On discriminant analysis techniques and correlation structures in high dimensions

Clemmensen, Line Katrine Harder

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

On discriminant analysis techniques and correlation structures in high dimensions

Line H. Clemmensen
Technical Report-2013-04
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark

March 14, 2013

Abstract

This paper compares several recently proposed techniques for performing discriminant analysis in high dimensions, and illustrates that the various sparse methods differ in prediction abilities depending on their underlying assumptions about the correlation structures in the data. The techniques generally focus on two things: Obtaining sparsity (variable selection) and regularizing the estimate of the within-class covariance matrix. For high-dimensional data, this gives rise to increased interpretability and generalization ability over standard linear discriminant analysis. Here, we group the methods in two: Those who assume independence between the variables and thus use a diagonal estimate of the within-class covariance matrix, and those who assume dependence between the variables and thus use an estimate of the within-class covariance matrix, which also estimates the correlations between variables. The two groups of methods are compared and the pros and cons are exemplified using different cases of simulated data. The results illustrate that the estimate of the covariance matrix is an important factor with respect to choice of method, and the choice of method should thus be driven by the nature of the problem at hand.

1 Introduction

Linear discriminant analysis (LDA) was first introduced in 1936 Fisher (1936). LDA is a widely used technique for supervised classification when the number of observations, n is larger than the number of variables, p ($n > p$). However, when the number of variables exceeds the number of observations ($p > n$), LDA fails to give accurate predictions as the within-class covariance matrix becomes singular. Recently, much emphasis has been put on developing new techniques which overcome this problem, see e.g. Hastie et al. (1995); Tibshirani et al. (2003); Guo et al. (2007); Clemmensen et al. (2011); Witten and Tibshirani (2011); Shao et al. (2011). These techniques focus on two things. First, they introduce sparsity where a number of parameters is set to zero in order to exclude variables irrelevant to the separation of classes. Second, they regularize the estimate of the within-class covariance matrix to achieve full rank. The previous papers give discussions and promote each their method based on the choice of algorithm due to speed, the choice of cost function, or the choice of sparsity measure. We illustrate that these choices are not the priority when the best classification rate is the goal.

In this paper, we show that it is the estimate of the within-class covariance matrix, which is the most important factor for good predictions. The choice of estimate of the within-class covariance matrix is based on an underlying assumption about the correlation structure between the covariates. If this assumption is right, the predictions are better. Therefore, we will not go into details on algorithms, object functions and speed of the methods, but focus on the estimate of within-class covariance matrix.

The paper is organized as follows. First, I briefly summarize linear discriminant analysis and the newly developed versions of this and give the estimates of the within-class covariance matrices used in each of the techniques (Section 2). Secondly, I describe the simulated data (Section 3) and the results obtained with the various techniques (Section 4), and finally the paper summarizes with a discussion of the results and the techniques (Section 5).

2 Methods

This section first reviews linear discriminant analysis (LDA), which is a widely used method for classification Fisher (1936). Secondly, it briefly re-

views the some of the newer classification techniques, which modify LDA to work in settings with more variables than observations ($p > n$). In particular, the estimates of the within-class covariance matrix are given for each of the models. It is not an extensive work of all methods which perform discriminant analysis in high dimensions. Other techniques could be considered, such as various shrinkage approaches of the covaraince estimates, see e.g. Schäfer and Strimmer (2005), but are left out in this analysis. The selected techniques have been split into two subsections. The first one includes the techniques that use a diagonal estimate of the within-class covariance matrix. The second one includes the techniques that also estimate off-diagonal elements of the within-class covariance matrix.

