Int. J. Advance Soft Compu. Appl, Vol. 7, No. 2, July 2015
ISSN 2074-2827

Effectiveness of Feature Weight Using BPSO In Text-Dependent Writer Identification

Khaled Mohammed bin Abdl[1,2], and Siti Zaiton Mohd Hashim[1]

[1]Faculty of Computing
Universiti Teknologi Malaysia, Malaysia
email: sitizaiton@utm.my

[2]Faculty of Applied Science
Hadhramout University for Science & Technology, Yemen
email: bin_abdel@hotmail.com

Abstract

Writer identification is an authorship authentication process based differences and similarities in handwriting. The main issue in writer identification is how to get the features that invariant to the writer. This study proposes Binary Particle Swarm Optimization (BPSO) based off-line text-dependent to investigate the effectiveness of feature weight in writer identification. BPSO has ability to perform such role since BPSO works on particle level and swarm level. The weight obtained by BPSO it's an average of feature selected times over 10 runs per writer. Then each feature multiplied by its corresponding weight so the features represented by their importance not their values. Off-line text-dependent words from IAM database are used. Moment and statistical features are extracted to represent the handwritten words. Experimental results show an improvement in writer identification performance based feature weight.

Keywords: *Binary Particle Swarm Optimization, Feature Weight, Offline Text-Dependent, Writer Identification.*

1 Introduction

Writer Identification (WI) is an authorship authentication process based the differences and similarities in handwriting. This process performs a one-to-many search in a database of known authorship to determine the author of a questioned handwritten document, see Fig. 1 for the typical steps based WI. WI rests on three facts: 1) writing is a skillful process that developed gradually, 2) visually, no two

people write exactly alike; and 3) no one person writes exactly the same way twice. In addition to, the intra-class (within-writer) variation is less than the inter-class (between-writers) variation. [1] established with a 98% confidence that the handwriting can be used as a biometric of a parson. WI methods fall into two categories text-dependent and text-independent. Text-dependent methods are constrained on comparisons between similar text, i.e. characters or words. On the other hand, text-independent methods use statistical features extracted from the entire image of a text block. In contrast, text-independent do not make any assumptions about the text while, text-dependent methods offer higher discriminative power using lesser amount of data. [2] concluded 1) text-dependent method achieves better results, but computation cost; 2) text-independent is quite simple strategy, but it shows bigger error rate; 3) if the application has few writers, the text-dependent should be considered and vice-versa. [3] suggested that, WI should be able to combine the advantages of text-independent (robust to forgery) and text-dependent (high accuracy with less data). WI can be off-line or on-line based on handwritten data, see Table 1 for the differences between on-line and off-line handwritten features. Off-line refers to scanned images of handwritten text i.e. real time information is lost. Dynamic features such as velocity, acceleration, pen-pressure, pen up-down movements, writing direction and strokes order are called on-line handwriting. On-line handwriting represents more individuality features [4]. However, it is often neglected for the limitation of the existing input device [5]. In addition to, off-line is commonly used since on-line not available in forensic applications and also it has high within class variation.

Table1: Example of on-line and off-line handwriting features

On-line features	**Off-line features.**
Pressure Acceleration	Horizontal Midpoint
Cartesian Displacement	Width of a word
Horizontal Displacement	Vertical Midpoint
Vertical Displacement	Height of a Word
Displacement in Pressure	Height of a Capital Letter
Number of Pen-Ups	Height of an Upper Zone
Duration of Writing	Height of a Middle Zone
Cartesian Acceleration	Height of a Lower Zone
Horizontal Acceleration	Height of Ascender
Vertical Acceleration	Height of Descender
Pressure Acceleration	Space between Word
Rotation	Slant

Due to the need of WI applications in forensic and civilian domains, numerous researches have been conducted towards WI. Handwriting individuality is proved by [1]. Threats in WI are discussed by [6]. Strength of evidence in WI investigated in [7]. [8] treat the writer identification task as a texture analysis

problem.[9] WI based feature selection. [1] has described a content-based information retrieval system for handwritten documents. Not too far [10] proposed textual based information retrieval model for the WI and Writer Verification (WV). [11] shows that handwritten words carry more individuality than handwritten allographs. In same way [12] concludes longer words provide better performance than shorter words. Capital letters that consist of several strokes bear more individual information than simple characters like “i” or “c” [13]. Not too far, writer invariant (set of similar patterns) provides higher discriminating power than any single character “d”, “y”, and “f”. [14] described a WI based recognizer. The study has concluded that: 1) there is a strong correlation between the text recognition (transforming handwriting to a machine print) and the WI, 2) applying normalization increases the recognition rate and decreases the identification accuracy because it removes writer-specific information.

