

# Analysis of Recognition Algorithms using Linear, Generalized Linear, and Generalized Linear Mixed Models

Geof H. Givens  
Dept. of Statistics  
Colorado State University  
Fort Collins, CO 80523

J. Ross Beveridge Bruce A. Draper David Bolme  
Dept. of Computer Science  
Colorado State University  
Fort Collins, CO 80523

## Abstract

*This paper discusses the design and use of linear, generalized linear, and generalized linear mixed models for evaluation and comparison of human face recognition algorithms. These models are introduced in a cohesive framework, and their benefits (compared to simple one-way descriptive comparisons) are reviewed. Several example analyses involving algorithm configuration and subject covariates illustrate the importance of using models that allow one to control for confounding variables, to estimate all important effects including interactions, and to isolate extraneous sources of variation.*

## 1. Introduction

Algorithm comparisons offer opportunities for high-quality experimental science that are too frequently ignored. The human face recognition problem in particular has a rich history [26, 3], has motivated many algorithms [3, 20], and has led to substantial evaluation efforts [20, 24, 1, 25, 5]. Nonetheless, two principal questions still facing the field are:

1. How should we compare algorithms to draw defensible, accurate, and detailed conclusions about performance, and for what conditions and sets of parameters will these conclusions hold?
2. What makes some people and images easier to recognize than others, and how should we properly draw such conclusions?

Unlike in many areas of science, scientists studying human ID algorithms have the opportunity—indeed the luxury—of being able to design precisely the experiments they wish to address such questions. They are therefore provided the freedom to model and analyze the resulting data in highly informative ways. Why, then, do so many analyses consist of simple one-way comparisons, charts, and tables with little use of the formal tools of statistical inference?

The experiments we refer to are algorithm tests, and must be distinguished from data (image) collection. Collecting large and representative sets of Human face imagery is difficult, time consuming and expensive; few institutions have the resources to do it. Fortunately, modest public data sets are already available [6, 9, 18], and larger data sets should become available with time. That said, often these datasets are what statisticians refer to as ‘samples of convenience’. For example, images are taken for every graduate student or co-worker who volunteered, and although attention is paid to controlling factors such as pose and illumination, very little control is imposed over factors having to do with the choice of human subjects. Therefore, from a statistician’s point of view, the resulting image collections are typically a mess: they are highly unbalanced relative to many subject factors and representativeness must be carefully studied. For precisely these reasons, simplistic analysis tools can often be misleading.

What the analyst may control is how various algorithms are tested on the image database. For example,

- Which subset(s) of Human subjects, and images, are used for algorithm training and algorithm testing?
- Which combinations of algorithm configuration variables are tested?
- Which factors are controlled when drawing inference about the influence of other factors?

These questions bear importantly on conclusions, and depending upon the specific goals of a particular analysis, different statistical techniques or combinations of techniques may be most appropriate.

This paper demonstrates three standard statistical techniques that are under utilized in computer vision: the linear model (LM), the generalized linear model (GLM) and the generalized linear mixed model (GLMM). The use of these statistical models to disambiguate the effects of possibly confounding factors is illustrated in two studies. The first study examines whether information about human subjects, such as age or gender, indicates the subject is easier

or harder to recognize [7]. The second study examines different configurations of a PCA+LDA recognition algorithm to see what design choices do, or do not, lead to improved recognition performance [8].

The human subject covariates study investigates whether specific covariates predict recognition difficulty for a standard PCA nearest neighbor classifier [13]. FERET data [20] is used in this study and we have prepared the following subject covariates: age, race, gender, skin complexion, glasses, facial hair, makeup, bangs, facial expression, mouth position, and eyelid position. Our primary study uses a linear model to relate recognition difficulty to these covariates. The study is novel in using distance between paired images of subjects to predict recognition difficulty, and a companion analysis using a generalized linear model establishes the necessary relationship between distance and recognition rank.

