BAYESIAN GAUSSIAN MIXTURE MODEL FOR SPATIAL-SPECTRAL CLASSIFICATION OF HYPERSPECTRAL IMAGES

Koray Kayabol

Gebze Technical University
Electronics Engineering
Turkey

ABSTRACT

We propose a Bayesian Gaussian mixture model for hyperspectral image classification. The model provides a robust estimation framework for small size training samples. Defining prior distributions for the mean vector and the covariance matrix, we are able to regularize the parameter estimation problem. Especially, we can obtain invertible positive definite covariance matrices. The mixture model also takes into account the spatial alignments of the pixels by using non-stationary mixture proportions. Based on the classification results obtained on Indian Pine data set, the proposed method yields better classification performance especially for small size training samples compared to state-of-the-art linear and quadratic classifiers.

Index Terms— Hyperspectral images, classification, Bayesian, Gaussian mixture models, auto logistic regression

1. INTRODUCTION

Hyperspectral images (HSIs) take place in many remote sensing applications including forest vegetation mapping and classification, urban and land usage, determination of the water resources and the crop species. The aim of this paper includes the contextual classification of hyperspectral images. A probabilistic model is proposed to include the spatial information into spectral classification problem.

An intuitive method used for spectral-spatial classification of hyperspectral image is the application of classification and segmentation processes successively as two independent processes. Mostly used classification methods are k-means, support vector machines, linear and quadratic discriminant analysis. After classification, a segmentation method need to be performed to obtain a smooth classification map. Rather than this kind of intuitive approaches, we use a Bayesian probabilistic model with inference algorithm that performs the classification and segmentation tasks together.

We propose to use a Bayesian Gaussian mixture model (GMM) for classification of HSI. GMM is a well-known probabilistic model widely used in data classification applications but it is not preferred in HSI classification due to large sizes of feature vectors and small sample size problem. For a L-dimensional data, there are $L + (L^2 - L)/2$ number of unknowns to be estimated for a single mixture component (L unknowns for mean vector and $(L^2 - L)/2$ unknowns for covariance matrix). Since there are less number of training samples compared to unknowns, the estimation problem becomes under-determined. To overcome the difficulty of under-determined problem, a dimension reduction before GMM classification is proposed in [1]. In [2], regularized linear discriminant analysis (LDA) is used. In [3], the covariance matrices are constrained to be in a particular structure to reduce the number of parameters to be estimated. In this study, we propose to use a Bayesian mixture model that allows to estimate parameters of under-determined problem without dimension reduction. In the proposed Bayesian GMM, we are able to define different covariance matrices for each class rather than using a single and unique covariance matrix as in LDA. In Bayesian framework, an under-detemined estimation problem can be regularized by defining appropriate prior distributions for the parameters. For example, using a few number of samples to estimate the covariance matrix of Gaussian may cause a non-invertible covariance matrix. Defining a prior distribution, we may obtain an invertible covariance matrix.

Other contribution of the paper is inclusion of the the spatial information into HSI classification problem by assuming that the mixture model is non-stationary. Non-stationarity is introduced into model by defining spatially varying mixture proportions. Non-stationary mixture models has been already used in image segmentation and classification applications, e.g. in [4], [5], [6], [7] a Markov Random Fields (MRFs) prior is used for mixture proportions to achieve a non-stationary mixture model. In [8] and [9], a latent Gaussian random field is proposed such that the mixture proportions are connected to the class labels by a Multinomial Logistic (MNL) function. In [10] and [11] an auto-logistic regression model is defined on class labels for synthetic aperture radar image classification. In this study, we use the auto-logistic regression model used in [10] and [11] for contextual classification of HSIs.

