Power Functions and Their Use In Selecting Distance Functions for Document Degradation Model Validation

Tapas Kanungo†, Robert M. Haralick† and Henry S. Baird‡

† Department of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195, USA
{tapas,haralick}@ee.washington.edu

‡ AT&T Bell Laboratories
600 Mountain Avenue, Room 2C-322
Murray Hill, NJ 07974, USA
hsb@research.att.com

Abstract

Two document degradation models that model the perturbations introduced during the document printing and scanning process were proposed recently. Although degradation models are very useful, it is very important that we validate these models by comparing the synthetically generated images against real images. In recent past, two different validation procedures have also been proposed to validate such document degradation models. These validation procedures are functions of sample size and various distance functions. In this paper we outline a statistical methodology to compare the various validation schemes that result by using different distance functions. This methodology is general enough to compare any two validation schemes.

Keywords: Document Degradation Models, Model Validation, Distance Functions, Robust Statistics

1 Introduction

Two document degradation models were proposed recently that model the perturbations introduced during the document printing and scanning process. One models the physical image formation process [2], while the other directly models the spatial structure statistics of the degraded document image [7, 8]. Such image formation models are extremely useful in designing optimal noise removal algorithms, OCR algorithm development, experimental performance evaluation, etc.. Although degradation models are very useful, it is very important that we validate these models by comparing the synthetically generated images against real images. Otherwise, the OCR performance we get on synthetic images are not reliable estimates of the OCR performance one would get on real images.

Interestingly, two different validation procedures have also been proposed to validate such document degradation models [6, 5, 10, 9]. It is important that we know which validation scheme is better, not just by looking at the results, but by statistical arguments. Furthermore, both these validation procedures are functions of the sample size and choice of various distance functions used in intermediate steps. In this paper we use the power function to study compare the effects of sample sizes and distance metrics on the validation procedure. We study how the validation procedure behaves when one of the samples is corrupted with outlier, which happens quite often in a real world OCR setting.

In section 2 we describe the document degradation model validation problem and give a non-parametric hypothesis methodology for rejecting a model. The power function approach to evaluating validation procedures is given in section 3. In section 4 we give the various robust and non-robust distance functions that can be used in the hypothesis testing methodology described in section 2. Experimental results of the application of the validation methodology is given in section 4.

2 The Document Degradation Model Validation Problem

In this section we describe a non-parametric validation procedure that can be used to statistically validate any document degradation model. The details of the document degradation model and estimation of the model parameters can be found in

[7, 8, 6, 5]. Suppose we are given a sequence of real degraded characters $X = \{x_1, x_2, \ldots, x_N\}$, and another sequence of artificially degraded characters $Y = \{y_1, y_2, \ldots, y_M\}$ that were created by perturbing an ideal character with a document degradation model. We can assume that the characters x_i and y_i are binary matrices of size (approximately) 30×30. The question that needs to be addressed is whether or not the x_i's and y_i's come from the same underlying population. At this point it does not matter where the x_i's and the y_i's came from – x_i's and y_i's could both be artificially generated, or both be real instances, or one of them could be artificial and the other real. A statistical hypothesis test can be performed to test the null hypothesis that the underlying population distributions of x_i's and y_i's are the same.

Standard parametric hypothesis testing procedures (chi-square test etc) are not applicable since (i) the dimensions of x_i and y_i are not fixed, (ii) the vectors are binary and in particular not Gaussian, and (iii) the size of the space to which they belong is very large (approximately 2^{900} if we assume each character to be of dimension 30×30) Instead, we now describe a non-parametric permutation test (see [4, 3]) that will perform this hypothesis test.