Our analysis of human subject covariates fails to find evidence that men are either more or less easily recognized than women. This contrasts with results presented in [21] using a simpler analysis of gender only. This apparent discrepancy illustrates how it is easy to be fooled by ‘statistical evidence’ when using an unsophisticated study design. Our data also indicates a gender effect when other subject covariates are not taken into account. Another conclusion of our study is that Asian, African-American and Other race subjects are more easily recognized than Caucasians (Whites). This effect persists even when training is balanced to compensate for the predominance of white subjects and illustrates how to control for issues of disproportionate representation.

A generalized linear mixed model is used to analyze the algorithm configuration space of a PCA+LDA [30, 1] face recognition algorithm. The generalized linear mixed model allows us to account for subject variation. Subjects are important because most of the observed variation in recognition performance is due to the fact that some humans subjects are harder than others. Indeed, variation attributable to subject accounts for roughly 70% of variation in rank-1 recognition rates.

The study of PCA+LDA has resulted in several important conclusion. Perhaps foremost, while the changes in algorithm configuration resulted in statistically significant changes in performance for subjects included in the training set, none of the alternative configurations were better or worse than any other for subjects outside the training set. From the standpoint of algorithm design, this is an important although perhaps disappointing conclusion. It nonetheless underscores the need for carefully designed experiments in order to truly measure advancements in algorithm design.

## 2. Linear, Generalized Linear, and Generalized Linear Mixed Models

Most readers will already be familiar with the linear modeling framework, i.e. multiple regression [16]. We review it here, as a special case of a generalized linear model (GLM) [14]. Let  $Y$  be a random variable representing a response, i.e. a quantity measured to evaluate the performance of a single attempt by a single algorithm faced with a single recognition task. For example,  $Y$  might be a binary variable, 1 for a correct match, and 0 otherwise. Alternatively,  $Y$  might encode recognition rank as a an integer-valued variable, or as shown below, it may be the distance between two images of the same subject that is used by a nearest neighbor classifier.

Let  $X$  denote a vector of independent variables with which we hope to predict the response. In a configuration study, these variables may be adjustable tuning parameters used to control the performance of the algorithm. In an environment- or subject-covariate study, these may be attributes of the images or subjects themselves: for example, illumination intensity, pose, race, or gender. Categorical predictors contribute vectors of binary indicator variables to  $X$  in the usual regression fashion [16]. An experiment will generally consist of  $n$  trials, resulting in a dataset of observations  $\{(y_1, x_1), \dots, (y_n, x_n)\}$ . We write the  $p$  components of the the  $i$ th predictor vector as  $X_i = (X_{i,1}, \dots, X_{i,p})$  and use upper case to denote random variables and lower case to denote observed values.

A GLM consists of three parts: a link function, a linear predictor, and a distributional model. The link function is  $g(\mu_{Y|X=x_i})$ , where  $\mu_{Y|X=x_i}$  is the conditional mean of the response given the  $i$ th set of observed values for the predictor vector. The function  $g$  may be nonlinear, and there are specific canonical choices to which we refer later. This function links the conditional mean to the predictors according to

$$g(\mu_{Y|X=x_i}) = x_i\beta = \beta_0 + \beta_1x_{i,1} + \dots + \beta_px_{i,p} \quad (1)$$

if there are  $p$  predictor variables in  $X$ . The vector of parameters  $\beta$  plays a role analogous to ordinary linear regression parameters: each  $\beta_j$  describes the magnitude and direction of relationship between the  $g(\mu_{Y|X=x_i})$  and the  $j$ th predictor variable.

The conditional distribution of  $Y$  given  $X$  is assumed to be  $Y | X = x_i \sim f(y; \mu_{Y|X=x_i})$ , independently for each  $i$ . The mean of the  $i$ th  $f$  is  $\mu_{Y|X=x_i}$ .

The simplest GLM takes  $g(z) = z$  and  $f(y; \mu_{Y|X=x_i}) = N(y; \mu_{Y|X=x_i}, \sigma^2)$ , where  $N(a, b^2)$  is the normal density with mean  $a$  and variance  $b^2$ . In this case the GLM reduces

to the ordinary multiple linear regression model (LM) for regressing  $Y$  on  $X$ .

