THE EVIDENTIAL VALUE OF MICROSPECTROPHOTOMETRY

Kenneth Annan

Introduction

This study seeks to investigate the evidential value of microspectrophotometry measurements made for pen inks) in Martyna et al. (2013). Our interest is in the three of the univariate LR's (corresponding to the "X", "Y", and "Z" measures of color) that Dr. Saunders constructed and their relationship to an omnibus LR. That is, the objective is to describe and characterize the relationship between the four forensic likelihood ratios, with the ultimate aim of predicting the omnibus Likelihood ratio in terms of the three univariate likelihood ratios. The basic goal is to investigate 1) whether or not there is a relationship between the marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) and the Omnibus LR (i.e. Omni.LLR.int) 2) characterize the relationship if any, and draw conclusions from it.

Data and methodology

The study employs the dat.LLR.int.csv dataset that Dr.Saunders constructed. The data set contains 6 variables and 820 observations. However, the present study was initiated by a data collection consisting of four forensic likelihood ratios (Omni.LLR.int, LLR.x, LLR.y, LLR.z) and Type(within-source comparison(wi) and between-source comparison(BW)). The within-source comparison (wi)consists of just 40 observations and the rest of the 780 observations are made up of between-source comparison (BW).In order to explore the dataset to know what model will best fit the data, I used the ggplot, the boxplot to check. I also used the aggregate function in R to find the means of the wi and by comparison. The boxplot shows that in all four measurements, the within-source comparison(wi) has a higher value than the between-source comparison(BW). This is also shown in the means of the four variables shown in Table 2. Also, the within-source comparison(wi) is less skewed as compared to the between-source comparison(BW) for all the four LLR measurements. The above discussion gives an indication of dividing the dataset for the analysis. Thus, I analyzed these effects separately for the within-source comparison(wi) and the between-source comparison(BW). Also, the correlation plot of the variables in Figure 2a and Figure 2b shows that all the independent variables (LLR.x, LLR.y, and LLR.z) have a non-linear relationship or pattern with the Omnibus LR. From the ongoing discussion, I posit that aside from the non-relationship between the variables, a simple linear model cannot be sufficient to characterize the relationship between the dependent variable and the three marginals. Thus, to better characterize this relationship, I used the generalized additive models (gam) model to measure this non-linear relationship. I used the gam. checker command that I learned in chapter 6 of this course to further check my results.

Discussion of results

within-group comparison

The summary result of the gam model for the within-group comparison shows that all the three marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) variables are statistically significant at the predefined 5% alpha level since their p-values are less than 0.05. The plot of the variables also shows that the marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) and the Omnibus LR (i.e. Omni.LLR.int) have a non-linear relationship. Thus, there exists a no linear relationship between the Omnibus and the three marginal LR's (ie, a polynomial pattern). Also, the variable "LLR.x" as seen from the plot and by the degrees of freedom nearing 8, can be explained by an 8th order smooth function, "LLR.y" by approximately 8th order smooth function, and "LLR.z" by a 9th order smooth function-see Table 2. The moderate GCV, AIC, and degrees of freedom for the model are 0.0000018,-447.0405324, and 25.5012195 respectively-See Table 3. The total degree of freedom (25.5012195) is approximately 26 to the nearest whole number indicates that we need to impose smoothing terms as well. This is to ensure that the degrees of freedom is equal to the number of smoothing terms to ensure an optimum model. However, the AIC of -447.0405324 is a bit low implying that there is no need to smooth the model. The summary of the gam.check() function also shows that none of the marginal LR's (i.e. LLR.y, and LLR.z) are significant at the at the predefined 5% alpha level.The diagnostics plot of the gam.check() function show that all the independent variables should be smoothed because all of the predictors show a non linear effect.Also, the residual plots show that residuals are approximately normally distributed albeit slightly right-skewed.Lastly, the choice of k is appropriate for the within group comparison because they are not strictly low.Thus, the smoothing is order 9 for all the independent variables.