2.1 Linear discriminant analysis

In LDA, we consider data which can be modelled by K classes of Gaussian normals, i.e. the k^th class has the distribution $C_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with mean value $\boldsymbol{\mu}_k$ and common covariance $\boldsymbol{\Sigma}$, $k = 1, ..., K$. The maximum likelihood estimate of the within-class covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T, \tag{1}$$

where $\hat{\boldsymbol{\mu}}_k = 1/n_k \sum_{i \in C_k} \mathbf{x}_i$ is the maximum likelihood estimate of the mean of the n_k observations in the k^{th} class. If data are normalized, such that each variable has zero mean and length one, we have that $\hat{\boldsymbol{\Sigma}}$ is an estimate of the within-class correlation matrix rather than the within-class covariance matrix.

A new observation $\mathbf{x}_{new}$ is classified using the rule $\max_{C_k} \{\boldsymbol{\mu}_k \boldsymbol{\Sigma}^{-1} \mathbf{x}_{new}^T - \frac{1}{2} \boldsymbol{\mu}_k \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k^T\}$ when assuming equal priors and losses, see e.g. Hastie et al. (2009).

However, when the number of variables is larger than the number of observations ($p > n$), the within-class covariance matrix becomes singular and LDA fails. In the next sections we review methods, which overcome this problem.

2.2 Assuming independence

One approach to regularization of the estimate of the within-class covariance matrix is to use a diagonal estimate. This is similar to a univariate regression

approach Sjöstrand et al. (2008), and thus assumes independence between the covariates. Two such methods are nearest shrunken centroids (NSC) and penalized linear discriminant analysis (PLDA).

2.2.1 Nearest shrunken centroids

In NSC Tibshirani et al. (2003) the covariance is estimated as the diagonal of the full covariance estimate $\hat{\mathbf{\Sigma}}_{NSC} = \text{diag}(\hat{\mathbf{\Sigma}})$, and the class means are shrunken using soft thresholding:

$$\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k^* = \text{sign}(\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k)(|\hat{\mathbf{\Sigma}}_{NSC}^{-1}\hat{\boldsymbol{\mu}}_k| - \Delta)_+, \tag{2}$$

where Δ is a constant, and tunes the degree of sparsity in the model and $(\cdot)_+$ denotes a thersholding to zero when the value is less than zero.

2.2.2 Penalized linear discriminant analysis

PLDA Witten and Tibshirani (2011) uses the Fisher's scoring problem as a starting point for their technique. Again, the diagonal estimate of the within-class covariance matrix is used, $\tilde{\mathbf{\Sigma}}_{PLDA} = \text{diag}(\hat{\mathbf{\Sigma}})$. The sparse discriminant directions are found using an L_1 penalty on the parameter estimates of the directions in the Fisher's discriminant problem, and the solution is found using a minorization algorithm which iteratively decreases the objective function until a local optimum is reached. The weight on the L_1 penalty, λ controls the degree of sparseness in the model.

This method can also be used with a fused lasso penalty Tibshirani and Saunders (2005) if an ordering of the variables is known a priori.

2.3 Assuming correlations exist

Another approach to regularization of the estimate of the within-class covariance matrix is to take into account the correlation structure between the covariates. These techniques in general range from an estimate with a full correlation structure to a diagonal estimate of the covariance matrix, depending on the weight of the penalization. Four such methods are: penalized discriminant analysis (PDA), regularized discriminant analysis (RDA), sparse discriminant analysis (SDA), and sparse linear discriminant analysis by thresholding (SLDAT).

2.3.1 Penalized discriminant analysis

PDA was proposed in Hastie et al. (1995), and uses a ridge-type Hoerl and Kennard (1970) regularization of the within-class covariance estimate:

$$\hat{\mathbf{\Sigma}}_{PDA}(\gamma) = \hat{\mathbf{\Sigma}} + \gamma \mathbf{I}, \tag{3}$$

where $\gamma \geq 0$. Letting $\gamma = 0$ gives a full estimate of the covariance matrix, and letting $\gamma \to \infty$ gives an identity matrix as the estimate of the covariance matrix. Hence, γ controls the degree of diagonalization of the within-class covariance matrix. PDA introduces no sparsity and therefore we have not included the method in the later comparisons.