Suppose  $Y$  is a binary random variable. The well-known logistic regression model [10] is established by using  $f(y; \mu_{Y|X=x_i}) = \text{Bern}(y; \mu_{Y|X=x_i})$ , where  $\text{Bern}(y; \pi) = \pi^y(1 - \pi)^{1-y}$  (with  $0 \leq \pi \leq 1$ ) is the Bernoulli distribution<sup>1</sup>. This assumption is paired with the canonical link  $g(z) = \text{logit}(z) = \log(z/(1 - z))$ .

A useful supplement to the ordinary LM is the notion of random effects, which provides the ‘mixed linear model’ [15]. Random effects can also be inserted in a GLM, yielding a ‘generalized linear mixed model’ (GLMM) [2, 28]. We define such models here, and describe their benefits below.

Suppose that the predictor variables are of two types. The first type, associated with fixed effects, are variables that we manipulate (e.g. configuration parameters) and variables whose influence on the response is of primary interest (e.g. subject race or gender). The second type of variable includes those that are known or believed to affect the response, but whose effect we are more interested in statistically controlling than estimating. These variables are fit using random effects. For example, if the experiment is replicated over several random training sets, we wish to control for the effect of training set, but it is sensible to treat these effects as mean-zero random influences whose actual values we don’t care to know. In other words, we are not interested in knowing if one specific random training set is or is not easier than another, but we do want to compensate for that fact that some random training sets will be easier or harder than others.

As another example, consider repeated measurements of  $Y$  on each subject (under different conditions). Since some subjects are clearly easier to recognize than others, there is a predictor variable (and corresponding  $\beta_j$  parameters) associated with subject identity that affects  $Y$ . Again, however, we would like to treat these subject effects as mean-zero random quantities. We wish to identify and separate the response variability associated with subject identity so that we don’t confuse it with the effects of the other variables such as subject attributes (race, gender) or configuration parameters.

Let  $X = (F, R)$  be decomposed into the variables associated with fixed and random effects. For simplicity, suppose in what follows that  $R$  contains the predictors resulting from a single random effect variable, such as subject identity<sup>2</sup>. There is a corresponding decomposition of  $\beta = (\beta_F, \beta_R)$ .

<sup>1</sup>There is another simple way to write this model that employs the Binomial distribution; see the references.

<sup>2</sup>There are straightforward extensions for multiple random effect predictors, but it requires greater notational complexity than space permits in this article.

We may then write the linear predictor for a GLMM as

$$g(\mu_{Y|X=x_i}) = x_i\beta = f_i\beta_F + r_i\beta_R \quad (2)$$

where  $x_i = (f_i, r_i)$ . Note that this is the same form as for the GLM. The distribution assumption is expressed in the same way as for the GLM, with an important addition:

$$\beta_{R,j} \sim N(0, \sigma_R^2) \quad (3)$$

for  $j$  indexing over (only) the random effects. Note that (3) allows *random parameters*: mean-zero effects attributable to the corresponding variables. Such models can be fit routinely using restricted pseudo-likelihood [22] or the MIVQUEO method [23, 12, 29].

It is worth stating one benefit of the LM, GLM, and GLMM at this point: all these methods provide a probabilistic framework under which defensible statistical inference can be made. In other words, we can generate standard errors, confidence intervals, and hypothesis tests, rather than relying on descriptive graphs and summaries.

### 3. LM & GLM Analysis of Human Subject Covariates

A total of 2,974 images from the FERET database [20] were scored on 11 factors related to subject. A single person scored all the images. The 11 factors were age, race, gender, skin complexion, glasses, facial hair, makeup, bangs, facial expression, mouth position, and eyelid position. Details of the scoring approach and the rest of the analysis are given by [7]. From these images, 1,072 pairs of images of separate subjects were used for analysis (i.e. 2,144 images altogether). The subjects chosen for this study were required to have two images taken on the same day and not to have removed or added glasses between the two photos. All the subject covariate factors were coded in discrete categories (e.g. old vs. young).