This work is supported by Scientific and Technological Research Council of Turkey (TUBITAK) under Project No. 114E535.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

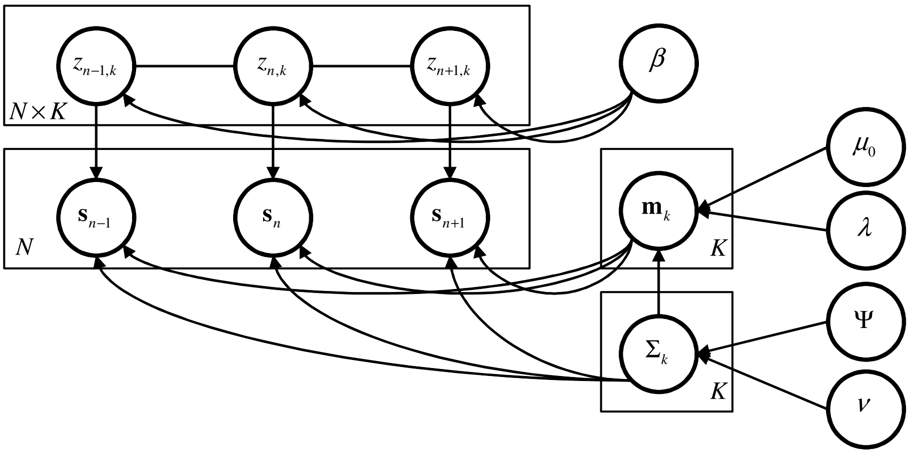


Fig. 1. Graphical representation of Bayesian GMM.

Organization of the paper is as follows. Section 2 and 3 respectively present the proposed Bayesian GMM-based model and related classification algorithm. The experimental results are reported in Section 4. Section 5 summarizes the conclusion and future work.

2. BAYESIAN GAUSSIAN MIXTURE MODEL FOR HYPERSPECTRAL IMAGES

We denote the HSI feature vector by $\mathbf{s}_n \in \mathbf{R}^L$ where $n = 1,\dots,N$ is the lexicographically ordered pixel indices. We assume that a feature vector at a pixel is produced from one of the K different multivariate Gaussian distributions. Each Gaussian distribution represents a class in HSI. Thus, the distribution of the kth class is given below

$$p(\mathbf{s}_n|\mathbf{m}_k,\Sigma_k) = \frac{1}{2\pi|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{s}_n-\mathbf{m}_k)^T\Sigma_k^{-1}(\mathbf{s}_n-\mathbf{m}_k)\right\} \quad (1)$$

where $\mathbf{m}_k$ and Σ_k are the mean vector and the covariance matrix of the kth class, respectively. For the kth class, the parameter set is defined to be $\theta_k = \{\mathbf{m}_k,\Sigma_k\}$. We define a normal inverse-Wishart prior for $\mathbf{m}_k$ and Σ_k, i.e.

$$p(\mathbf{m}_k,\Sigma_k) = \mathcal{N}\left(\mathbf{m}_k\middle|\mathbf{m}_0,\frac{1}{\lambda}\Sigma_k\right)\mathcal{W}^{-1}\left(\Sigma_k|\Psi,\nu\right) \quad (2)$$

Assuming that there are K number of land cover class in HSI, we define a K-dimensional label vector $\mathbf{z}_n \in \{0,1\}^K$ for each pixels. The binary label vector $\mathbf{z}_n$ has the property that $\sum_{k=1}^K z_{n,k} = 1$ which means it indicates only one of the K superpixel by assigning its related element to 1. We can write $\mathbf{z}_n \in \{[1,0,\dots,0],[0,1,\dots,0],\dots,[0,0,\dots,1]\}$. We also assume that $\mathbf{s}_n$'s are conditionally independent given the labels, $\mathbf{z}_n$'s. In this study, we assume that $\mathbf{s}_n$'s are independent but the hidden labels, $\mathbf{z}_n$'s, are spatially dependent. The joint conditional density of $\mathbf{s}_n$ and $\mathbf{z}_n$ is written as $p(\mathbf{s}_n,\mathbf{z}_n|\theta_{1:K},\beta) \propto p(\mathbf{s}_n|\mathbf{z}_n,\theta_{1:K})p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)},\beta)$ where $p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)},\beta)$ is the prior density of the class label, $\mathcal{M}(n)$ is the set of pixels around the nth pixel and β is the smoothing parameter. We define an auto-logistic regression model over hidden labels. According to auto-logistic regression, the conditional probability of a class label can be given as follows [12]:

$$p(z_{n,k}|z_{\mathcal{M}(n),k},\beta) \propto e^{\beta\left(z_{n,k}+z_{n,k}\sum_{m\in\mathcal{M}(n)} z_{m,k}\right)} \quad (3)$$

