta(u_i) = (d(u_i) - pf\beta(u_i))/pf\beta(u_i) \), find \( \beta \) such that \( \sum_{i=1}^{n} pf\beta(u_i)((1+\delta(u_i))^{\lambda+1} - 1)/(\lambda((\lambda+1)) \) is minimized. Maximum likelihood, minimum Hellinger distance, and minimum Kullback-Leibler Divergence correspond to \( \lambda = 0, -1/2, \) and 1, respectively (see Lindsay (1994)). Note that ML is an MD estimator. If \( n = 1 \), \( d(u) = 1 \) when \( u = u_{\text{obs}} \) and 0 otherwise.

In the cheetah example above, negative Hellinger distance is used to measure agreement with the sample of size one, i.e., \( g_S(\beta) = -2(1 - \sqrt{pf(f(u_{\text{obs}})))^2} \). Also for this example, due to limited computational resources, \( g_H(\beta) \) measures only the agreement in the first moment between the consistent and hypothesis distributions. This is accomplished by defining \( g_H(\beta) \) to be the negative Euclidean length of the difference between two sample mean vectors. The first is computed over a Monte Carlo sample of 20 realizations from \( U_\beta \) and the second from a size-20 sample drawn from \( U_{\beta_H} \). These definitions of \( g_S(\beta) \) and \( g_H(\beta) \) makes consistency analysis a penalized MD estimator.

2 Interpretation and Assignment of \( c_H \)

Consistency analysis uses a point assignment of the parameters to represent the substantive science theory, and \( c_H \) to represent the analyst’s priority weights of having the consistent distribution agree with the hypothesis and empirical distributions. Because these two objectives are being simultaneously optimized, consistency analysis does not force the analyst to “put his/her eggs all in one basket,” i.e., \( \hat{\beta} \) is not solely a function of prior knowledge nor solely a function of the sample.

The following heuristics can be used to guide the selection of a value for \( c_H \):

1. Setting \( c_H \) to 1.0 is equivalent to the opinion that a sample is unnecessary and hence is not gathered. Consequently, no consistency analysis is performed.
2. If $\beta_H$ values are taken from peer-reviewed articles reporting on experiments very similar to the one being modeled, and the sample is partly the result of interpretation by the collector, $c_H$ should be high, say between .5 and .9.

3. If $\beta_H$ values are based on analyst “hunches” and some element of science-based reasoning, and the sample is known to be partly based on interpretation, $c_H$ should be moderate, say between .3 and .7. An example would be when a review of the literature reveals disagreement over the impact of a driver variable on an ecosystem health indicator.

4. If $\beta_H$ is specified with little scientific justification, and the sample is known to be highly reliable and solely the result of calibrated instruments, $c_H$ should be small, say between 0.0 and .4.

5. If $\beta_H$ is completely unreliable relative to the sample, $c_H$ should be set to zero.

3 Consistency Analysis Procedures

Before an model should be trusted to aid any type of real-world decision, its reliability should be assessed. A model whose forecasts have excessively high variance (low predictive validity) should not be used to formulate policy. To this end, the RMSPE of policy-relevant output variables is always calculated to provide an assessment of the model’s prediction skill.

The consistency component of consistency analysis always entails the computation of (1) parameter estimates that are as consistent as possible with both prior theory and the sample, and (2) the RMSPE.

If an assessment of individual parameter estimate uncertainty is needed and computational resources permit, the Coefficient of Variation (CV) is also computed for each parameter estimate using the delete-$d$ jackknife (see Shao and Tu (1995)) where $d < n$. The $CV(\hat{\beta})$ is defined to be $S_{\hat{\beta}}/\hat{\beta}$ where

$$ S_{\hat{\beta}} = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left( \tilde{\beta}_{-i} - \frac{1}{n} \sum_{k=1}^{n} \tilde{\beta}_{-k} \right)^2} $$

(2)
and $\tilde{\beta}_{-i}$ is the consistency analysis estimate of $\beta$ when datum $i$ is temporarily removed from the sample (Shao and Tu 1995, p. 333). $CV(\tilde{\beta})$ values larger than (say) 70% indicate so much uncertainty as to render the estimated value of $\beta$ unreliable. Shao and Tu give recommendations for selecting $d$.

The analysis component of consistency analysis involves the following activities:

1. Examination of the RMSPE to determine the reliability of predictions computed from the fitted model.

2. If computed, examination of parameter estimate $CV$’s to assess parameter estimate reliability.

3. Assessment of the differences between hypothesis parameter values and the associated consistent values. Large absolute differences indicate areas of prior theory that do not agree with observation. Such theory should be critically reviewed and possibly modified.