The response variable is the subspace distance between images used by a PCA-based nearest neighbor classifier [13]. Specifically, for a given subject it is the distance between the pair of images of that subject. The link function is the identity function, i.e.  $g(z) = z$ , a normal distributional assumption is made and we use an ordinary LM (in this case an ANOVA) to analyze the data.

The specific distance metric used is a variant of a measure proposed by Moon and Phillips for PCA [19] and subsequently refined by Wendy Yambor [27]. It is a correlation like measure with Mahalanobis like normalization. In past work [1], this distance<sup>3</sup> has performed better than  $L1$  distance,  $L2$  distance or correlation, and therefore it is the

<sup>3</sup>Previous papers have called this measure ‘Mahalanobis Angle’. However, this name is somewhat misleading and is being dropped.

appropriate choice if studying the best a PCA classifier can do.

The FERET data are too sparse with respect to human covariates to model interactions, so we fit a purely additive model. The model yielded  $R^2 = 0.39$ , indicating that about 39% of the total variation in distance can be explained by the subject covariates. When compared to baseline runs of the PCA algorithm, in which about 75% of subjects can be correctly recognized at rank 1, this  $R^2$  is notable.

The results are summarized in Figure 3. Base-case settings are indicated down the center of the diagram, with the degree and direction of effects noted. Effects are expressed as percent change from base-case, and rescaled in terms of similarity (1 minus distance) so that positive effects correspond to easier recognition. The threshold of a two-sided 95% confidence interval is shown as a thin vertical line. Solid bars indicate statistically significant changes in distance measure.

To illustrate, the pairwise distance between images of subjects always wearing glasses is reduced by nearly 33.5% relative to the base-case of subjects not wearing glasses. In this case, the thin vertical line indicating statistical significance appears at 6%; thus the effect is highly significant. A reduction in distance suggests the subjects are more easily recognized, hence the bar for ‘Glasses Always On’ is shown on the right side of the chart. Conversely, consider the subjects whose eyes are open in one image and closed in another. In the base case, subjects have eyes open in both images. The effect for ‘Eyes Open/Closed’ is the top black bar on the side of Figure 3 corresponding to more difficult recognition, and it indicates a 11% increase in relative distance between pairs of subjects. Thus, not surprisingly, our study suggests that subjects are significantly harder to recognize if they close their eyes in one image but not the other.

A more startling outcome is the race effect. In our set of 1,072 FERET subjects, 720 are White, 143 are Asian, 121 are African-American and 88 are Other. Relative to the majority of the subjects (which are White), Asians, African-Americans and Others are all significantly easier to recognize. This is not what we expected going into this experiment. To the contrary, our expectation was that a PCA space trained primarily on white subjects would favor those subjects. Frul et al. [17] have observed a similar result for a smaller subset of the FERET data and looking only at the distinction between White and Asian.

### 3.1. Questioning the Response Variable: Using a GLM

An immediate question arising from our choice of response variable is whether the distance values relate to the primary response of interest: rank-1 recognition. To further study this question, rank-distance between pairs of images is computed. Specifically, for the  $i$ th subject, one image of that

subject is selected to serve as a probe and then the remaining 2,143 images are sorted by increasing distance. Rank-distance  $R_i$  is the position in this sorted list of the other image of the subject<sup>4</sup>.

Visual inspection of a log-log plot suggests a strong relationship between the distance  $Y_i$  and rank distance  $R_i$ . This is reassuring because it suggests that inferences about subject covariates based on distance are likely to hold for recognition rank, too. However, it is important to go beyond a simple visual inspection and to design an experiment that will explicitly test the extent to which  $Y_i$  predicts whether subject  $i$  will or will not be correctly recognized at rank 1. This is done using a GLM to carry out a logistic regression analyses. For this GLM, the predictor variable is the distance  $Y_i$  between the pair of images of subject  $i$ . The response variable is  $Z_i$ , where  $Z_i = 1$  if  $R_i = 0$  and  $Z_i = 0$  otherwise. The GLM model itself can be summarized as  $Z_i|Y_i \sim \text{Bernoulli}(p_i)$  where  $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 Y_i$ .

