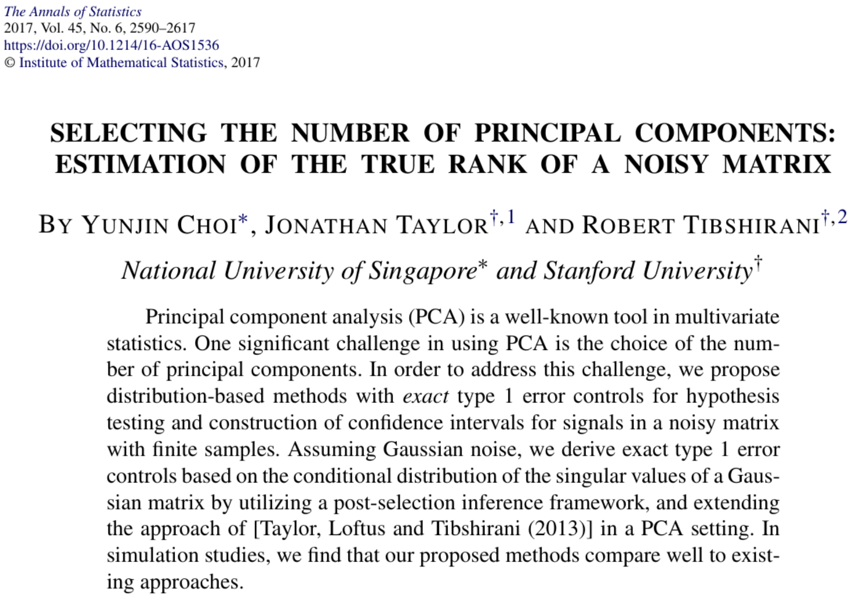
Choi, Taylor and Tibshirani (2017, AoS)

Prelim. for Matsumura & Tachibana

Yasuyuki Matsumura (Kyoto University)

Last Updated: September 22, 2025

yasu0704xx.github.io



The Annals of Statistics
2017, Vol. 45, No. 6, 2590–2617
https://doi.org/10.1214/16-AOS1536
© Institute of Mathematical Statistics, 2017

SELECTING THE NUMBER OF PRINCIPAL COMPONENTS: ESTIMATION OF THE TRUE RANK OF A NOISY MATRIX

BY YUNJIN CHOI*, JONATHAN TAYLOR†,1 AND ROBERT TIBSHIRANI†,2

National University of Singapore and Stanford University†*

Principal component analysis (PCA) is a well-known tool in multivariate statistics. One significant challenge in using PCA is the choice of the number of principal components. In order to address this challenge, we propose distribution-based methods with *exact* type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples. Assuming Gaussian noise, we derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of [Taylor, Loftus and Tibshirani (2013)] in a PCA setting. In simulation studies, we find that our proposed methods compare well to existing approaches.

Choi, Taylor, and Tibshirani (2017, AoS)

- In using principal component analysis (PCA), the choice of the number of principal components is a significant challenge.
- Choi, Taylor and Tibshirani (2017) propose distribution-based methods with exact type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples.
- Assuming Gaussian noise, Choi, Taylor and Tibshirani (2017) derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of Taylor, Loftus and Tibshirani (2016) in a PCA setting.

Contents

- Introduction
- Proposed Distribution-Based Methods
  - Review of the Kac-Rice Test
  - General Hypothesis Testing
  - Confidence Interval Construction
- Rank Estimation
- Estimating the Noise Level
- Additional Examples
- Conclusions

Introduction

Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) is a commonly used method in multivariate statistics.
  - a descriptive tool for examining the structure of a data matrix
  - a pre-processing step for reducing the dimension of the column space of the matrix (Josse and Husson, 2012)
  - matrix completion (Cai, Candés and Shen, 2010)
- One important challenge: How to determine the number of principal components to retain in PCA?
- A summary by Jolliffe (2002)
  1. subjective methods (e.g., scree plot)
  2. distribution-based test tools (e.g., Bartlett's test)
  3. computational procedures (e.g., cross-validation)

Example (Mardia, Kent and Bibby, 1979)

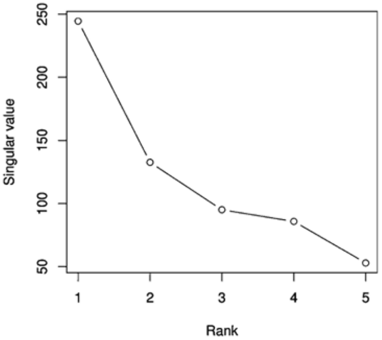


FIG. 1. *Singular values of the score data in decreasing order. The data consist of exam scores of 88 students on five different topics (Mechanics, Vectors, Algebra, Analysis and Statistics).*

This Paper

- Choi, Taylor and Tibshirani (2017) propose a class of statistical methods utilizing hypothesis testing framework for determining the rank of the signal matrix in a noisy matrix model.
- The estimated rank here corresponds to the number of principal to retain in PCA.
- Under Gaussian assumption, the proposed hypothesis testing method provides exact type 1 error controls along with exact confidence intervals of signal parameters in finite samples.