4. Examination of dependency links that appear weak – indicated by consistent parameter values that change little as the conditioning variable values of a random variable are varied. Such a situation suggests a random variable is actually independent of its conditioning variables.

4 Computational Implementation of Consistency Analysis

Consistency analysis is implemented by the author as follows. When the model is an influence diagram, the logic sampling (Henrion 1988) algorithm is used to simulate realizations. The “complex” constrained optimization algorithm of Box (1965) is used to maximize $g_{CA}(\tilde{\beta})$ subject to the constraints of valid parameter values in $\beta$ and valid conditional distributions defined by $\tilde{\beta}$. The same sequence of random numbers is used for each evaluation of $g_{CA}(\tilde{\beta})$ to ensure smoothness of this objective function.
5 Comparison of Consistency Analysis With Other Methods

Consistency analysis can be viewed as a penalized goodness-of-fit parameter estimation method. See Easterling (1976) for a discussion of parameter estimation via maximization of a goodness-of-fit statistic. Viewed this way, the penalty function is the deviation from the hypothesis distribution.

Consistency analysis has similarities with Minimum Discrimination Information (MDI) parameter estimation (Gokhale and Kullback 1978). When the model is an influence diagram, the function relating model parameters to probabilities of the model’s joint events (called external constraints by Gokhale and Kullback) is given by the recursive factorization of the influence diagram’s joint probability distribution. Consistency analysis differs from MDI in that no functional relationships between model event probabilities and the empirical PDPF (called internal constraints by Gokhale and Kullback) need be satisfied. Also, in MDI the discrimination information statistic measures the distance between the estimated model and the maximum entropy model (constant PDPF over all joint events). The idea of MDI is to minimize this distance subject to satisfying the external and internal equality constraints. In consistency analysis, the maximum entropy distribution is replaced by the hypothesis distribution supplied by the analyst. Further, by manipulating the value of $c_H$, the analyst can change the relative emphasis consistency analysis gives to finding parameter estimates that agree with this hypothesis distribution relative to agreeing with the empirical distribution. Hence, in consistency analysis, no apriori relationships between the estimated model’s distribution and the empirical distribution need be satisfied.

In a multi-parameter estimation problem, classical bayesian methods require the analyst to specify an apriori joint distribution for the parameters (for convenience, independence is often assumed). For example, Dickey et al. (1987) require the analyst to specify a joint Dirichlet distribution for the parameters of a contingency table. Hence, not only is the expected value of each parameter needed, but also all higher moments and cross-moments. If the sample is small and/or incomplete, this very strong distributional assumption can have more influence than the sample on bayesian parameter estimates. In this case, justifying this prior, joint distribution of the parameters with arguments
from relevant scientific knowledge is critical to establishing the credibility of the resulting parameter estimates. Several areas of scientific inquiry however, lack sufficient theoretical knowledge to completely specify such a distribution. Instead, available theory can often only support the assignment of point values (“best guesses”) to the parameters with no reliable judgements about the the prior uncertainty of such values. One such area is the use of large, complex environmental process models to support environmental management decision making. In such situations, a bayesian approach requires more from the analyst than the analyst can justify.

Consistency analysis is one way to incorporate prior knowledge into parameter estimation without inheriting some of the criticisms of the application of bayesian methods to the model assessment problem. In addition to the above concern of excessive demands on the analyst’s prior knowledge, other criticisms of the bayesian approach to parameter estimation include (a) what constitutes a noninformative prior, (b) interpretability of an improper prior, and (c) reliable elicitation of priors. See Dennis (1996) for further discussion on the use of bayesian inference in ecological modeling.

From a bayesian perspective, MML is based on the belief that, before observing the sample, all parameter value combinations have the same chance of generating the sample (see Berger (1985, pp. 27, 132)). Further, MMLE’s of parameters that define unobserved random variables are the result of the likelihood of the data on the observed random variables being made as large as possible, i.e., since no information on the unobserved random variables is available, parameter estimates for the unobserved random variables are based only on the model’s structure and the observed random variables. In the absence of any substantive theory for the process, these characteristics of MMLE’s are reasonable. On the other hand, with prior substantive theory (including previous empirical studies), an MMLE completely ignores any prior or qualitative knowledge about the process and allows absurd combinations of parameter values to contribute the same weight to parameter estimation as combinations having substantive science justification. Through the use of $g_H(\beta)$, consistency analysis downweights such absurd combinations of parameter values without any need to place a prior joint distribution on the parameters.

References


Lindsay, B. (1994), “Efficiency versus Robustness: the Case for Minimum Hellinger Dist-