1. Given (i) real data $X = \{x_1, x_2, \ldots, x_N\}$, (ii) synthetic data $Y = \{y_1, y_2, \ldots, y_M\}$, (iii) a distance function on sets, $\rho(X, Y)$, (iv) a distance function on characters, $\delta(x, y)$, (iv) size of test ϵ, (i.e. misdetection rate = 0.05).
2. Compute $d_0 = \rho(X, Y)$.
3. Create a new sample $Z = \{x_1, \ldots, x_N, y_1, \ldots, y_M\}$. Thus Z has $N + M$ elements labeled z_i, $i = 1, \ldots, N + M$.
4. Randomly permute (reorder) Z.
5. Partition the set Z into two sets X' and Y' where $X' = \{z_1, \ldots, z_N\}$ and $Y' = \{z_{N+1}, \ldots, z_{N+M}\}$.
6. Compute $d_i = \rho(X', Y')$.
7. Repeat steps 4, 5 and 6 K times to get K distances $d_1, \ldots, d_K$.
8. Compute the distribution of d_i's empirically: $P(d \geq v) = \#\{k | d_k \geq v\}/K$
9. Compute the P-value: $\epsilon_0 = P(d \geq d_0)$.
10. Reject the null hypothesis that the two samples come from the same population if $\epsilon_0 < \epsilon$.

The above procedure computes the null distribution of the distance function $\rho(X, Y)$ nonparametrically. In the standard parametric hypothesis testing procedure, the form of the distributions of x and y are known (usually Gaussian) and hence the null distribution of $\rho(X, Y)$ is known. In contrast, we compute the null distribution by randomly permuting the data set Z and creating a histogram of d_i's.

By design, the size of the test, ϵ, is fixed. Thus, irrespective of the distance function $\rho(X, Y)$, the percentage of times that the validation procedure will reject a true null hypothesis that the two samples are from the same underlying population is ϵ. In other words, the probability of mis-detection is ϵ. What is not fixed is the probability of false alarm, γ. Thus, although the use of various distance functions ρ and δ will give rise to the same probability of mis-detection, ϵ, each has a different probability of false alarm, γ, which is the probability that the procedure claims that X and Y come from the same underlying populations when, in fact, they come from different underlying populations.

It is important to note that if two samples X and Y pass the validation procedure, it does not mean that we accept the null hypothesis. Rather, it means that we do not have enough statistical evidence to reject the null hypothesis. But, when we reject a null hypothesis, it *does* mean that we have enough statistical evidence to reject the null hypothesis.

3 Power Functions

Let us assume that the x_i's are distributed as $F(\theta_X)$ and the y_i's are distributed as $F(\theta_Y)$, where θ_X and θ_Y are the parameters of the distributions. And, let the null hypothesis, H_N, and the alternate hypothesis, H_A, be:

$$H_N \quad : \quad \theta_X = \theta_Y \tag{1}$$
$$H_A \quad : \quad \theta_X \neq \theta_Y \tag{2}$$

The size of the test, ϵ, is fixed by the algorithm designer and is given as

$$\epsilon = P(H_A | H_N \text{ is true}) \; . \tag{3}$$

The plot of 1 minus the probability of false alarm as a function of θ is the power function. Thus, if we fix the distribution parameter of the x_i's at $\theta_X = \theta_0$, and vary the distribution parameter value $\theta_Y = \theta$ for y_i's, the power function is denoted by $\gamma_{\theta_0}(\theta)$, and is given by:

$$\gamma_{\theta_0}(\theta) = P(H_A | \theta_X = \theta_0 \text{ and } \theta_Y = \theta) \; . \tag{4}$$

Thus $1 - \gamma_{\theta_0}(\theta)$ is the probability of false alarm. The power function should have a minimum at $\theta_X = \theta_Y = \theta_0$, with $\gamma_{\theta_0}(\theta_0) = \epsilon$, and should increase on either side and go up to 1 when $\theta_Y = \theta$ is very far from θ_0.