Results from the model show that the probability of a rank 1 match decreases sharply with increasing distance. In fact, the estimate of  $\beta_1$  is  $-10585.3$  with standard error 809.5, and the negative relationship between these variables is strongly significant. This supplemental analysis comparing rank-1 recognition rate to distance provides confidence that our conclusions are generalizable to recognition rank and rank-1 recognition rate.

### 3.2. Under-represented Groups: Balanced Experiments

Many of the results seen in Figure 3 could be explained by arguing that groups of subjects under-represented in training appear closer in PCA subspace because PCA is proportionally under-representing the region of the space in which these groups lie. Consequently, they appear more tightly clustered than do the majority groupings. If this hypothesis is true, it would be of considerable concern, since, for example, we are drawing the conclusion that Asians are easier to recognize based upon the decreased average distance between pairs of images of Asians relative to whites. This hypothesis was directly addressed by a series of balancing experiments in [7], where it was shown not to be supported by the available data.

### 3.3. Risks of Simpler Methods

What has this analysis gained us, compared to simpler methods? Consider the lack of a significant gender effect shown in Figure 3. Many researchers engaged in face recognition work have at one time or another been part of informal discussions of whether men or woman are more easily recognized, and it is intriguing how often researchers

<sup>4</sup>We subtracted 1 so that the ideal outcome corresponds to a rank-distance of 0.

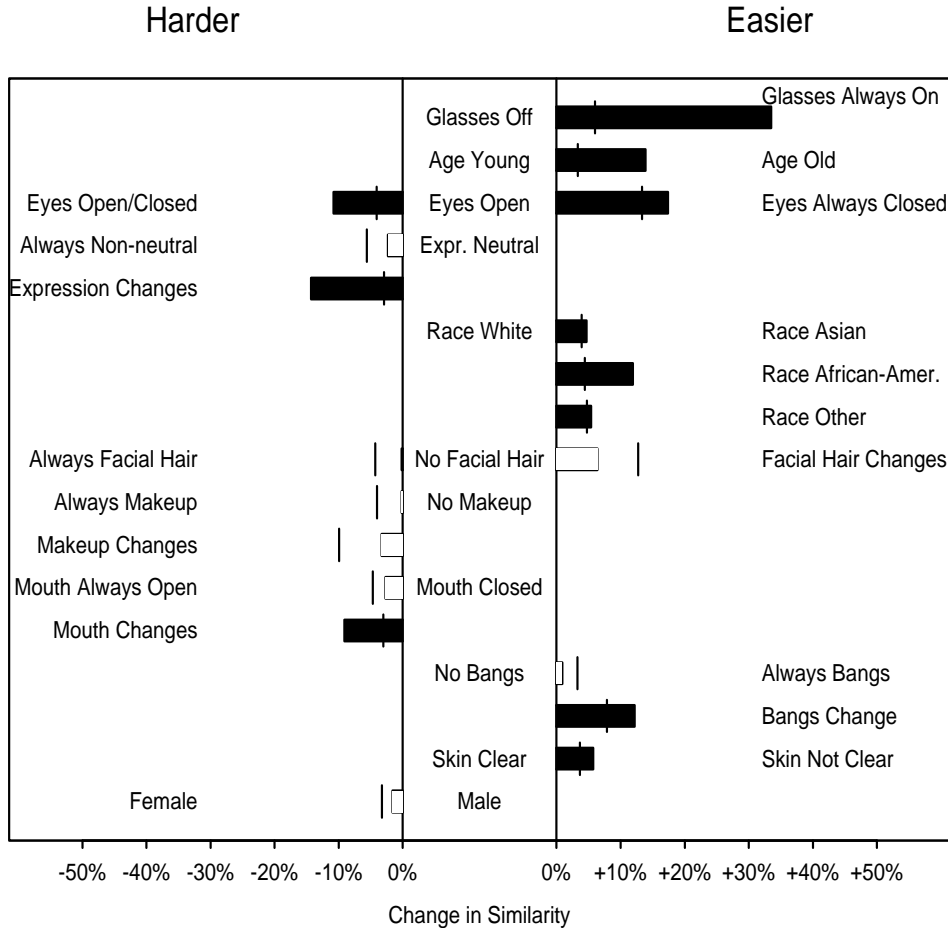


Figure 1: Summary of the results of [7]. Interpretation is discussed in the text.

have an opinion on this question. Nevertheless, there is little prior formal evaluation of gender. One important exception is the work of Gross et al. [21]. They report recognition rates of 87.6 for males and 93.7 for females. However, the results are for the commercially developed FaceIt system [4] and use a different set of face images. Also, the analytical technique in [21] is to compute recognition rates over whole galleries<sup>5</sup> and compare these: a process notably different than ours.