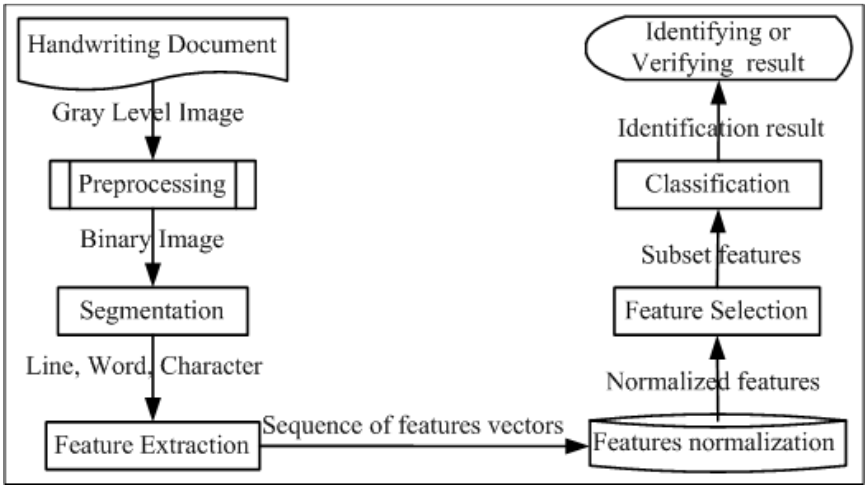

Fig. 1. Typical steps based WI

The main issue in WI is how to get the features that invariant to the writer. It has been shown in literature that there are a significant number of feature extraction techniques developed for WI. However, the importance of a specific feature in WI has not been entirely investigated. This paper intended to investigate the effectiveness of feature weight in WI using BPSO. BPSO has ability to perform such role and learn the features weights because it works on local level (particle level) and global level (swarm level) [15]. In addition to, BPSO is still not used for WI. However, the continuous-values version of Particle Swarm Optimization (PSO) has shown high performance in some related fields like pattern classification [16, 17], signature verification [18], and handwriting digit recognition [19]. This paper proposed off-line text-dependent WI-based on BPSO to examine feature weight in WI. Section 2 feature ranking based WI. Section 3 BPSO based WI. Section 4 discuses experimental and results. Section 5 draws the conclusion.

2 Feature Ranking Based Writer Identification

Among all the phases of WI, feature ranking play very important role. Undoubtedly, features quality and quantity affects several aspects such as complexity, accuracy, and identification time [11] revealed that feature selection can significantly improve the WI rate using Sequential Forward Search (SFS), Genetic Algorithm (GA), Principal Component Analysis (PCA) and Multiple Discriminate Analysis (MDA). [20] WI accuracies reach acceptable levels using optimal subset features produced by Genetic Algorithm (GA). The main issue in WI is how to get the features that reflect the varieties of handwriting. It has been shown in literature that there are a significant number of feature extraction techniques developed and employed for WI; however, the importance of a specific feature in WI has not been entirely investigated. WI is presented as a feature ranking model by this study. Let D is a handwritten document and that D contains texts each text represented by features. As D has $T_i(i = 1, 2, \dots, n)$ text. Each T_i is presented by m features $(1 \leq m < \infty)$ denoted by $f_j(j = 1,2, \dots, m)$ and they are subject to the following constraints:

$$D = T_1 \cup T_2 \cup \dots T_n \quad (1)$$
$$f^i \neq \emptyset, \ (\forall_i = 1, \ 2, \ \dots, \ n) \quad (2)$$
$$f_j \cap f_k = \emptyset, \ \left(\forall_{j,k} = 1, \ 2, . , \ m\right), \ \forall \ j \neq k \quad (3)$$