Let us say there are two validation schemes A and B with test size ϵ and power functions $\gamma^A_{\theta_0}(\theta)$ and $\gamma^B_{\theta_0}(\theta)$. Since the size ϵ is same for both the schemes, A is better than B if the false alarm rate of A is less than the false alarm rate of B. That is A is better than B if $1 - \gamma^A_{\theta_0}(\theta) < 1 - \gamma^B_{\theta_0}(\theta)$ or $\gamma^A_{\theta_0}(\theta) > \gamma^B_{\theta_0}(\theta)$. If the above relation is true for all values of θ, then the procedure A is said to be uniformly more powerful than B. That is, the scheme A is better than scheme B if the power function plot of A is above the power function plot of B for all values of θ. Generalizing, if there are many validation schemes, the one whose power function is above all other power functions, is the best scheme. There is no clear winner if the power functions intersect – for some regions in the parameter space on scheme is better while in other regions the other is better.

For a given validation scheme, if we increase the sample sizes N and M, the power function changes and the new power function is higher than the old power function, and so by definition is more powerful.. Thus, the sensitivity, i.e, the width of the notch at the minimum, is a function of the sample sizes N and M. When the sample size is small, the notch is broader and when the sample size is large, the notch is sharper. This fact is used in deciding what sample size should be used: choose the sample size such that the desired probability of false alarm is attained when the parameters θ_X and θ_Y vary by a small (specified) amount $\Delta\theta$.

Finally, since our validation scheme described in the previous section is dependent on two distance functions ρ and δ. Each choice of ρ and δ gives rise to a power function. The combination that has the highest power function, is the best choice. See [1] for details on power functions.

4 Distance Functions, Outliers, and Robust Statistics

Various distance functions $\rho(X, Y)$ can be used for computation the distance between the sets of characters X and Y. We used the following symmetric distance functions for ρ.

Mean Nearest Neighbor Distance:

$$\rho_{Mean}(X,Y) = \frac{(\rho_{Mean}(Y;X) + \rho_{Mean}(X;Y))}{(N+M)}$$

where,

$$\rho_{Mean}(Y;X) = \sum_{x \in X} \left(\min_{y \in Y} \delta(x,y) \right)$$

$$\rho_{Mean}(X;Y) = \sum_{y \in Y} \left(\min_{x \in X} \delta(x,y) \right)$$

Trimmed Mean Nearest Neighbor Distance:

$$\rho_{Trim}(X,Y) = (\rho_{Trim}(Y;X) + \rho_{Trim}(X;Y))/2$$

where,

$$\rho_{Trim}(Y;X) = \mathrm{Trim}_{x \in X} \left(\min_{y \in Y} \delta(x,y) \right)$$

$$\rho_{Trim}(X;Y) = \mathrm{Trim}_{y \in Y} \left(\min_{x \in X} \delta(x,y) \right)$$

Here the *Trim* function accepts as input a set of real numbers, orders them in an increasing order, discards the top and bottom 10%, and returns the mean of the rest 80%.

Median Nearest Neighbor Distance:

$$\rho_{Med}(X,Y) = (\rho_{Med}(Y;X) + \rho_{Med}(X;Y))/2$$

where,

$$\rho_{Med}(Y;X) = \mathrm{Median} \left(\min_{y \in Y} \delta(x,y) \right)$$

$$\rho_{Med}(X;Y) = \mathrm{Median} \left(\min_{x \in X} \delta(x,y) \right)$$

Notice that the mean NN distance is not a robust distance measure. Thus, if for some reason, one of the data values becomes very large, the P-value computation will become very sensitive to this data point. This can occur if one of the characters in the real data set X is actually a 'c' (instead of being an 'e,'), and has been identified wrongly as 'e'. Yet another outlier source is connected characters: when characters are segmented from a real document, they might be touching other characters, pieces of which might slip in. The Median and the Trimmed Mean distance measures are robust against outliers since they do not look at the tails of the distribution. One would expect that these should work better in the case there are outliers.

The $\delta(x, y)$ mentioned in the distance between two individual characters x and y. We use the Hamming distance for δ. This is computed by counting the number of pixels where the characters x

and y differ after the centroids of x and y have been registered. A variety of other character distances, $\delta(x,y)$, and set distance functions, $\rho(X,Y)$, could have been used. (e.g. the Hausdorf distance, rank ordered Hausdorf distance, etc.) The combination of character distance $\delta(x,y)$, and set distance, $\rho(X,Y)$, that give rise to the best power function is the best pair of distances to use for the validation procedure.