The Gross et al. [21] analysis was a simple one-way, marginal comparison, without a measure of uncertainty. They identified “a surprising trend: better recognition rates are consistently achieved for female subjects”. Setting aside the question of statistical significance, consider the model. Effectively, this conclusion is drawn from a comparison of two mean recognition rates—roughly analogous to a LM

<sup>5</sup>The terms ‘gallery’ and ‘probe’ derive from the FERET protocol [20]. Briefly, a probe image is a new example that is to be matched against a known image gallery, typically by sorting gallery images by increasing distance from the probe image.

like (1) with a single predictor: gender. No consideration is given to possible confounding factors. The result is entirely dependent on how well or poorly the sexes are balanced with respect to other important covariates (e.g. race, age, and glasses) in the analyzed dataset.

Compare this to the appropriate conclusion from our model. In a multiple regression model like ours, the effect for any predictor is interpreted as its impact on  $Y$  holding all other predictor variables fixed. In other words, the  $\beta_j$  estimate is not confounded by the potential effect of any other variable in the model. Thus, our conclusion is that, after adjusting for important effects due to glasses, age, eyelid position, race, facial expression, bangs, mouth position, and skin complexion, there is no significant gender effect ( $t = 0.98, p = 0.33$ ).

This conclusion is stronger than Gross’ conclusion because it controls for potential confounding variables. This lesson is repeated in every text on linear modeling [16]. It is far too much to hope that a ‘sample of convenience’ will be balanced with respect to every confounding variable. It

is far safer, therefore, to explicitly account for covariates in the modeling.

To illustrate the danger, we fit our data with a simple one-way ANOVA LM, analogous to the comparison of [21]. The results show a ‘statistically significant’ gender effect: females are about 13% more difficult ( $t = 6.2$ ,  $p < 0.0001$ ). But we already know this result to be wrong, having completed the more reliable analysis presented above. This example shows how easy it can be to be fooled when the experimental design and analysis are over-simplified.

## 4. GLMM Analysis of a PCA+LDA Algorithm’s Configuration

We summarize here a few results from a larger GLMM analysis [8] of the configuration of a standard PCA+LDA face recognition algorithm [30]. This study was carefully designed using a subset of images from the FERET database [20]. Algorithms were trained on a set of 192 images (3 images each of 64 FERET subjects), plus up to 192 supplemental images generated using various supplementation algorithms described below. A single experiment consisted of configuring the PCA+LDA recognition algorithm and attempting to recognize 256 probes at rank 1. Thus, some probes were included in the training and some were not. The response variable was binary, indicating whether rank-1 recognition was successful.

This experiment was replicated for each of 30 randomly selected galleries. Each experiment with each gallery was further replicated for 8 different sets of training subjects. This entire replicated process was repeated for 32 different algorithm configurations. The same 256 probes and the same 30 gallery sets were used throughout.

The recognition algorithm configurations were generated through full factorial combination of four configuration parameters: image size, distance metric, PCA dimensionality cutoff, and training supplementation. Supplementation of training sets consisted of inserting certain additional images to the baseline training set. The supplemental images were generated by: shifting each baseline image one pixel left, reflecting each baseline image about its centerline, or adding white noise to each baseline image. No supplementation was also an option. Further details of this experiment are given by [8].