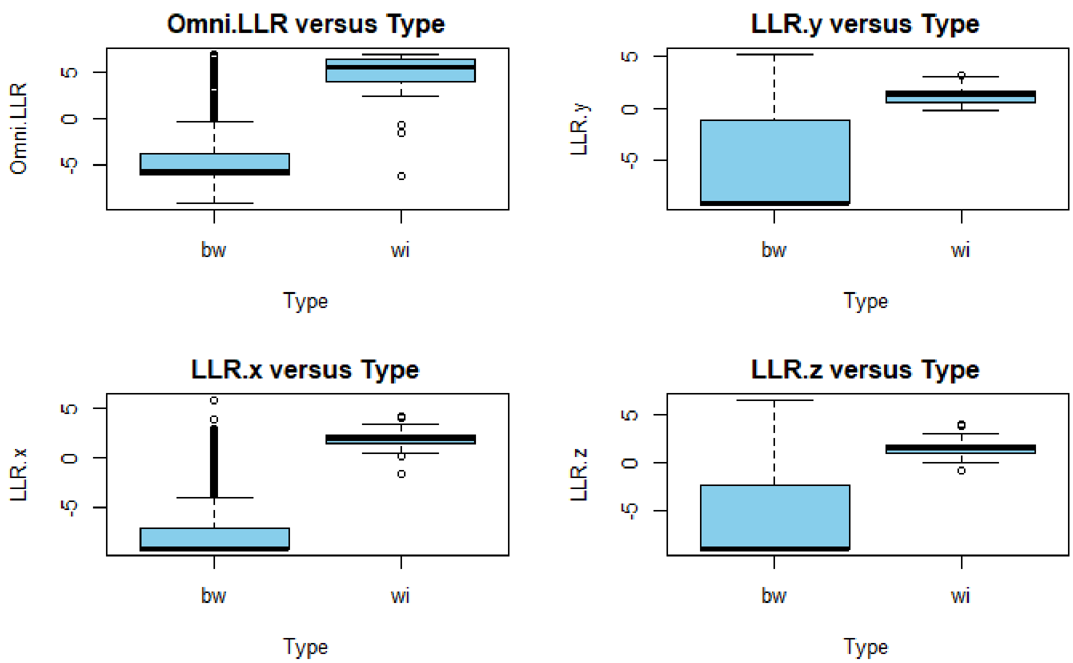
Between group comparison

The summary result of the gam model for the between-source comparison shows that all the three marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) variables are statistically significant at the predefined 5% alpha level since their p-values are less than 0.05. The plot of the variables also shows that the marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) and the Omnibus LR (i.e. Omni.LLR.int) have a non-linear relationship. Thus, there exists no linear relationship between the Omnibus and the three marginal LR's (ie, a polynomial pattern). Also, the variable "LLR.x" as seen from the plot and by the degrees of freedom nearing 9, can be explained by a 9th order smooth function, "LLR.y" by approximately 8th order smooth function, and "LLR.z" by 8th order smooth function. The moderate GCV, AIC, and degrees of freedom for the model are 0.0373758, -349.0346052, and 25.5817165 respectively-See Table 4. The total degree of freedom (25.5817165) is approximately 26 to the nearest whole number indicates that we need to impose smoothing terms as well. This is to ensure that the degrees of freedom is equal to the number of smoothing terms to ensure an optimum model. However, the AIC of -349.0346052 is a bit low implying that there is no need to smooth the model. The summary of the gam.check() function also shows that two of the marginal LR's (i.e. LLR.y, and LLR.z) are significant at the predefined 5% alpha level while the variable LLR.x is not significant. The diagnostics plot of the gam.check() function shows that all the independent variables should be smoothed because all of the predictors show a non-linear effect. Also, the residual plots show that residuals are approximately normally distributed albeit slightly left-skewed. Lastly, the choice of k is appropriate for the variable because they are not strictly low.