2.3.2 Regularized discriminant analysis

Closely related to PDA, but introducing sparsity in the estimates of the class means is shrunken centroids regularized discriminant analysis (RDA) proposed in Guo et al. (2007). In RDA, the within-class covariance matrix is a weighted average of the full estimate and the diagonal estimate

$$\hat{\mathbf{\Sigma}}_{RDA}(\alpha) = \alpha \hat{\mathbf{\Sigma}} + (1 - \alpha)\text{diag}(\hat{\mathbf{\Sigma}}), \tag{4}$$

where $0 \leq \alpha \leq 1$. Letting $\alpha = 0$ gives a diagonal estimate of the within-class covariance matrix, and letting $\alpha = 1$ gives a full estimate of $\hat{\mathbf{\Sigma}}$. When data is normalized $\hat{\mathbf{\Sigma}}$ is the estimate of the correlation matrix, and we see that it is equivalent to the correlation matrix estimate in PDA. The sparsity is introduced by shrinking the class means as

$$\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k^* = \text{sign}(\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k)(|\hat{\mathbf{\Sigma}}_{RDA}^{-1}\hat{\boldsymbol{\mu}}_k| - \Delta)_+, \tag{5}$$

like in NSC, but with a different estimate of $\hat{\mathbf{\Sigma}}$, where Δ is a positive constant. The feature selection properties of this form of shrunken centroids are considered conservative as it in general includes a large number of variables. Here, α, and Δ control the degree of diagonalization of the within-class covariance matrix and the degree of sparsity, respectively.

2.3.3 Sparse discriminant analysis

In the *sparse discriminant analysis* (SDA) algorithm proposed in Clemmensen et al. (2011) the discriminant problem is recast as optimal scoring

which is a regression type problem. The discriminant directions β_k, $k = 1, ..., K-1$ are found through an optimal scoring problem using both the L_2, and L_1 penalties through the elastic net Zou and Hastie (2005). The discriminant directions are penalized with the model parameters λ and *stop*, where λ is a positive constant which controls the degree of diagonalization, and *stop* controls the degree of sparsity. A full covariance estimate is used based on the sparse discriminant directions $\hat{\mathbf{\Sigma}}_{SDA} = 1/n \sum_{k=1}^{K} \sum_{i \in C_k} (\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_k)(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_k)^T$, where $\tilde{\mathbf{x}}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\mu}}_k = 1/n_k \sum_{i \in C_k} \tilde{\mathbf{x}}_i$. As the sparse directions are both ridge and lasso penalized, the covariance estimate will depend only on the active variables and will be ridge regularized as in PDA.

This method can also be used for mixtures of Gaussians and thus a non-linear separation of classes.

2.3.4 Sparse linear discriminant analysis by thresholding

Another variant of LDA that introduces sparsity and goes from a full estimate of the covariance matrix to a diagonal estimate is called *sparse linear discriminant analysis by thresholding* (SLDAT) and was proposed in Shao et al. (2011). SLDAT uses thresholding to induce sparsity into the estimate of the covariance matrix in the following manner

$$\hat{\mathbf{\Sigma}}_{ij,SLDAT} = \hat{s}_{ij} I(|\hat{s}_{ij}| > t_1), \text{ with } t_1 = M_1 \sqrt{\log p}/\sqrt{n}, \tag{6}$$

where M_1 is a positive constant in general, and specifically $0 \leq M_1 \leq \sqrt{n}/\sqrt{\log p}$ when $\hat{\mathbf{\Sigma}}$ is the correlation matrix. Additionally, $\hat{s}_{ij}$ is the $(i,j)^{th}$ element of $\hat{\mathbf{\Sigma}}$, and $I(\mathcal{A})$ is the indicator function of the set $\mathcal{A}$. Letting $M_1 \to \infty$ gives a diagonal estimate of $\mathbf{\Sigma}$, and letting $M_1 = 0$ gives a full estimate of $\mathbf{\Sigma}$. In the case where $\hat{\mathbf{\Sigma}}_{,SLDAT}$ is not invertible a generalized inverse is used.