This experiment was carefully designed to provide useful information about configuration factors while controlling for possible confounding covariates. In particular, the repeated use of the same probes allows statistical isolation of subject factors from configuration factors. Using different probes for different configurations would have clearly introduced potential biases if the types of subjects used in each case differed in some known or unknown way. The repeated measures design requires that we account for the

resulting covariance structure of the response data, which is done by using a GLMM with subject as a random effect variable. This helps us avoid distraction due to between-subject variation while introducing a correlation between different response observations from the same subject. In particular, in the GLMM model, the correlation between  $Y_i$  and  $Y_j$  is nonzero when these are responses to probes of the same person, and 0 for different people. The degree of correlation depends on the relative magnitude of  $\sigma_R^2$  and the Bernoulli variation.

The results indicate that variation attributable to subject accounted for roughly 70% of variation in rank-1 recognition rates. In contrast, the random training sets accounted for virtually no variability in rank-1 recognition rates. On the logit scale, within-person correlation was about 0.7. These results indicate that subject identity has a dominant influence on recognition, and that one should be cautious before relying on the results of any simpler analysis that fails to account for the correlation structure of the responses.

The best configuration used the soft  $\mathcal{L}_2$  metric [31], 114 PCA dimensions, and training supplemented with shifted images. It had estimated rank-1 recognition rate of 0.87 and 0.75 for subjects in and not in the training set, respectively. In comparison, baseline configuration had rates of 0.75 and 0.74, respectively. In [8] we describe a number of interesting effects that image size, distance metric, PCA dimensionality, and training supplementation have on rank-1 recognition rate. For example, supplementing training with shifted images can be helpful, but supplementing with reflected images actually degrades performance.

The benefit of using a GLMM is that it enables the replicated experiment while improving resolution. Roughly speaking, statistical inference in linear models (including ANOVA, LM, GLM, and GLMM) is driven by the signal to noise ratio for various effects. To detect, say, that old people are easier to recognize at rank 1 than younger people, the effect of age on rank-1 recognition must be of a substantially greater magnitude than the typical variation in rank-1 recognition. Models that incorporate random effects for individuals decompose the noise variation into two parts: noise attributable to between-subject differences (i.e. some people are easier to recognize than others) and unexplained noise. When using such models, the denominator in the signal to noise ratio can be reduced to only the unexplained noise portion. Therefore, since using the random effects allows us to reduce the denominator of the signal to noise ratio from what it would otherwise be, the ratio increases in magnitude, thereby allowing greater statistical power to detect effects that would otherwise be undetectable.

A further benefit of all the modeling approaches discussed in this paper is that they allow estimation of interactions between variables. For example, in this experi-

ment, a key result involved interactions between each predictor mentioned thus far and the additional binary predictor that recorded whether or not each probe was in the training set. Our result was that every single configuration benefit found in our experiment vanished for subjects not included in training. Thus, for example, the soft  $L_2$  metric mentioned above was shown only to improve rank-1 recognition performance when probing subjects that were included in the training. For subjects not seen in training, soft  $L_2$  offers no improvement.

As another example, we found an important and statistically significant three-way interaction. Being in the training set was helpful when PCA dimensionality was fixed at 60% of the number of basis vectors; but when PCA dimensionality was fixed at 90% energy [11, 27], being in the training set was helpful when using the soft  $L_2$  metric but actually detrimental when using the standard  $L_2$  metric. Complex interactions of this type are rarely if ever found when relying on simpler methods.

## 5. Summary and Conclusions

In few disciplines but computer science do researchers have such unfettered opportunity to conduct well-designed experiments on topics of their choosing, with virtually no cost (aside from computation time). This presents human ID researchers with an opportunity to conduct powerful, informative analyses that effectively evaluate and compare algorithms. We have described a variety of models, all well-known (and all but GLMM frequently used) by mainstream applied statisticians. We hope that the experiments and results described above illustrate the importance of a cohesive experimental design and analysis strategy using models that allow one to control for confounding variables, to estimate all important effects including interactions, and to isolate extraneous sources of variation.