The joint density of the binary image $z_{n,k}$, $n = 1,\dots,N$, can be constituted by multiplying the individual conditionals. Unlike [12], we use long distance interactions in $\mathcal{M}(n)$. Since we use an homogeneous β in the regression, we can still find a proper joint density for the binary image. This auto-logistic model has been already used in [10] and [11] for amplitude based SAR image classification.

After these definitions, we may write the probability of $\mathbf{s}_n$ as the marginalization of the joint probability density $p(\mathbf{s}_n,\mathbf{z}_n|\theta_{1:K},\mathbf{z}_{\mathcal{M}(n)},\beta) = p(\mathbf{s}_n|\mathbf{z}_n,\theta_{1:K})p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)},\beta)$ w.r.t. hidden label vector $\mathbf{z}_n$ as follows:

$$p(\mathbf{s}_n|\theta_{1:K},\mathbf{z}_{\mathcal{M}(n)},\beta) = \sum_{\mathbf{z}_n}\prod_{k=1}^K[p(\mathbf{s}_n|\theta_k)\omega_{n,k}]^{z_{n,k}} \quad (4)$$

where $\omega_{n,k}$ is the non-stationary mixture proportions and can be obtained from (3) as

$$\omega_{n,k} = p(z_{n,k}=1|z_{\mathcal{M}(n),k},\beta) = \frac{e^{\beta v_{n,k}}}{\sum_{j=1}^K e^{\beta v_{n,j}}} \quad (5)$$

where

$$v_{n,k} = 1 + \sum_{m\in\mathcal{M}(n)} z_{m,k}. \quad (6)$$

The overall Bayesian GMM is graphically shown in Fig. 1.

3. INFERENCE

We use a supervised classification method. The details of the training and classification phases of the method are given in the following two sections.

3.1. Training

We need to perform a posterior inference for the proposed probabilistic model using the labeled training data. Given the labeled training data $\mathbf{z}_{train}$, we may find the maximum-a-posteriori (MAP) estimates of the mean vectors and the covariance matrices of Gaussian distributions. The posterior of the parameters are given by

$$\mathbf{m}_k|\mathbf{s}_{1:N}, \Sigma_k \sim \mathcal{N}\left(\mathbf{m}_k \middle| \frac{N_k\bar{\mathbf{s}}_k + \lambda\mathbf{m}_0}{N_k+\lambda}, \frac{1}{N_k+\lambda}\Sigma_k\right) \quad (7)$$

and

$$\Sigma_k|\mathbf{s}_{1:N} \sim \mathcal{W}^{-1}\left(\Sigma_k \middle| \Psi + N_k\mathbf{S}_k + \frac{N_k\lambda}{N_k+\lambda}(\bar{\mathbf{s}}_k - \mathbf{m}_0)(\bar{\mathbf{s}}_k - \mathbf{m}_0)^T, N_k+\nu\right) \quad (8)$$

where the sample mean vector $\bar{\mathbf{s}}$ and the sample covariance matrix $\mathbf{S}$ are calculated as follows:

$$\bar{\mathbf{s}}_k = \frac{1}{N_k}\sum_{n\in D_k}\mathbf{s}_n \quad (9)$$

$$\mathbf{S} = \frac{1}{N_k}\sum_{n\in D_k}(\mathbf{s}_n - \bar{\mathbf{s}}_k)(\mathbf{s}_n - \bar{\mathbf{s}}_k)^T \quad (10)$$