An evaluating function (θ) assigns to each feature f_j of T_i a score value v_j^i , the value v_j^i is a measurement of confidence that f_j is an individuality representative feature of a writer $W_g, g = 2, 3, \dots, r, (2 \leq r < \infty)$

$$\left(W_g^i, v_j^i\right) = \theta(f_j^i), \qquad v_j^i \in [0,1] \qquad (4)$$

3 BPSO based Writer Identification

Handwriting features are the cornerstone in constructing of any WI system, however, the identification accuracy is sensitive for those features in terms of how the writers are scored. This presents us with a feature weight problem in the WI. Therefore, we used BPSO [15] to investigate the effect of the feature weight and the feature selection in the WI problem. The BPSO has ability to perform such role and learn the feature weights, since BPSO works on local level (particle level) and global level (swarm level), where many solutions are suggested for the problem and the best solution among them is selected. The continuous-values version of particle swarm optimization was applied successfully in the feature weight problem; [21] used PSO to select subset of features for classification and training of neural network. [16] used PSO for feature selection in the classification problem where support vector machines with one-versus-rest method were used as fitness function. The good performance of PSO in literature

promises that BPSO can do well for WI as well, see Fig. 2 a proposed framework of BPSO-based WI. In addition to, BPSO is still not used for WI. However, the continuous-values version of Particle Swarm Optimization (PSO) has shown high performance in some related fields like pattern classification [16, 17], signature verification [18], and handwriting digit recognition [19].

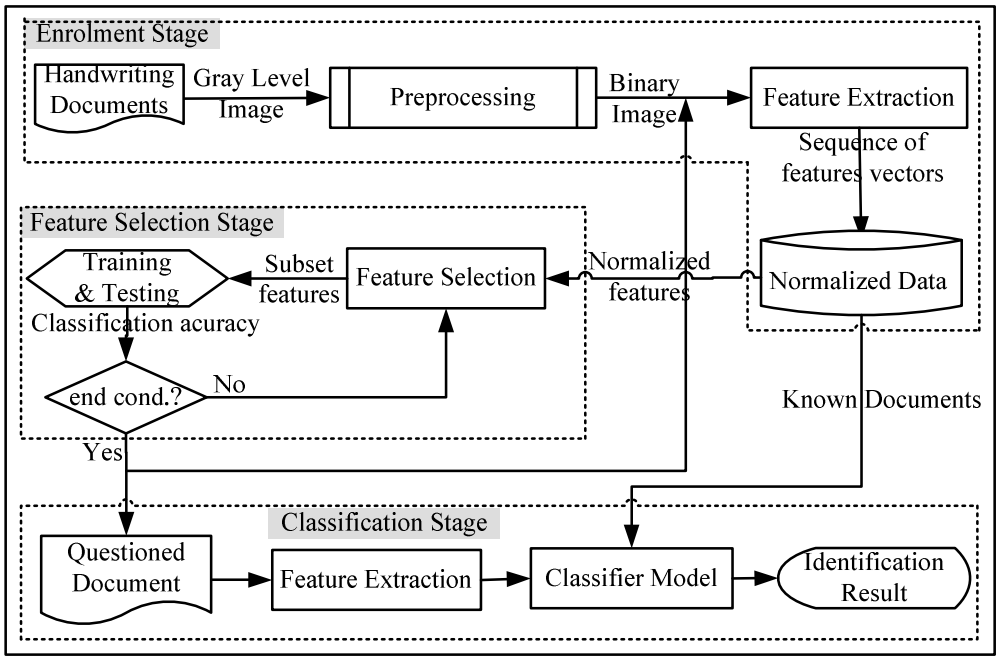

Fig. 2. Typical steps based WI

3.1 BPSO Algorithm

Particle Swarm Optimization (PSO) is an adaptive algorithm simulates the social behavior such as bees, birds or a school of fish. Originally PSO is developed as an optimization technique for continuous-values search spaces. However, lots of practical issues are formulated as discrete optimization problems. [15] developed a discrete version of PSO for binary problems. They proposed a model wherein the particle position is represented as bit string rather than real numbers. In the search space each single solution is a "bird" and it called "particle". Each particle uses its own and companions' flying experience to move in the search space and retains the best position. Each particle is updated at any iteration based on two values. First one is the local best *(pbest)* which is the best solution it has achieved so far. Another "best" value is the global best, obtained so far by any particle and called *(gbest)*. Optimization process is then carried out a fixed number of iterations. Particle velocity and position are updated at any iteration using Eq. (5) and Eq. (6) respectively. Finally, after several runs the optimal or near optimal solution is found. The basic pseudo-code for the BPSO algorithm can be described as follows:

Start

For each particle
Initialize particle with random numbers
End
While maximum number of iterations is not met
For each particle
Calculate fitness value
If the fitness value is better than the best fitness value (*pbest*) in history
Set current value as the new *pbest*
End
End
Choose the particle with the best fitness value of all the particles in the history as the *gbest*
For each particle
Calculate particle velocity according to equation (5)
Update particle position according to equation (7)
End
End

Finish

$$V_i(t+1) \leftarrow w * V_i(t) + c_1 r_1 (P_b(t) - X_i(t)) + c_2 r_2 (P_g(t) - X_i(t)) \qquad (5)$$

Where $V_i(t)$, $X_i(t)$, $c_{1,2}$, $P_b(t)$, $P_g(t)$, w, and $r_{1,2}$ respectively are particles velocity, particles position, acceleration parameters, *pbest*, *gbest*, inertia weight, random numbers.

$$Xi(t+1) \leftarrow Xi(t) + Vi(t+1) \qquad (6)$$

Where $x_i(t+1)$, $x_i(t)$, $V_i(t+1)$ respectively are particle *i* new position, particle *i* current position, particle *i* new velocity.

$$X_{ij}(t+1) = \begin{cases} 0 \; if \; p_{ij}(t) \geq \dfrac{1}{1+e^{-v_{ij}}}(t) \\ 1 \; otherwise \end{cases} \qquad (7)$$

In continuous PSO, position update is done directly by adding the velocity to the previous position while in BPSO, the velocity is used only in the sigmoid function to calculate the probability of the bit value to be changed to 0 or 1, where the value retrieved from the sigmoid function is compared with random generated value in the range between zero and one. Fig. 3 and Eq.(7) show the sigmoid function. Our literature survey has shown that BPSO has not investigated in WI.

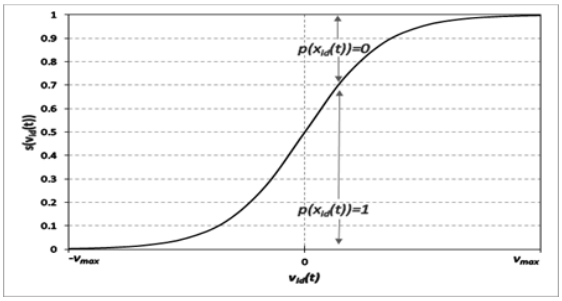

Fig. 3. The sigmoid function

4 Experimental and Results

Handwriting features are writer's characteristics to individuality [22]. These features appear in the literature under different names such as moments, statistical, macro, micro, directional, and texture-level features. Geometrical moments [22] proved to be most useful features corresponding to aspects of the shape of handwriting. [23] proposed WI using statistical features such as writing width, slant, height, writing zones, connected components, enclosed regions, lower and upper contour.

This study used geometrical (moment and statistical) feature. The geometric moment of $(p+q)^{th}$ order are extracted to represent inertial ratio, aspect ratio, spreadness horizontal skewness, vertical skewness, horizontal extensio, and vertical extension of handwritten images. The statistical features are used to measure length, height and the area of handwriting zones. Then find the suitability between the writing zones e.g. aspect ratio of word height to middle zone height; Fig. 4 depict the writing zones.

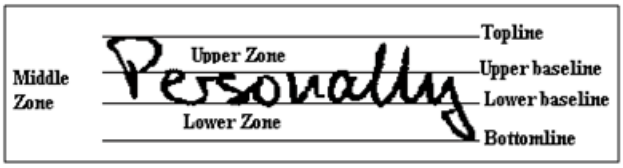

Fig. 4. English writing zones

For the screening experiments, IAM-G06 data set is chosen from the off-line IAM data (English handwriting) (http://www.iam.unibe.ch/fki/databases/iam-handwriting-database, 2010), see Table 2.