Conclusion

Both the between-source comparison and the within-source comparison of the gam model indicate that all the three marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) variables are statistically significant in predicting the Omnibus LR (i.e. Omni.LLR.int). Also, all the three marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) have a non-linear relationship with the Omnibus LR (i.e. Omni.LLR.int) for both comparisons. Lastly, the study sheds light that we can apply the gam model to estimate the omnibus LLR value given the values of the three marginal LLR's (i.e. LLR.x, LLR.y, and LLR.z). Also, the gam model fits the data well for between-source comparison and the within-source comparison as shown in Table 5.

Appendix



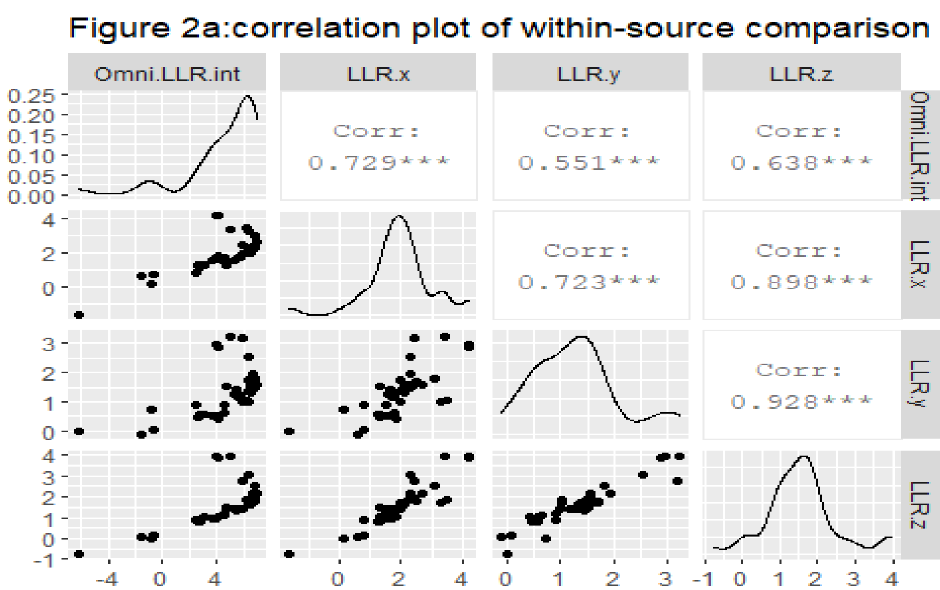


Figure 2b:correlation plot of between-source comparison

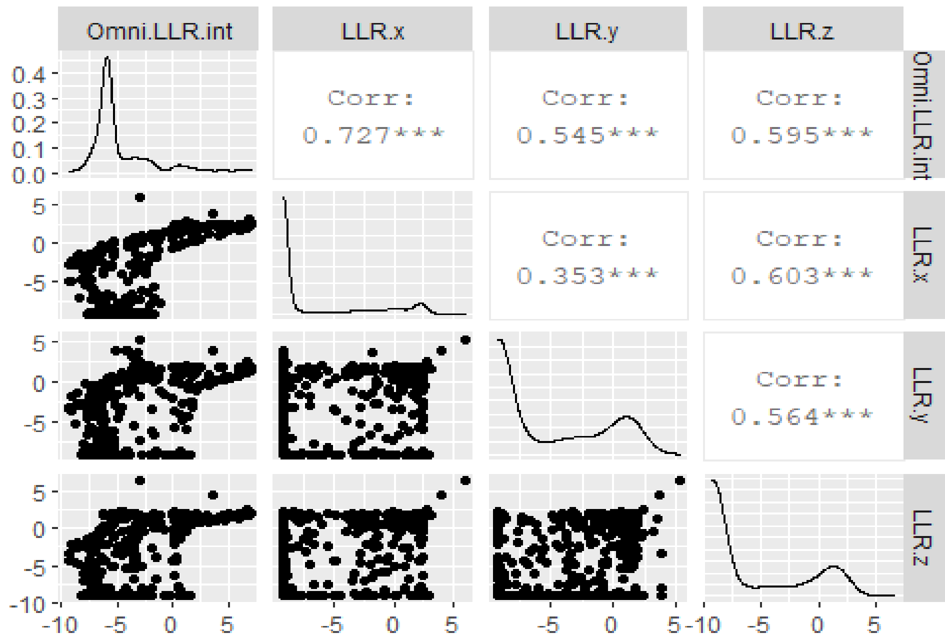


Table 1:Means of the four variables by type

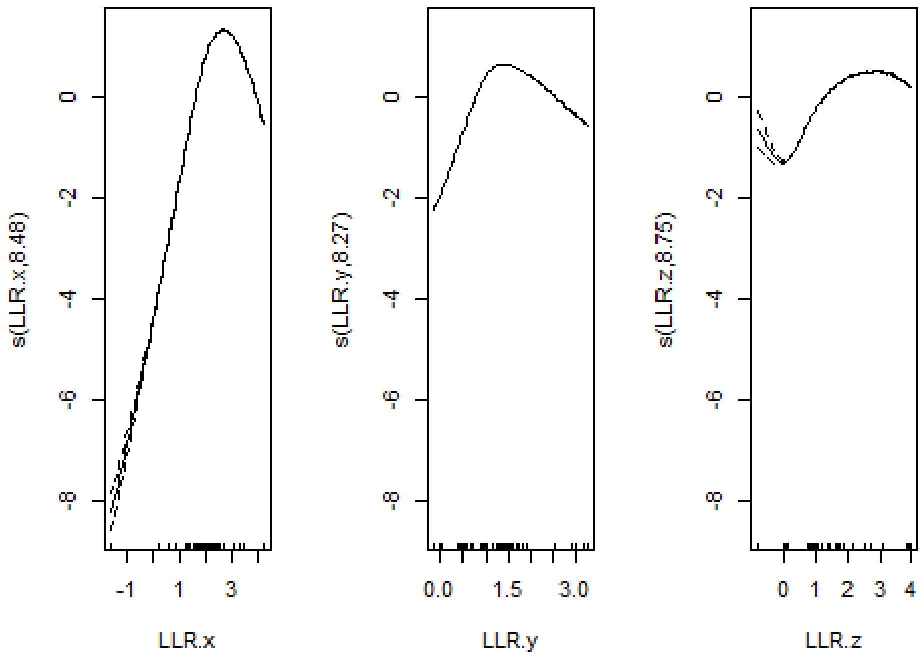
Group.1	Omni.LLR.int	LLR.x	LLR.y
bw	-4.576042	-6.963264	-5.699931
wi	4.576042	1.944720	1.284734

```
## [1] -447.0405

## [1] 0.9999999
```