As seen from (7) and (8), posteriors are well-known distributions, normal and inverse-Wishart respectively. From these posteriors MAP estimates of the parameters are found as follows:

$$\hat{\mathbf{m}}_k = \frac{N_k\bar{\mathbf{s}}_k + \lambda\mathbf{m}_0}{N_k+\lambda} \quad (11)$$

$$\hat{\Sigma}_k = \frac{1}{N_k+\nu+L+1} \cdot \left[\Psi + N_k\mathbf{S}_k + \frac{N_k\lambda}{N_k+\lambda}(\bar{\mathbf{s}}_k - \mathbf{m}_0)(\bar{\mathbf{s}}_k - \mathbf{m}_0)^T\right] \quad (12)$$

3.2. Classification

After the learning of the parameters of the Gaussians, $\hat{\theta}_{1:K} = \{\hat{\mathbf{m}}_{1:K}, \hat{\Sigma}_{1:K}\}$, we can use the model for classification. We perform the classification step by maximizing the posterior of the class labels $\mathbf{z}_{test}$. The posterior of the class labels is factorized as follows

$$p(\mathbf{z}_{1:N}, \beta|\mathbf{s}_{1:N}, \hat{\theta}_{1:K}) \propto p(\mathbf{s}_{1:N}|\mathbf{z}_{1:N}, \hat{\theta}_{1:K})p(\mathbf{z}_{1:N}|\beta) \quad (13)$$

where the likelihood term (the first term on the righthand side) is given by

$$p(\mathbf{s}_{1:N}|\mathbf{z}_{1:N}, \hat{\theta}_{1:K}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathbf{s}_n|\hat{\theta}_k)^{z_{n,k}}$$

The joint distribution of the pixel labels $p(\mathbf{z}_{1:N}|\beta)$ in (13) can be constructed by using the conditional probabilities $p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}, \beta)$ defined in (3). Based on the conditional independence assumption that $p(\mathbf{z}_{n'}|\mathbf{z}_{\{1:N\}\backslash n'}, \beta) = p(\mathbf{z}_{n'}|\mathbf{z}_{\mathcal{M}(n')}, \beta)$, we define the joint probability of the random field as follows

$$p(\mathbf{z}_{1:N}|\beta) = \frac{\prod_{k=1}^{K}\exp\left\{\beta\sum_{n=1}^{N} z_{n,k}\left(1+\frac{1}{2}\sum_{m\in\mathcal{M}(n)} z_{m,k}\right)\right\}}{\mathcal{Z}(\beta)} \quad (14)$$

where $\mathcal{Z}(\beta)$ is the normalization term.

In order to perform a posterior inference, we use the ICM algorithm. We update the variables along the iterations in the following order:

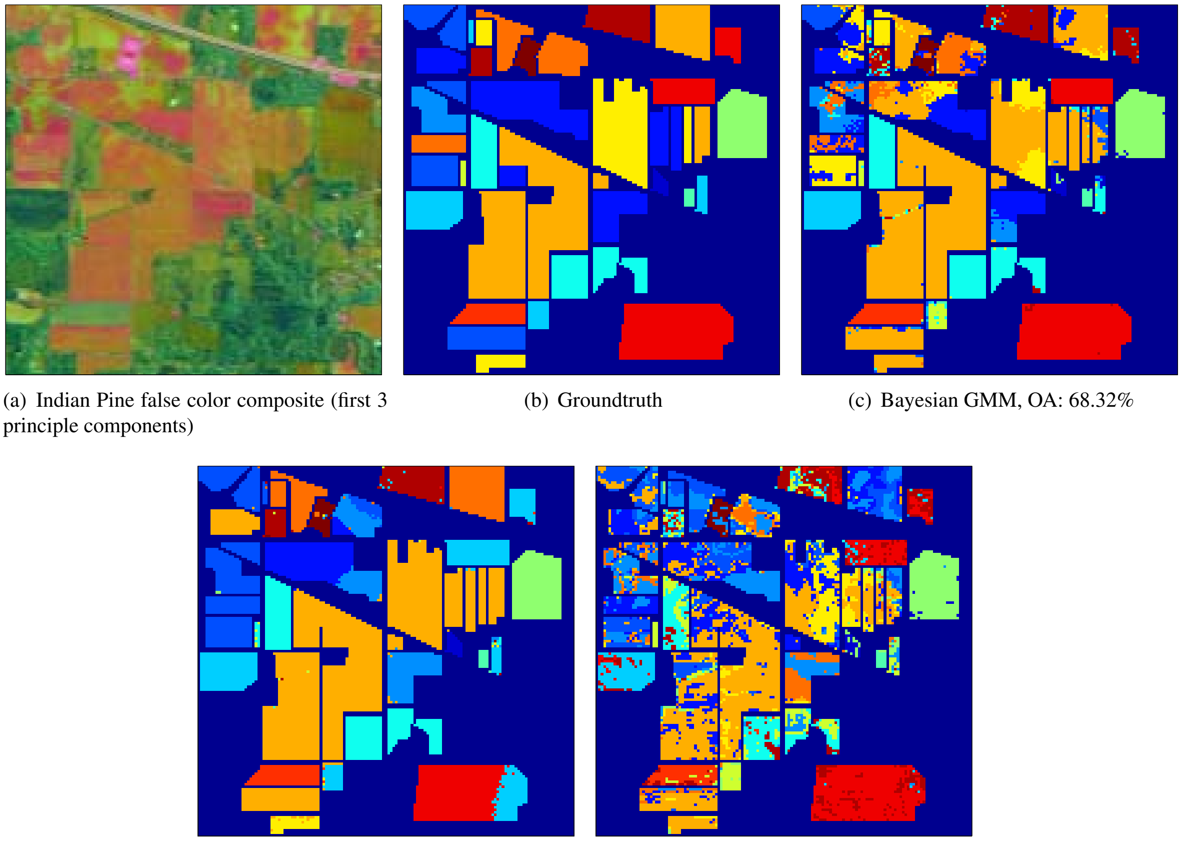
$$\begin{aligned} \mathbf{z}_n^t &\leftarrow \max_{\mathbf{h}_n} p(\mathbf{h}_n|\mathbf{z}_n, \phi_{1:K}^{t-1})p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}^{t-1}, \beta^{t-1}) \\ \beta^t &\leftarrow \max_{\beta}\prod_{n=1}^{N} p(\mathbf{z}_n^t|\mathbf{z}_{\mathcal{M}(n)}^t, \beta) \end{aligned} \quad (15)$$

where $n = 1,..,N$, $k = 1,..,K$ and t is the pseudo time index. We use the Besag's pseudo-likelihood [12] approach to estimate the smoothing parameter β, because it provides a tractable and computationally cheap estimator for β.

4. EXPERIMENTAL RESULTS

We present a comparison of the proposed Bayesian GMM-based method and two state-of-the-art methods, namely psuedo-LDA and diagonal-QDA. Due to small sample size problem, estimated covariances are not positive definite. Thus, conventional LDA and QDA are not applicable to HSI image classification. Psuedo-LDA method uses singular value decomposition (SVD) for covariance matrix inversion. In diagonal-QDA, covariances are approximated by only using the diagonal terms.

For experiments, we use the well-known HSI data set Indian Pines that is obtained by Airborne Visible Infrared Imaging Spectrometer (AVIRIS) over Northern Indiana on June 12, 1992. The data set contains a 145 × 145 pixels and 220 bands HSI and a 16-class ground-truth map. We remove the 20 noisy bands and use 200 spectral bands in our experiments. The ground-truth set contains 10249 labeled pixels from 16 classes. We randomly divide the groundtruth set into a set of 5128 training samples and 5121 test samples. We learn the



(a) Indian Pine false color composite (first 3 principle components)

(b) Groundtruth

(c) Bayesian GMM, OA: 68.32%

(d) pseudo-LDA, OA: 62.99%

(e) diagonal-QDA, OA: 50.14%

Fig. 3. HSI classification maps obtained by proposed Bayesian GMM, pseudo-LDA and diagonal-QDA methods.