Table 2: Genuine and Imposter per Training and Testing Dataset

		Training		Testing		
Data set	**Features**	**Genuine**	**Impostor**	**Genuine**	**Impostor**	**Total**
IAM-G07	8	515	575	559	499	2148
IAM-H07	21	567	626	459	400	2052
IAM-A01	29	451	1026	377	630	2484
IAM-G06	144	349	5600	230	2506	8685

4.1 BPSO Setup

BPSO variables set as follows: 5 particles are used, *V max*, *V min*, C_1, and C_2 respectively are 4,- 4, 2, and 2. The range of *w* is 0.4 to 0.9. Iterations set to 1000 and runs to 10. BPSO works as: first, subset feature is selected by a particle and evaluated using the Euclidian Distance (ED) Eq. (8). The *pBest* value for the corresponding particles is calculated and the *gBest* among those five particles is chosen. Second, *pBest* and *gBest* are updated by comparing the new values with the prior ones. Third, at each run, the best features selected as vector based on the particle position with the *gBest* value. Finally, the feature weight is average of the created vectors.

$$ED = \sqrt{\sum_{i}^{n} (r_i - q_i)^2} \qquad (8)$$

Where: *n* is number of features, r_i, q_i respectively are referenced and questioned images.

Table 3: Experimental results with and without feature weight

Data set	Without feature weight			With feature weight		
	1 - FMR%	1 - FNMR%	ACC%	1 - FMR%	1 - FNMR%	ACC%
G07	92.63	97.58	95.11	93.84	94.78	95.71
H07	93.26	98.45	95.86	93.98	95.10	96.21
A01	94.19	96.98	95.59	97.27	97.50	97.12
G06	94.84	98.09	96.47	97.35	97.54	97.72

This study uses the K-fold cross-validation method [24]. We separate the data into *n* parts according to the number of writer in each dataset $W_1, W_2, \ldots, W_n$. Then we carried out experiments a total of *n* times. So, *n* times of classification accuracies

are produced in each dataset. Finally, the classification accuracy of WI is averages of these *n* accuracies. Table 4 shows a small improvement in text-dependent WI accuracy using feature weight. A slight improvement has shown by G06 and A01 which represented by large number of features. Group G07 which represented by moment features showed the lowest improvement.

5 Conclusion

This paper investigated the effectiveness of feature weight in WI process. BPSO is used as feature weighting mechanism and ED is used as an evaluation function. At each run, the best features selected as vector based on the particle position with the *gBest* value. Then, the feature weight is average of the created vectors. Finally, each feature multiplied by its corresponding weight so the features represented by their importance not their values. Text-dependent handwritten words are used to validate the proposed approach. Experimental results show slightly improvement in WI accuracy based feature weight. This study can be extended as feature selection based WI. An optimal subset feature will be selected based on top *n* high weight features.

ACKNOWLEDGEMENTS.
This work was supported by Ministry of Education (MOE), Malaysia and Universiti Teknologi Malaysia (UTM) through Research University Grant (GUP), the authors wish to thank the editor of IJASCA, as well as the anonymous reviewers for their helpful comments.

References

[1] Srihari, S. Cha, S. Arora, H.; and Lee. 2002. Individuality of Handwriting, *Journal of Forensic Sciences*, 47(4), July 2002, 1-17

[2] Gupta, S. and Namboodiri A. 2008. Text Dependent Writer Verification using Boosting. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition(ICFHR'08)*, Montreal, Canada

[3] Pavelec, D., Justino, E., Batista, L. and Oliveira, L. 2008. Author Identification using Writer-Dependent and Writer-Independent Strategies. In *proceedings of the 2008 ACM symposium on Applied computing*, March, 414-418.

[4] Yamazaki,Y. Nakashima, A. Tasaka, K. Komatsu, N. 2005. A Study on Vulnerability in On-line Writer Verification System. *ICDAR 2005*, 640-644.

[5] Meng, M. Wu, Z. Fang, P. Ge, Y. Yu, Y. 2004. On-line writer verification using force features of basic strokes, *Lecture notes in computer science*, *2004*, ISSU 3338, 646-653.

[6] Uludag, U. and Jain, A. 2004. Attacks on biometric systems: a case study in fingerprints, *Proc. SPIE-EI 2004, Security, Steganography and Watermarking of Multimedia Contents VI*, 622-633, San Jose, CA, January 18-22.

