Evaluation of Probabilistic Models for Categorical Variables

Sahyun Hong and Clayton V. Deutsch

Evaluation procedures are essential when justifying the reasonableness of the used probabilistic model or model parameters. Although there is no way to attain complete objective criteria, some important measures are described for the purpose of checking a categorical variable model. Checking the reproduced global mean is first. Closeness measures how close estimated probabilities are to true value. Checking the fairness of estimated probabilities is performed when first two criteria are met. Closeness and fairness measures are both related to the distribution shape of estimated probabilities and their behavior is depicted by changing the probability distributions.

Introduction

It is impossible to fully validate a model when applied to unknown true values and thus a model can be partially evaluated with actual observation (Oreskes et al, 1994). Although a model is checked and has a high agreement with observations, the estimated values or probabilities are always uncertain because a model may be too simplistic and data we have at hand contain errors. Nevertheless, constructed model is evaluated based on some objectivity. There is no way to attain complete objectivity, but there are some related considerations that are important.

Leuangthong et al (2004) reviewed minimum acceptance criteria for continuous variable model. The discrete nature of categorical variables and the lack of an ordering require different cross validation techniques than continuous variables (Deutsch, 2002). The prediction of categorical variable may be checked by (1) reasonable reproduction of input global proportions, (2) high correlation between true facies and estimated probability of true facies, (3) and fairness of the estimated probabilities. In addition to these basic criteria, reproduced variograms from multiple realizations should be checked with input variogram in a simulation setting. Large scaled facies proportion modeling or secondary data integration; however, do not have multiple realization of facies. They will be anchored into the hard data in a simulation mode for generating facies realizations. Before using them in simulation, the modeled proportion trends or secondary derived probability estimates should be assessed because invalid results can make wrong final geostatistical models.

Global Proportions

A quick way of validating the probabilistic model is to check the reproduction of global proportions. A declustering or debiasing must be considered if necessary in order to obtain representative proportions and these are compared with the reproduced ones from models. Reproduced global proportions are arithmetic average of the estimated probabilities of facies k at sample locations $\mathbf{u}_\alpha$, α=1,…,n:

$$p_k^{repro} = E\{p^*(\mathbf{u}_\alpha; k)\},\ \alpha = 1, ..., n \quad (1)$$

The reproduced proportions usually do not exactly honor the input proportion; however, large departures indicate that the considered probabilistic model had better be revisited.

Although global proportions are well reproduced, one should check the distribution shape of $p^*(\mathbf{u}_\alpha; k)$ as well since the uncertainty is more related to the probability distributions. Figure-1 shows some examples of distributions of $p^*(\mathbf{u}_\alpha; k)$. Representative global proportion of facies k is assumed to be 0.3. Averages of three different probability distributions $p^*(\mathbf{u}_\alpha; k)$ are all 0.3. Three cases may be equally acceptable in terms of good reproduction of global proportions. Further checks based on the probability distributions should be performed.

Closeness to True Facies

In addition to checking the global proportion reproduction, we should evaluate that how close the local estimated probabilities are to the true facies. There is no doubt that the predicted probabilities of facies k should be high at true facies k, and should be low at facies *not* k. Deutsch (1999) proposed a measure for this purpose termed as *closeness*:

$$C_k = E\left\{ p^*(\mathbf{u}_\alpha; k) \mid \text{true} = k \right\},\ k = 1, \ldots, K$$
$$\text{or} \qquad (2)$$
$$C_{\text{overall}} = C_k / K$$

whereby $p^*(\mathbf{u}_\alpha;k)$ are estimated probabilities of facies k=1,…,K at well locations $\mathbf{u}_\alpha$, α=1,…,n. This measure can be interpreted as local accuracy of estimated probabilities. Because the calculated C_k is an average of probabilities, it varies from 0 to 1 theoretically. The case of C_k = 0 represents the results completely contradict to the true values. This case; however, rarely happens since it is better to take a global proportions as a local estimate when C_k=0. The worst case is then $C_k = p_k$. The closeness measure will have the following bounds:

$$p_k \le C_k \le 1 \qquad (3)$$

As C_k gets close to 1, the estimated local probabilities of true facies become close to 1 and the probabilistic model under consideration would be regarded as of being accurate. Recall three cases shown in the figure-1. Probability distributions are separated based on the true facies from samples and they are shown in the figure-2. For examples, top row in the figure-2 represents the distributions of probability $p^*(\mathbf{u}_\alpha;k)$ from samples of true facies k, and middle row represents the distributions of probability $p^*(\mathbf{u}_\alpha;k)$ from samples of true facies *not* k. Bottom row represents the distribution of $p^*(\mathbf{u}_\alpha;k)$ from all samples. Sum of two distributions shown in the top and middle amounts to the distributions in the bottom row. Distribution shapes located in the top and middle row are examples; thus, any shape of distribution is possible as long as their sum amounts to the distribution in the bottom. By the definition of closeness in equation (2), C_k is a mean of the probability distribution shown in the top row of the figure-2. The figure-2 gives an idea that that closeness measure becomes higher (moves to the right) when the extracted probabilities are toward to 1 leading to the bimodal shape of overall probability distribution as shown in the bottom. Due to good reproduction of global proportion and high closeness, the case (c) would be considered as a good model.

There is a classical measure of accuracy that assigns the most likely facies with a maximum probability and then count the actual and predicted pixel number. Confusion matrix summarizes these results (Johnson and Wichern, 2002):

	Predicted pixel number		
Actual pixel number		Facies 1	Facies 2
	Facies 1	n_{11}	n_{12}
	Facies 2	n_{21}	n_{22}

For example, n_{11} is number of pixels correctly classified as facies 1 and n_{21} is number of pixels misclassified. Classification accuracy of facies 1 is then calculated by $n_{11}/(n_{11}+n_{12})$. This simple evaluation method does not consider the uncertainty or probabilities. Figure-3 illustrates drawback of accuracy calculation using the classical confusion matrix. Small 4 grid example is shown for illustration. Probabilities around 0.5 are assigned at each grid which introduces high uncertainty when assigning facies 1 or 2 with the largest probability. Classification accuracy is calculated as 0% for both facies. Based on 0% classification accuracy, it might be concluded that the secondary data is useless when one wants to use them for identifying facies or the considered probabilistic model is completely invalid. The classical accuracy does not account for the high degree of uncertainty indicated by the estimated probability just below and above 0.5.

Fairness

The closeness is an indication of how the locally estimated probabilities close to the true value. It is worth to evaluate the relative accuracy between facies being assessed and facies not being assessed. The locally estimated probabilities can be said to be "*fair*" if they reflect the true fraction of times the predicted facies occurs (Deutsch, 2002). For example, consider all locations where we predict 0.3 probability of sand then most fair case is that 30% of those locations are sand. Below or above from this 30% may be a problem in terms of fairness. Fairness check is conducted by comparing the actual fractions and predicted p. In its general form, the fraction is computed at a given predicted probability p by:

$$\text{Actual Fraction} = E\{I(\mathbf{u}_\alpha;k) \mid p(\mathbf{u}_\alpha;k) = p\},\ \forall p,\ k = 1,\ldots,K \quad (4)$$

where $p^*(\mathbf{u}_\alpha;k)$ are estimated probabilities at sample location $\mathbf{u}_\alpha$, α=1,…,n. p is a predicted probability. Indicator function I(·) takes either 0 or 1 depending on facies type k what we want to evaluate. Actual fraction at p is an average of indicator value under the predicted p. This actual fraction is calculated at every p values and then plotted against p. Figure-4 shows a schematic illustration of how to calculate the actual fraction at a given predicted p. Identify what facies is to be evaluated and build the distribution of probabilities $p^*(\mathbf{u}_\alpha;k)$ extracted from samples of true facies k (identify them as code 1) and from samples of true facies *not* k (identify them as code 0). In figure-4, densely and sparsely dashed lines represent each extracted probability distribution and they are negatively and positively skewed with varying degree of skewness, respectively. At every p value averaging indicators described in equation (4) is equivalent to calculating the ratios of frequencies ($f_1(p)$) from densely dashed line to the summed frequencies from densely and sparsely dashed lines ($f_1(p)+f_0(p)$). These ratios are computed at every specified p and plotted against p. Plotting the actual fractions versus predicted probability allows us to check fairness in one plot as shown in the right of figure-4. The proximity of the dots to the 45 degree line attests to the *fairness* of the local estimated probabilities.