5 Experimental Results

In this section we give experimental results on synthetic data: we use the validation procedure to distinguish two samples X and Y of degraded characters that were simulated with different parameter values. We change the sample size of the datasets and study the behavior of the power function. Next, we fix the sample size and run the validation procedure for various set distance functions $\rho(X,Y)$. The power functions of these procedures are then compared to choose which choice of distance function is the best. Finally, we corrupt one of the samples with outliers and then see how robust the power functions are against outliers.

The following protocol used for creating the samples X and Y. The distribution parameter θ_X of sample X was fixed with the following parameter component values: $\theta_X = (\eta_f, = \eta_b = 0, \alpha_0 = \beta_0 = 1$, $\theta_X = (\eta_f, \eta_b, \alpha_0, \alpha, \beta_0, \beta, k) = (0, 0, 1, 1.5, 1, 1.5, 5)$. For details of the degradation model, and the parameter values, please see [7, 8]. The distribution parameter θ_Y for sample Y was varied by changing $\alpha = \beta$. Other distribution parameter components of θ_Y – $\eta_f, \eta_b, \alpha_0, \beta_0, k$ – were same as that in the model parameter θ_X. In all cases the noise free document that was degraded using the model was the same (a Latex document page formatted in IEEE Transaction style) and the same set of 340 character 'e' (Computer Modern Roman 10 point font) were extracted from the page, for creating the sample X and the sample Y.

The validation procedure protocol was as follows: the significance level ϵ was fixed at 0.05; the sample sizes $N = M$ used were 10, 20, and 60; the number of permutation K for creating the empirical null distribution was 1000; the number of trials T for estimating the misdetection rate was 100.

A degraded document generated with model parameter θ_X is shown in Figure 2 (a). The power function for the sample sizes 10, 20, 60 are shown in Figure 1. The power function corresponding to sample size 10 is the widest, and the power function corresponding to sample size 60 is the narrowest. Note all the three power functions give a misdetection (reject) rate close to $\epsilon = 0.05$ when the sample Y has a parameter value close to that of X, which has θ_X such that $\alpha = \beta = 1.5$. Furthermore, when the θ_Y has $\alpha = \beta$ far from 1.5, the misdetection rate is close to 1, which implies that the validation procedure can distinguish the two samples with high probability. An images of a document generated with parameter values $\alpha = \beta = 2.0$, which the validation procedure could distinguish from one that was generated with $\alpha = \beta = 1.5$, (shown in Figure 2 (a)) is shown in Figure 2 (b). In these experiments the data sets X and Y did not contain any outliers.

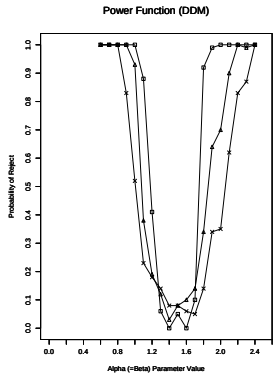
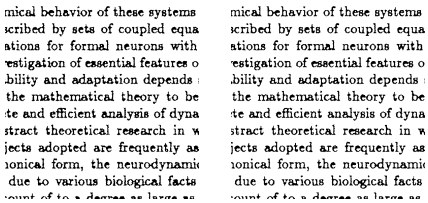


Figure 1: Power plots as a function of sample size $N = M = \{10, 20, 60\}$. The sample population parameter had $\alpha = \beta = 1.5$. Notice that the power function has a minimum near $\alpha = \beta = 1.5$. There were no outliers in either of the samples.



(a) (b)

Figure 2: (a) Subimage of a document. degraded with parameter values $\alpha = \beta = 1.5$. (b) degraded document, simulated with $\alpha = \beta = 2.0$. The two samples were judged dissimilar (null hypothesis was rejected) by the validation procedure. Sample size used was 60.

Next, we studied the sensitivity of the validation procedure to the set distance $\rho(X,Y)$ as follows.