SLDAT additionally introduces sparseness on the difference between the class means, likewise by thresholding parameter estimates at a level t_2, where $t_2 = M_2(\log p/n)^{0.3}$, and M_2 is a positive constant. The difference between the means of class k and l is then given as $\tilde{\delta}_{i,kl} = \hat{\delta}_{i,kl} I(|\hat{\delta}_{i,kl}| > t_2)$, where $\hat{\boldsymbol{\delta}}_{kl} = \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_l$ and $\hat{\delta}_{i,kl}$ is the i^{th} element of $\hat{\boldsymbol{\delta}}_{kl}$.

Here, M_1 and M_2 control the degree of diagonalization and the degree of sparsity, respectively. In Shao et al. (2011), it is shown that if $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is bounded, then SLDAT is asymptotically optimal.

3 Simulations

This section describes the simulations conducted to illustrate and examine the limits of the assumptions of independent versus correlated variables.

The aim is to simulate different correlation structures as the claim is that the techniques will perform different according to the correlation structures in the data. Two simulation settings are considered with different correlation structures between the variables, and the case where the variables are independent (the correlation is set to zero in either of the two cases).

3.1 Data description

In the first case, there are four classes of Gaussian distributions with independent or dependent features, $C_k : \mathbf{x}_i \sim N(\boldsymbol{\mu_k}, \boldsymbol{\Sigma})$, $k = 1, ..., 4$. The means separate the classes in blocks of 100 variables $\boldsymbol{\mu_{jk}} = 0.7 \times 1_{((k-1)\times 100+1 \leq j \leq k\times 100)}$. The covariance structure of $\boldsymbol{\Sigma}$ is block diagonal with 100 variables in each block, and the blocks have the $(j, j')^{th}$ element $r^{|j-j'|}$, where $0 \leq r \leq 1$. These simulations are similar to those in Witten and Tibshirani (2011), but here with varying correlation degrees r.

The experiments were

- S1: Independent variables with $p = 500$, and $r = 0$.
- S2: Correlated variables with $p = 500$, and $r = 0.99$.
- S3: Correlated variables with $p = 1000$, and $r = 0.99$.
- S4: Correlated variables with $p = 1000$, and $r = 0.9$.
- S5: Correlated variables with $p = 1000$, and $r = 0.8$.

Another simulation was conducted with the same settings as above, except from the correlation structure $\boldsymbol{\Sigma}$ where all correlations in the off-diagonal were equal to ρ. In this setting, the following simulations were performed

- X1: Correlated variables with $p = 1000$ and $\rho = 0.8$.
- X2: Correlated variables with $p = 1000$ and $\rho = 0.6$.
- X3: Correlated variables with $p = 1000$ and $\rho = 0.4$.
- X4: Correlated variables with $p = 1000$ and $\rho = 0.2$.

- X5: Independent variables with $p = 1000$ and $\rho = 0$.

The data were simulated using the package `mvtnorm` in R by drawing data from a multivariate Gaussian distribution using `rmvnorm` with the above mentioned parameters.

3.2 Procedure

An outcome of 1200 observations were simulated given the distributions in the previous section. One hundred observations were used to train a model, and another hundred observations were used to validate the model and tune model parameters, and finally another 1000 observations were used to test the model and estimate the prediction error. This setting was repeated 25 times and mean values and standard deviations were calculated.