Drawn distributions shown in the figure-4 are examples and fairness plot based on that distributions looks acceptable. Pattern of fairness plot depends on the distribution of estimated probabilities. Figure-5 demonstrates various pattern of fairness plots derived from different probability distributions. Artificial probability distribution funcitons are smoothly generated and actual fractions are calculated using these distributions at every 0.01 predicted probabilities that are shown as smooth curve in the right column of the figure-5. Top first and second cases have high estimated probabilities about true facies (negatively skewed) and their fairness plots are more or less resemble 45 degree line. Among two cases, first case has higher closeness than the second case; however, second case is more fair than the first case. Third case has uniform distribution of estimated probabilities that is little or no informative for recognizing true facies. The last in the figure-5 shows the case of highly contradictory to true facies and perhaps this case has very low closeness measure. Fairness plot is highly departed from ideal line.

The closeness can be described on the probability distribution plots: it is a mean of the probability distribution of code 1 in figure-5. The top case has the highest closeness and it may be concluded as the best model in terms of closeness measure. Fairness plot; however, looks departed from the ideal plot shown as thick gray line. In the second case, closeness is slightly shifted to the left because of relatively long tail of probability distribution and fairness plot becomes close to the diagonal line. Fairness check is aimed at evaluating whether or not the estimated probabilities *unfairly* tend to be extreme, very close to either 0 or 1 in which case the probabilistic model is regarded as too optimistic.

Fairness plot examples with real data are shown in Figure 6. Large scaled binary facies proportion maps are generated using moving window method. Size of moving window is set differently, e.g. 10%, 30% and 60% of domain size, which makes three different proportion maps and probability distributions. Bar charts in the left column represent the distribution of extracted probabilities and plotted dots in the right column represent the calculated actual fractions from each case of probability distributions. The first case is the most accurate case by closeness measure ($C_{k=1}$=0.9), for the probabilities are highly concentrated on either 0 or 1. Fairness calculation shows either 0% or 100% actual fractions for this case. There are no proportions from the probability distributions when predicted p is between 0.4 and 0.6. Actual fractions; thus, cannot be calculated for that range (shown as unfilled circle). It does not show close inclination to the 45 degree line despite of the highest closeness among three cases. The second is a moderately accurate case ($C_{k=1}$=0.81). The probability distributions are skewed to either 0 or 1, but they have non-zero frequencies over entire probability ranges leading to a long tail. Its fairness plot more tends to be diagonal than the first case. The third example is that the estimated probabilities are almost evenly distributed. Closeness measure becomes smaller than other two cases and fairness plot looks erratic.

Closeness and Fairness

Defined closeness and fairness are affected by the distribution of estimated probabilities. The former is to evaluate the absolute accuracy of local estimated probabilities and it is maximized when distribution

shape of the estimated probabilities tends to be bimodal. The latter is aimed at assessing whether or not estimated probabilities are overly confident. The highest closeness does not always indicate the most fair model. When the closeness is emphasized too much the fairness would be impaired.

Discussions

It is impossible to comprehensively validate a model with true values because true values are limitedly known. Instead, we can validate the robustness and consistency of models themselves. For this, some objective criteria for categorical variable model are described: reproduction of global proportion, closeness to true value, and fairness of the estimated probabilities. Checking the reproduced global proportions is a first step. Checking the closeness becomes a priority before evaluating the fairness. The model may be revisited if closeness is far less than the minimum, that is global proportions. Once the measured closeness is reasonably acceptable then the fairness check would be performed further for model validation. Trying to increase closeness makes the estimated probabilities be too close to 0 or 1 (overly confident) resulting in an impair of fairness. Preserving the fairness precludes a model being too optimistic and it could drop closeness measure. To conclude more fair model will be selected unless the closeness is significantly dropped.

References

N. Oreskes, K. Shrader-Frechette, and K. Belitz, 1994, Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, Vol. 35, No. 2.

O. Leuangthong, J. A. McLennan, and C. V. Deutsch, 2004, Minimum acceptance criteria for geostatistical realizations, *Natural Resources Research*, Vol. 13, No. 3.

R. A. Johnson and D. W. Wichern, 2002, *Applied multivariate statistical analysis*, Prentice Hall.

C. V. Deutsch, 2002, *Geostatistical reservoir modeling*, Oxford University Press, New York.

C. V. Deutsch, 1999, A short note on cross validation of facies simulation methods, *Centre for Computational Geostatistics*.