Table 2:summary for within-source comparison

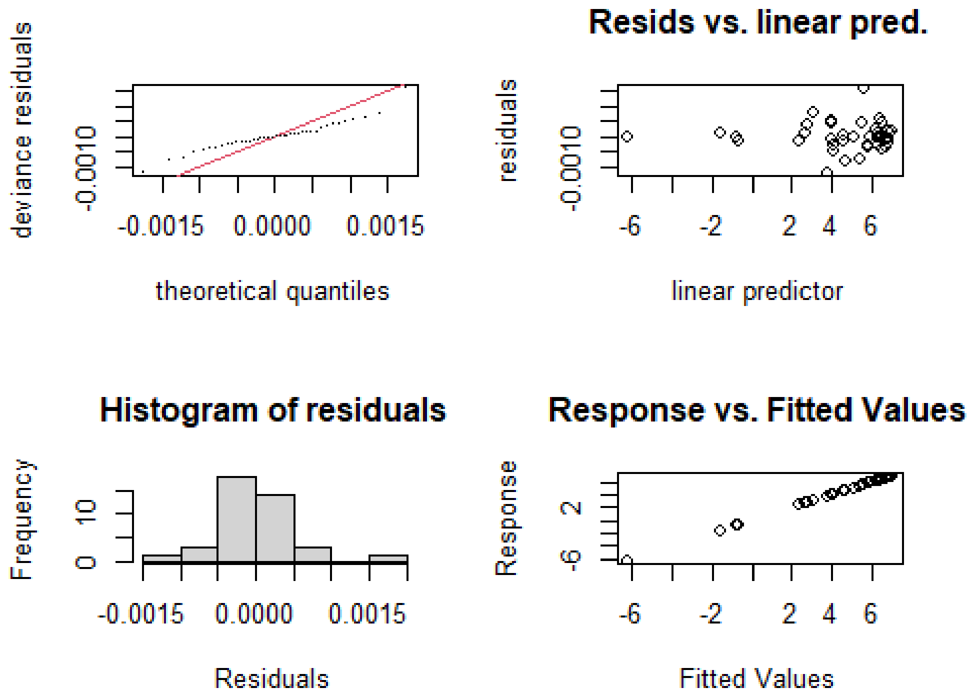
	edf	Ref.df	F	p-value
s(LLR.x)	8.477581	8.705807	522555.99	0
s(LLR.y)	8.269693	8.519370	114977.75	0
s(LLR.z)	8.753945	8.851928	23484.36	0




```
##       GCV.Cp
## 1.820515e-06

## [1] -447.0405

## [1] 25.50122
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 14 iterations.
## The RMS GCV score gradient at convergence was 1.083115e-07 .
## The Hessian was positive definite.
## Model rank =  28 / 28
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##              k'  edf k-index p-value
## s(LLR.x) 9.00 8.48    1.12    0.72
## s(LLR.y) 9.00 8.27    1.06    0.61
## s(LLR.z) 9.00 8.75    1.11    0.74
```

Table 1:Results for the GCV,AIC,R-squared,DF

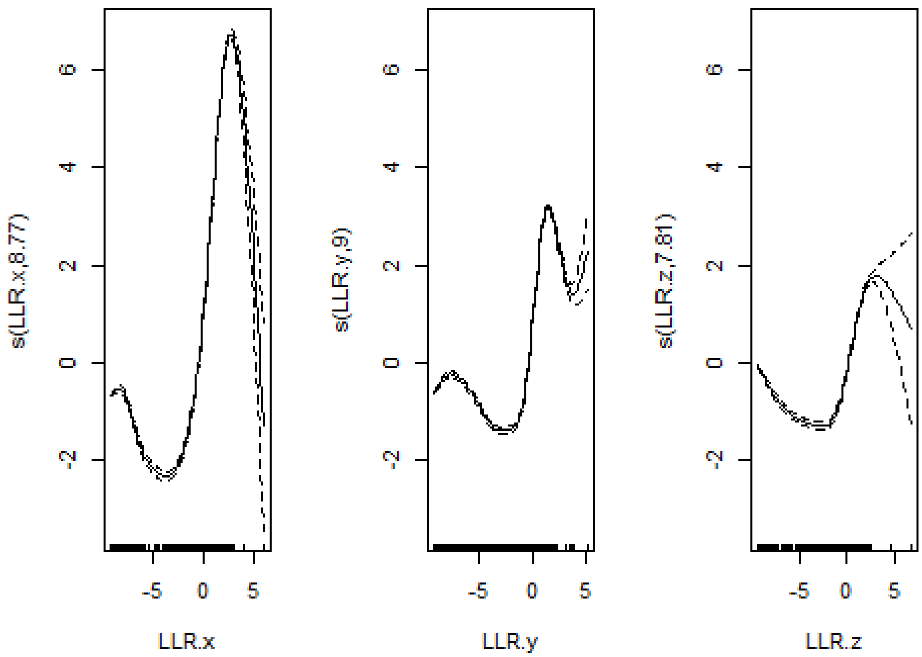
GCV	0.0000018
AIC	-447.0405324
R-squared	0.9999999
DF	25.5012195

```
## [1] -349.0346

## [1] 0.9960881
```