A grid was spanned for the model parameters for each of the methods. The grids used were as follows:

- PLDA: $\lambda \in \{0 : 0.03 : 3\}$
- NSC: δ The default 30 values tried in the `pamr` package which range from 0 to the relation between the maximum absolute difference between the centroids of each group and the overall centroids over the standard deviation of these differences.
- SDA: stop $\in \{-500 : 100 : -300, -250 : 50 : -150, -120 : 20 : -80, -70 : 10 : -50\} \times \lambda \in \{10^2, 10^3, 10^4, 10^6\}$ chosen similar to the values in Clemmensen et al. (2011).
- RDA: $\delta \in \{0 : 0.11 : 0.99\} \times \alpha \in \{0 : 0.33 : 3\}$ the default in the `rda` package.
- SLDAT: $M_1 \in \{0.001, 0.01, 0.1 : 0.475 : 2, 3 : 5\}$ and $M_2 \in \{0.001, 0.01, 0.1 : 0.475 : 2, 3 : 5\}$ Note that the values are in a different range than those used in Shao et al. (2011). The size of the model parameters for SLDAT can differ a lot from problem to problem depending on the variance of the variables.

4 Results

R was used to conduct the simulations and estimate prediction errors for each of the techniques. The following R packages from CRAN (2009) were used: `penalizedLDA` (PLDA), `pamr` (NSC), `rda` (RDA), `sparseLDA` (SDA), and the author's own implementation of SLDAT.

The results are summarized in Table 1. RDA performs well for all simulations reflecting that the model parameters can tune the model from a diagonal estimate of the within-class covariance matrix to a full estimate. SDA does not perform well when the variables are completely independent, but gives the best estimates when a strong degree of correlations exists. NSC and PLDA perform best when the variables are independent. SLDAT performed best when the variables were correlated, but in general did not perform quite as well as RDA and SDA, though considerably better than PLDA and NSC in the cases where correlations exist.

The results of NSC performing better than PLDA is consistent with results in Witten and Tibshirani (2011), where it was penalized LDA with a fused lasso penalization which in general gave the lowest errors. Results of SLDAT performing worse than RDA are not consistent with the results reported in Shao et al. (2011). This may be caused by the problems analyzed or the values of model parameters examined. It is noted that the selected model parameters generally were within the examined range, though the examined grid did not have as high a resolution as in the original paper. Results of SDA performing similar to or slightly better than RDA in problems with highly correlated covariates are consistent with results in Clemmensen et al. (2011).

It is worth noting that for SDA, RDA and SLDAT the erros drop considerably when some of the variables are correlated (compare simulations S1 with S2-5). This is also the case for NSC and PLDA when all variables are correlated (compare simulations X5 with X1-4).

5 Discussion

The performance results on the different test data matched the estimates of the within-class covariance matrix. Thus, the methods with a diagonal estimate, assuming independence between variables, performed best when data indeed were simulated from a distribution with independence (penal-

Table 1: Summary of mean errors and mean number of non-zero features in the solutions for each of the methods and each of the simulations. The means are taken over 25 simulations, and the standard errors are given in the parentheses. The lowest numbers of errors for each simulation is in bold.