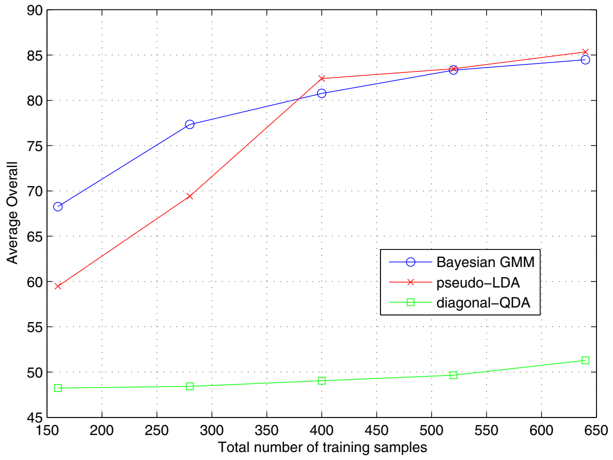


Fig. 2. Average OA (in percent) as functions of total number of training samples.

model using randomly selected 10, 20, 30, 40 and 50 samples per class from training set. For small sample size classes i.e. alfaalfa (46), grass-pasture-mowed (28), oats (20) and stone-steel-towers (93), the number of training samples is set to 10. Overall (OA) accuracy is computed using the all test samples.

Fig. 2 shows the performance of three algorithms. The OA values are obtained by averaging the results of 20 random runs of the algorithms. As seen from the plots in Fig. 2, Bayesian GMM-based method yields better results for the sample size smaller than about 400 that approximately corresponds to 25 samples per class. Diagonal-QDA method produces the worst results since it requires bigger sample size to estimate the unknown parameters. In Fig. 3, classification maps are demonstrated. These maps are obtained using 160 training samples. As denoted in Fig. 3, OA of Bayesian GMM is higher than the others.

5. CONCLUSION

In this study, we propose a Bayesian GMM model for HSI image classification. The results show that the proposed model provides an alternative solution for HSI image classification for small sample size training data set. The propsed Bayesian mixture model can be a competitor for linear and quadratic classifiers as its current form. Using robust distributions such as student's t instead of Gaussian, the model may be improved to compete with kernel-based classifiers. The inference part of the study can be easily modified for unsupervised classification, but initialization of the class parameters and determination of the number of classes are two difficulties of unsupervised classification.

REFERENCES

[1] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using Gaussian mixture models and Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 153–157, 2014.

[2] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, 2009.

[3] A. Berge and A. H. S. Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, 2006.

[4] S. Sanjay-Gopal and T. J. Hebert, "Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm," *IEEE Trans. Image Process.*, vol. 7, no. 7, pp. 1014–1028, 1998.

[5] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[6] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A spatially constrained mixture model for image segmentation," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 494–498, 2005.

[7] A. Diplaros, N. A. Vlassis, and T. Gevers, "A spatially constrained generative model and an EM algorithm for image segmentation," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 798–808, 2007.

[8] C. Fernandez and P. J. Green, "Modelling spatially correlated data via mixtures: A Bayesian approach," *J. R. Stat. Soc. B*, vol. 64, no. 4, pp. 805–826, 2002.

[9] M. A. T. Figueiredo, "Bayesian image segmentation using Gaussian field priors," in *Energy Minimization Methods Computer Vision and Pattern Recognition*, 2005, vol. LNCS 3757, pp. 74–89.

[10] K. Kayabol, A. Voisin, and J. Zerubia, "SAR image classification with non-stationary multinomial logistic mixture of amplitude and texture densities," in *Int. Conf. Image Process. ICIP*, 2011, pp. 173–176.

[11] K. Kayabol and J. Zerubia, "Unsupervised amplitude and texture classification of SAR images with multinomial latent model," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 561–572, 2013.

[12] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Stat. Soc. B*, vol. 36, no. 2, pp. 192–236, 1974.