The data sets X and Y are collections of (synthetic) degraded character 'e'. Degradation parameter values for X were fixed at $\alpha = \beta = 1.5$, but the degradation parameters for Y were varied from 0.6 to 2.4. Hamming distance was used for the character-to-character distance, $\delta(x, y)$. Sample size of X and Y was fixed at $N = M = 60$. The mean, trimmed mean and median distances were used to compute the power function, both, in the presence and in the absence of outliers.

Figures 3 (a), 4 (a), and 5(a), show the power functions in the absence of outliers when the mean, trimmed mean distances were used. Next, we introduced outliers in the dataset X but substituting 5 degraded 'e's with degraded 'c's. The Y data set was unchanged. Figures 3 (b), 4 (b), and 5(b), show the power functions in the presence of outliers. Clearly the median and trimmed mean nearest neighbor distances are more robust against outliers, since the corresponding power functions are not affected. Furthermore, it can be seen that the median NN distance function, in the outlier-free case, it is less 'powerful' than the mean distance function since the function lies below the mean NN power function plot. Finally, it can be seen that the 10 % trimmed NN distance function is superior to the other two distance functions, since the corresponding power function is robust against outliers and at the same time higher.

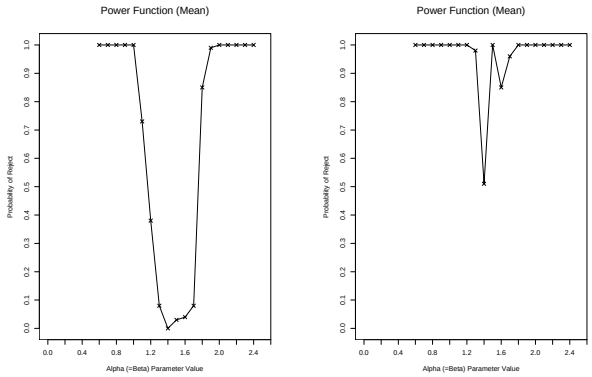


Figure 3: Power functions of the validation procedure when mean nearest neighbor distance is used for the set distance functions $\rho(X, Y)$. Figure (a) is when there are no outliers. Figure (b) corresponds to the situation when there are 5 outliers in one of the data sets.

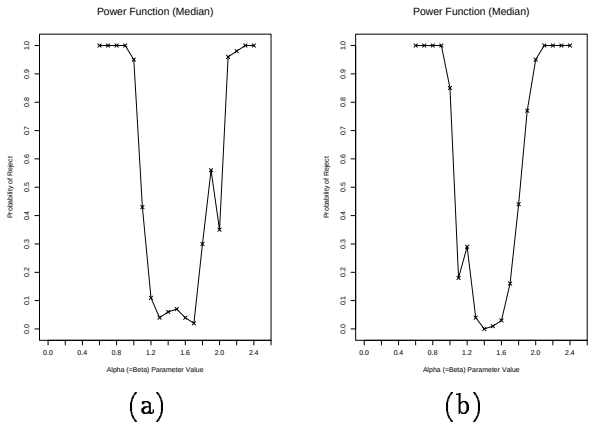


Figure 4: Power functions of the validation procedure when median nearest neighbor distance is used for the set distance functions $\rho(X, Y)$. Figures (a) is when there are no outliers. Figure (b) corresponds to the situation when there are 5 outliers in the X data set.

6 Summary

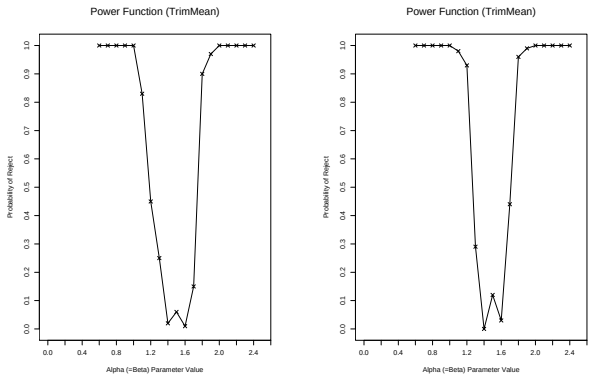
We proposed the use of power functions for comparing validation procedures for document degradation models and for selecting distance functions used in the validation procedure. Power functions allow us to compare validation procedures with same significance level by comparing their false alarm rates. Furthermore, since there is no justification for assuming a parametric form for the null distributions, we adopt a non-parametric hypothesis testing methodology. The method is general enough that any two validation schemes can be compared. Experimental results show that in the presence of outliers, the trimmed mean nearest-neighbor distance is the best.