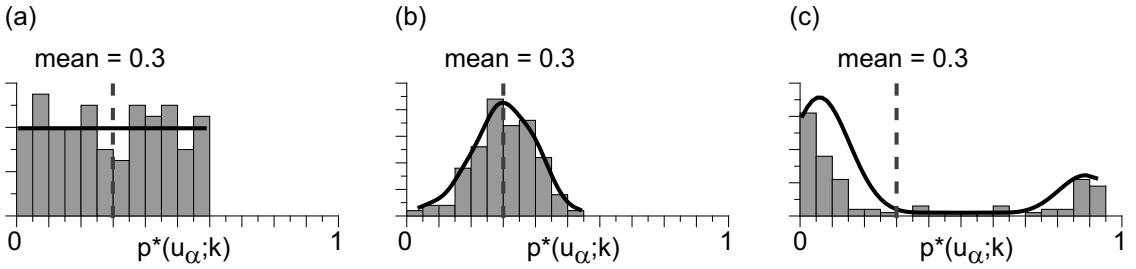

Figure-1: Three different distributions of the estimated probabilities $p^*(\mathbf{u}_\alpha;k)$. Averages of these distributions are all 0.3 that exactly reproduces the input global proportion.

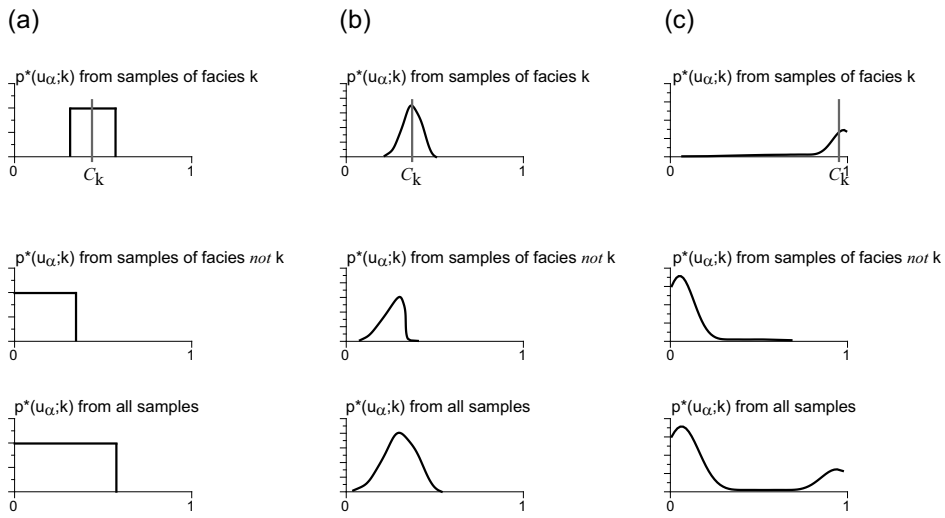

Figure-2: Probability distributions shown in the figure-1 are separated based on samples of true facies: distributions of the estimated probabilities $p^*(\mathbf{u}_\alpha;k)$ using samples of facies *k* (top), samples of facies *not k* (middle) and all samples (bottom).

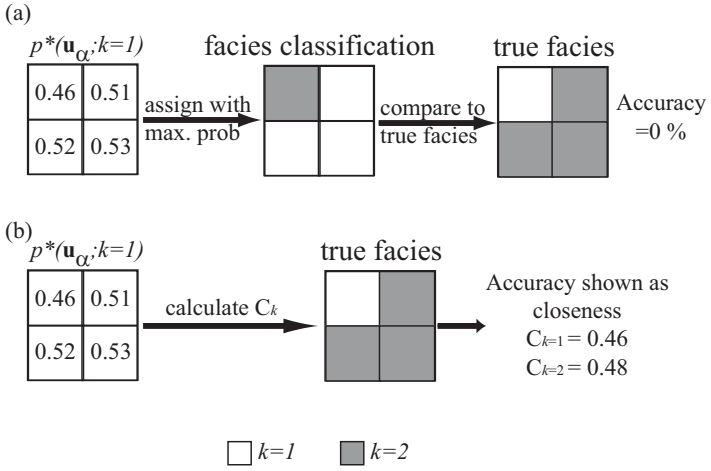

Figure-3: An example of accuracy measurement using traditional confusion matrix. Accuracy is calculated based on the fraction of pixels assigned to true values and it cannot quantify the uncertainty indicated by the estimated probability.