## References

- [1] J. Ross Beveridge, Kai She, Bruce Draper, and Geof H. Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535 – 542, December 2001.
- [2] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, 8:9–25, 1993.
- [3] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *PIEEE*, 83(5):705–740, May 1995.
- [4] Identix Corporation. Faceit system homepage. [www.identix.com](http://www.identix.com), 2002.
- [5] M. Teixeira D. Bolme, R. Beveridge and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *Proceedings of the Third International Conference on Vision Systems*, page (to appear), Graz, Austria, April 2003.
- [6] FERET Database. <http://www.itl.nist.gov/iad/humanid/feret/>. NIST, 2001.
- [7] Bruce A. Draper Geof Givens, J. Ross Beveridge and David Bolme. A statistical assessment of subject factors in the pca recognition of human faces. Technical report, Computer Science, Colorado State University, 2003.
- [8] Bruce A. Draper Geof H. Givens, J. Ross Beveridge and David Bolme. Using a generalized linear mixed model to study the configuration space of a pca+lda human face recognition algorithm. Technical report, Computer Science, Colorado State University, 2003.
- [9] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.
- [10] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, NY, 2000.
- [11] M. Kirby. *Dimensionality Reduction and Pattern Analysis: An Empirical Approach*. Wiley, 2000.
- [12] L. R. LaMotte. Quadratic estimation of variance components. *Biometrics*, 29:311–330, 1973.
- [13] M. A. Turk and A. P. Pentland. Face Recognition Using Eigenfaces. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586 – 591, June 1991.
- [14] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983. i-xiii+216pp.
- [15] "C. E. McCulloch and S. R. Searle". *Generalized, Linear, and Mixed Models*. Wiley, New York, 2001. i-xxi+325pp.
- [16] J. Neter, W. Wasserman, and M. H. Kutner. *Applied Linear Statistical Models*. Irwin, Boston, 1990. i-xvi+1181pp.

- [17] P. Jonathon Phillips, Nicholas Furl, and Alice J. O’Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26:797–815, 2002.
- [18] Olivetti Face Data. <http://www.uk.research.att.com/facedatabase.html>. AT&T Laboratories Cambridge, Cambridge University Engineering Department, 2003.
- [19] J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation. In H. Wechsler, J. Phillips, V. Bruse, F. Soulie, and T. Hauhg, editors, *Face Recognition: From Theory to Application*. Springer-Verlag, Berlin, 1998.
- [20] P.J. Phillips, H.J. Moon, S.A. Rizvi, and P.J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *T-PAMI*, 22(10):1090–1104, October 2000.
- [21] Jeffrey F. Cohn, Ralph Gross, Jianbo Shi. Quo vadis face recognition?: The current state of the art in face recognition. Technical Report TR-01-17, Carnegie Mellon University, June 2001.
- [22] Ramon C. Littell, George A. Milliken, Walter W. Stroup, Russell Wolfinger. *SAS System for Mixed Models*. SAS Publishing, Cary NC, 1996.
- [23] C. R. Rao. Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.*, 67:112–115, 1972.
- [24] Ross J. Micheals and Terry Boult. Efficient evaluation of classification and recognition systems. In *IEEE Computer Vision and Pattern Recognition 2001*, pages I:50–57, December 2001.
- [25] D.A. Socolinsky and A. Selinger. A comparative analysis of face recognition performance with visible and thermal infrared imagery. In *ICPR02*, pages IV: 217–222, 2002.
- [26] D. Valentin, H. Abdi, A.J. O’Toole, and G.W. Cottrell. Connectionist models of face processing: A survey. *Pattern Recognition*, 27(9):1209–1230, September 1994.
- [27] Bruce A. Draper, Wendy S. Yambor, and J. Ross Beveridge. Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. In *Second Workshop on Empirical Evaluation in Computer Vision*, Dublin, Ireland, July 2000.
- [28] Russ Wolfinger and M. O’Connell. Generalized linear models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.
- [29] Russ Wolfinger, Randy Tobias, and John Sall. Computing gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal of Scientific Computing*, 15:1294–1310, 1994.
- [30] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *In Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, editors, Face Recognition: From Theory to Applications*, pages 73–85, 1998.
- [31] W. Zhao, R. Chellappa, and P.J. Phillips. Subspace linear discriminant analysis for face recognition. In *UMD*, 1999.