Acknowledgement: Authors would like to thank Werner Stuetzle and David Madigan for discussions on permutation tests.

References

[1] S. F. Arnold. *Mathematical Statistics.* Prentice-Hall, New Jersey, 1990.

[2] H. Baird. Document image defect models. In *Proc. of IAPR Workshop on Syntactic and Structural Pattern Recognition*, pages 38–46, Murray Hill, NJ, June 1990.



(a) (b)

Figure 5: Power functions of the validation procedure when 10% trimmed mean nearest neighbor distance is used for the set distance functions $\rho(X,Y)$. Figures (a) is when there are no outliers in the data X and Y. Figure (b) corresponds to the situation when there are 5 outliers in the X data set.

[3] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993.

[4] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New York, 1994.

[5] T. Kanungo and R. M. Haralick. Morphological degradation parameter estimation. In *SPIE Proceedings*, San Jose, CA, February 1995.

[6] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, and D. Madigan. Document degradation models: Parameter estimation and model validation. In *Proc. of Int. Workshop on Machine Vision Applications*, Kawasaki, Japan, December 1994.

[7] T. Kanungo, R. M. Haralick, and I. Phillips. Global and local document degradation models. In *Proc. of Second International Conference on Document Analysis and Recognition*, pages 730–734, Tsukuba, Japan, October 1993.

[8] T. Kanungo, R. M. Haralick, and I. Phillips. Non-linear local and global document degradation models. *Int. Journal of Imaging Systems and Technology*, 5(4), 1994.

[9] Y. Li, D. Lopresti, and A. Tomkins. Validation of document defect models for optical character recognition. In *Proc. of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 137–150, Las Vegas, Nevada, April 1994.

[10] G. Nagy. Validation of ocr data sets. In *Proc. of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 127–135, Las Vegas, Nevada, April 1994.

A Null Distributions and Power Functions for Gaussian Samples

In this appendix we compute the null distributions and power function associated with the inter cluster mean distance $\rho(X,Y)$ when x_i and y_i are Gaussian distributed. Let $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in R$, and $x_i \sim N(\mu_X, \sigma^2)$. Similarly, let $Y = \{y_1, y_2, \ldots, y_N\}$, where $y_i \in R$, and $y_i \sim N(\mu_Y, \sigma^2)$. Here σ^2 is known, and μ_X and μ_Y are unknown. The problem is to test the null hypothesis,H_N, that $\mu_X = \mu_Y$, against the alternate hypothesis, H_A, that $\mu_X \neq \mu_Y$.

Now, we know that

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \sim N(\mu_X, \sigma^2/N)$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \sim N(\mu_Y, \sigma^2/N) .$$

Therefore,

$$\bar{x} - \bar{y} \sim N(\mu_X - \mu_Y, 2\sigma^2/N) .$$

Now, let $d = \rho(X,Y) = N(\bar{x} - \bar{y})^2/(2\sigma^2)$. Thus under the null hypothesis that $\mu_X = \mu_Y$, we have

$$d \sim \chi^2_1 .$$

Thus, instead of empirically computing the distributions as described in section 2 we can use the above distance function and the corresponding analytic form of the null distribution to test the null hypothesis.

The power function is the distribution of the test statistic under the alternate hypothesis. Here the two sample test statistic d is distributed as the noncentral chi-square distribution $\chi^2_{1,b}$ where the noncentrality parameter $b = N(\mu_X - \mu_Y)^2/(2\sigma^2)$.