[7] Srinivasan, H. Kabra, S. Srihari, S. and Huang, C. 2007. On computing strength of evidence for writer verification, Proceedings of the 9th *International Conference on Document Analysis and Recognition*, Sep 2007, 844-848.

[8] Said, H. Tan, T. and Baker, K. 2000. Personal identification based on handwriting. *Pattern Recognition*, 33: 149-160, Jan. 2000

[9] Bar-Yosef, I. Beckman, I. Kedem, K. and Dinstein I. 2007. Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. *International Journal on Document Analysis and Recognition, (IJDAR)*. 89-100

[10] Bensefia, A. Paquet, T. and Heutte, L. 2005. Handwritten Document Analysis for Automatic Writer Recognition, *Electronic Letters on Computer Vision and Image Analysis*, Vol. 5, No. 2, 72-86, Aug 2005.

[11] Zhang, B. and Srihari, S. N. 2003. Analysis of handwriting individuality using word features. *In Proceedings of the Seventh International IEEE Conference on Document Analysis and Recognition (ICDAR 2003)*. 3-6 August. Edinburgh, Scotland, 1142–1146

[12] Tomai, C., Zhang, B., and Srihari, S. N. 2004. Discriminatory power of handwritten words for writer recognition. *In proceedings of the 17th International IEEE Conference on Pattern Recognition (ICPR'04)*. Cambridge, UK. 638–641

[13] Pervouchine , V. and Leedham G. 2007 . Extraction and analysis of forensic document examiner features used for writer identification. *Pattern Recognition Society*, Elsevier, vol. 40. March. 1004–1013

[14] Schlapbach A, Bunke H. 2005. Writer identification using an HMM-based handwriting recognition system: to normalize the input or not? In: *Proceedings of the 12th Conference Of The International Graphonomics Society*, 138–142

[15] Kennedy J & Eberhart RC. 1997. A discrete binary version of the particle swarm algorithm. Systems. Man. and Cybernetics. Computational Cybernetics and Simulation. 1997 *IEEE International Conference on. (IEEE)*. 4104-4108.

[16] Tu, C. Chuang, L. Chang, J. and Yang, C. 2006. Feature selection using PSO-SVM. *International Journal of Computer Science.*

[17] Huang, B.Q. and Kechadi, M.T. 2006. A fast feature selection model for online handwriting symbol recognition. *Proceedings of the 5th International IEEE Conference on Machine Learning and Applications*, Dec. 2006, IEEE Xplore Press, Orlando, FL., pp: 251-257. DOI: 10.1109/ICMLA.2006.6

[18] Das, M.T. and Dulger, L.C. 2007. Off-line signature verification with PSO-NN algorithm. *Proceeding of the 22nd International IEEE Symposium on Computer and Information Sciences*, Nov. 7-9, IEEE Xplore Press, Ankara, pp: 1- 6. DOI: 10.1109/ISCIS.2007.4456842

[19] Sahel, O. and Shamsuddin, S.M. 2008. Handwritten digits using recognition with particle swarm optimization. *Proceeding of the 2nd Asia International Conference on Modeling and Simulation*, May, 13-15, IEEE Xplore Press, Kuala Lumpur, pp: 615-619. DOI: 10.1109/AMS.2008.141

[20] Gazzah, S.and. Ben Amara E. 2006. Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection, Lecture Notes in Computer Science, *International Symposium on Neural Networks*, China, Vol. 3972, 2006, 271-276

[21] Liu, Y., Qin, Z., Xu, Z. and He, X. 2004. Feature Selection with Particle Swarms. In Zhang, J., He, J.-H. and Fu, Y. (Ed.). *Computational and Information Science*, LNCS 3314, 425–430. Heidelberg: Springer-Verlag

[22] Muda, A.K., S.M. Shamsuddin, S. M. Darus, M. 2008. Invariants discretization for individuality representation in handwritten authorship. *Lecture Notes Comput. Sci.*, 5158: 218-228. DOI: 10.1007/978-3-540-85303-9_20

[23] Hertel, C. and Bunke, H. 2003. A set of novel features for writer identification, *Proc. of 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, UK, 679–687

[24] Stone, M., 1974. Cross-Validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society* B, 36, 1974, 111-147