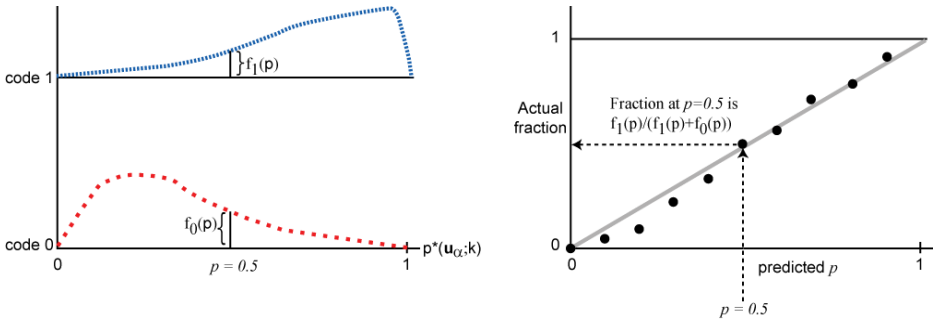

Figure-4: A schematic illustration of drawing the fairness plot. Based on distributions of estimated probabilities (left), ratios of these distributions are calculated at a given probability interval *p* and they are plotted against the considered *p* (right).

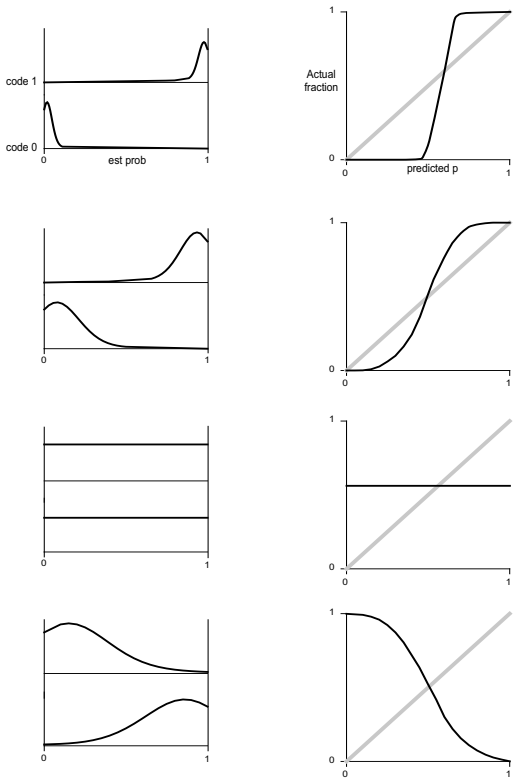

Figure-5: Various fairness plots based on the different probability distributions. Top first and second are acceptable cases and third and last are problematic cases. Among first two cases, closeness is high in the first case, but the second case is more fair.

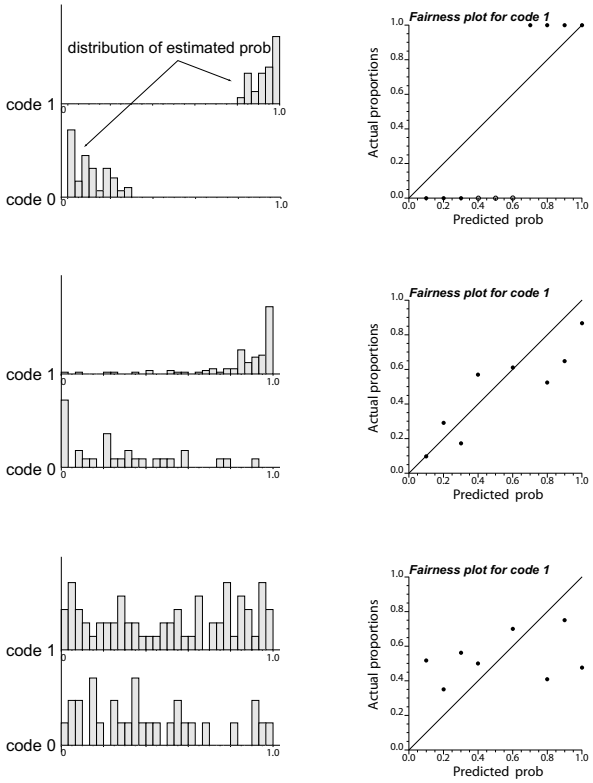

Figure-6: Fairness plots with test examples.