Table 3:summary for between-source comparison

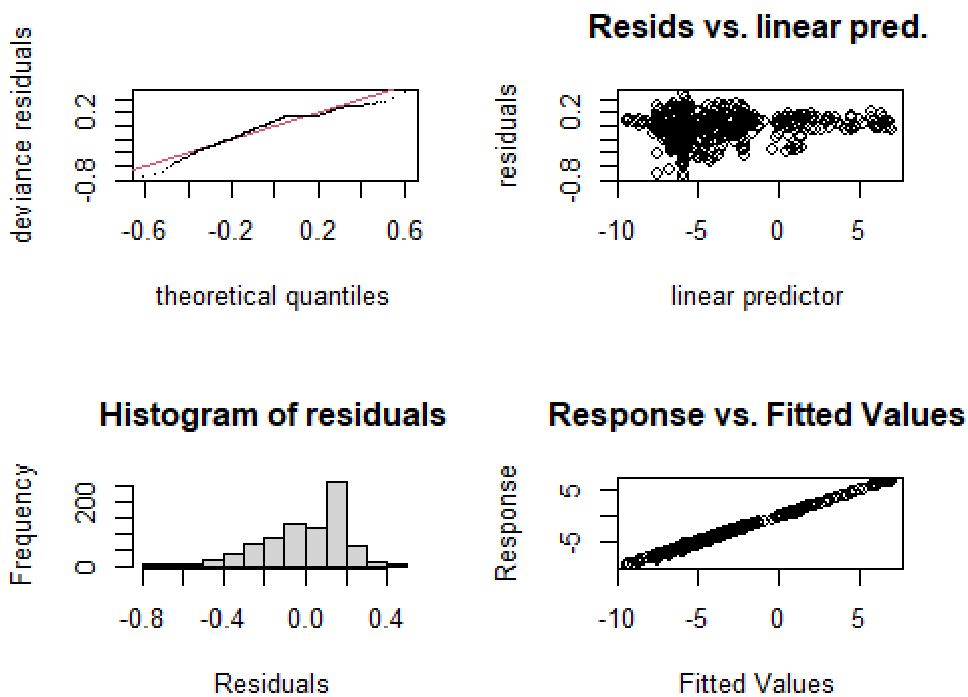
	edf	Ref.df	F	p-value
s(LLR.x)	8.774363	8.933637	8846.259	0
s(LLR.y)	9.000000	9.000000	3725.410	0
s(LLR.z)	7.807354	8.490809	1057.077	0




```
##      GCV.Cp
## 0.03737576

## [1] -349.0346

## [1] 25.58172
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 19 iterations.
## The RMS GCV score gradient at convergence was 1.350631e-07 .
## The Hessian was positive definite.
## Model rank =  28 / 28
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##             k'  edf k-index p-value
## s(LLR.x) 9.00 8.77    0.96    0.11
## s(LLR.y) 9.00 9.00    0.64  <2e-16 ***
## s(LLR.z) 9.00 7.81    0.45  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2:Results for the GCV,AIC,DF

GCV	0.0373758
AIC	-349.0346052
R-squared	0.9960881
DF	25.5817165

```
## [1] 0.9960881
```

Table 5:within comparison between comparison summary

	within comparison	between comparison
GCV	0.0000018	0.0373758
AIC	-447.0405324	-349.0346052
R-squared	0.9999999	0.9960881
DF	25.5012195	25.5817165

References

1. Zadora, G., Martyna, A., Ramos, D., & Aitken, C. (2013). Statistical analysis in forensic science: evidential value of multivariate physicochemical data. John Wiley & Sons.

2. Hothorn, T., Everitt, B. S., Data II, C. A. L., Scaling, C. M., & Partitioning, C. R. (2017). HSAUR3: A Handbook of Statistical Analyses Using R.