		PLDA	NSC	SDA	RDA	SLDAT
S1:	#errors	116.6(4.3)	**88.5**(2)	124.4(4.6)	**90.9**(2.4)	141.2(5.9)
	#features	348(18.8)	276(17.1)	261.7(18.1)	218.1(12.3)	292(23.1)
S2:	#errors	539.72(23.9)	424.84(26.6)	**0**(0)	**0.36**(0.3)	13.2(10.8)
	#features	264.32(34)	143.92(12.3)	500(0)	449.52(14.7)	473.28(14.8)
S3:	#errors	602.1(18.8)	449.2(24.9)	**0**(0)	**0.04**(0)	18.6(6.5)
	#features	444.4(69.5)	170.2(27.3)	847.6(1.6)	715.9(39.2)	890.8(43.5)
S4:	#errors	622.4(18.2)	440.2(21)	**0.12**(0.1)	3.1(0.8)	256.9(24.7)
	#features	566.9(66.6)	153.5(23)	841.4(10.8)	955.7(35.8)	711(76.9)
S5:	#errors	550.7(22.7)	412.9(26.9)	**2.2**(0.4)	5(1.4)	397.4(21.9)
	#features	436.2(68)	161.6(21.1)	814.3(18.2)	867.7(62.4)	585(85.2)
X1:	#errors	166.9(10.1)	58.4(10.4)	**0**(0)	2.2(0.6)	12.5(1.5)
	#features	133.7(16.6)	125.6(24.8)	857.4(1.7)	376.4(86.7)	725.8(73.3)
X2:	#errors	134.7(7.9)	29(6.2)	**0**(0)	6.72(2.1)	42.4(6.6)
	#features	155.2(6.6)	141(14.3)	857.3(2.1)	293(81.1)	218.3(53.9)
X3:	#errors	106.3(7.8)	17.4(3.4)	**0.04**(0)	7.12(1.5)	21.4(6.1)
	#features	192.2(6.5)	161.6(30.6)	858.3(1.8)	477.4(94.2)	125.6(6.3)
X4:	#errors	36(4.3)	5.6(1.1)	**0.08**(0.1)	6.4(1.4)	5(1.5)
	#features	245.2(36.4)	363.5(47.8)	862.4(1.7)	594.9(93)	181.2(16.8)
X5	#errors	166.3(6.7)	**116.7**(3.3)	174.6(4.2)	**120**(5.1)	211.7(6.2)
	#features	418.2(45.1)	320.6(33.4)	339.6(27)	296(22.3)	357.4(50.8)

ized linear discriminant analysis - PDA, nearest shrunken centroids - NSC). Similarly, the methods which estimate the off-diagonal as well, assuming correlations between variables exist, performed best when data were simulated from a distribution with correlations (sparse discriminant analysis - SDA, shrunken centroids regularized discriminant analysis - RDA, sparse linear discriminant analysis by tresholding - SLDAT).

In practice a correlation matrix of high dimensions cannot be calculated, but a correlation matrix of a subset of say 100 variables is feasible to calculate and may reveal the correlation structures of the given data. A choice of method could then be based on the observed correlation structure, or simply by a priori knowledge of the associations in the problem at hand.

Another practical issue to consider is the interpretability of the given models, where models which give low-dimensional projections of data can be very useful. The methods which provide such low-dimensional projections are SDA and PLDA. Finally, if speed is an issue, then PLDA, NSC and RDA are faster alternatives than SDA and SLDAT.

References

Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B., 2011. Sparse discriminant analysis. Technometrics(To appear).

CRAN, 2009. The comprehensive r archive network.
URL `http://cran.r-project.org/`

Fisher, R., 1936. The use of multiple measurements in axonomic problems. Annals of Eugenics 7, 179–188.

Guo, Y., Hastie, T., Tibshirani, R., 2007. Regularized linear discriminant analysis and its applications in microarrays. Biostatistics 8 (1), 86–100.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. The Annals of Statistics 23 (1), 73–102.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd Edition. Springer.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Schäfer, J., Strimmer, K., 2005. A shrinkage appraoch to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications inGenetics and Molecular Biology 4 (1), 32.

Shao, J., Wang, G., Deng, X., Wang, S., 2011. Sparse linear discriminant analysis by thresholding for high dimensional data. The Annals of Statistics 39 (2), 1241–1265.

Sjöstrand, K., Carden, V. A., Larsen, R., Studholme, C., 2008. A generalization of voxel-wise procedures for highdimensional statistical inference using ridge regression. In: .Reinhardt, J. M., Pluim, J. P. W. (Eds.), SPIE. SPIE 6914, Medical Imaging.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statistical Science 18, 104–11.

Tibshirani, R., Saunders, M., 2005. Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society - Series B 67 (1), 91–108.

Witten, D., Tibshirani, R., 2011. Penalized classification using fisher's linear discriminant. Journal of the Royal Statistical Society, Series B.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society - Series B 67 (Part 2), 301–320.