Related Works (Polyhedral Approach)

- Tibshirani, Taylor, Lockhart and Tibshirani (2017)
  - the truncated Gaussian test conditioning on the event of selecting active variables
  - utilizing the fact that the selection event can be characterized by an observed data vector y falling into a polyhedral set
- In PCA, the event of selecting principal components cannot be characterized by an observed data vector y falling into a polyhedral set.
  - Selecting a variable is a discrete event of forward stepwose regression.
  - On the other hand, a principal components in PCA are chosen from a continuum for Y being a matrix.
  - As the domain of selection event is a continuum, the resulting null distribution conditional on the selection event is defined on a measure zero domain rather than being a truncated Gaussian distribution.

Kac-Rice Test

- The *Kac-Rice test*: an exact method for testing and constructing confidence intervals for signals under a global null hypothesis in adaptive regression.
- Under the global null scenario, one of the proposed methods corresponds to the application of the *Kac-Rice test* to a penalized regression minimizing the Frobenius norm with a nuclear norm penalty.
- Choi, Taylor and Tibshirani (2017) extend the *Kac-Rice test* and the construction of confidence intervals to not only to the global null scenario but the general case.
- Also in the global null scenario, one of the extended methods provides stronger power than the *Kac-Rice test*.
- The exact property of the *Kac-Rice test* is preserved in extension to a general step by incorporating a post-selection inference framework.
- The resulting statistics use a conditional survival function of the eigenvalues of a Wishart matrix.

Literature & Contribution

- Inference based on the distribution of eigenvalues
  - Muirhead (1982, Theorem 9.6.2) : a likelihood ratio test with the asymptotic Chi-square distribution
  - Kritchman and Nadler (2008) : the asymptotic distribution of the largest eigenvalue of a Wishart matrix (the Tracy-Widom law), incorporating the result of Johnstone (2001)
- These test methods show conservative results and thus lose signal detection power in the general stage.
- Choi, Taylor and Tibshirani (2017) provide exact type 1 error controls and decent detection power at the same time, and additionally provide a method for constructing confidence intervals in addition to hypothesis testing.

Proposed Distribution-Based Methods

Setup

- Assume that the observed data matrix $Y \in \mathbb{R}^{N \times p}$ is the sum of a low-rank signal matrix $B \in \mathbb{R}^{N \times p}$ and a Gaussian noise matrix $E \in \mathbb{R}^{N \times p}$ as follows:

$$\begin{aligned} & Y = B + E, \\ & \text{rank}(B) = \kappa < \min(N, p), \\ & E_{ij} \sim \text{Normal}(0, \sigma^2) \text{ for } i \in \{1, \cdots, N\}, j \in \{1, \cdots, p\} \end{aligned}$$

that is,

$$Y \sim \text{Normal}(B, \sigma^2 I_N \otimes I_p). \tag{1}$$

- Assume $N > p$ w.l.o.g.

- Following the notation from Muirhead (1982, p.73), $Y \sim \text{Normal}(B, \sigma^2 I_N \otimes I_p)$ in (1) denotes

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} B_1 \\ \vdots \\ B_N \end{pmatrix}, \begin{pmatrix} \sigma^2 I_p & 0 & \cdots & 0 \\ 0 & \sigma^2 I_p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 I_p \end{pmatrix} \right),$$

  where Y_j and B_j represent the jth row vectors of Y and B, respectively, and the Kronecker product is denoted by $\otimes$.

- We focus on finding κ, the rank of signal matrix B, and the construction of confidence intervals for the signals in B.

For the Case of a Centered Data Matrix

- With a centered data matrix, the proposed approaches with this model assumption are valid for the popular spiked covariance model as well.
- Both traditional spiked covariance approaches and the proposed tests with a centered data matrix are examining basically the same statistics.
  - Note that the centering operation of a data matrix does not change the rank of B.
- The statistics in interest for the spiked covariance model approaches are the eigenvalues of a covariance matrix, and the proposed methods investigate the singular values of a data matrix.
  - The eigenvalues from the covariance matrix are the same as the squared singular values of a centered data matrix.

Taylor, Loftus and Tibshirani (2016)

- We first review the global null test and confidence interval construction of the first signal of Taylor, Loftus and Tibshirani (2016) and its application in matrix denoising problem, which corresponds to the PCA setting.
- Then we extend the global null test to a general test procedure for testing

$$\begin{aligned} &\mathbb{H}_{k,0} : \text{rank}(B) \leq k-1 \text{ vs } \mathbb{H}_{k,1} : \text{rank}(B) \geq k \\ &\text{for } k = 1, \cdots, p-1, \end{aligned}$$

  and describe how to construct confidence intervals for the kth largest signal parameters.

Review of the Kac-Rice Test: Global Null Hypothesis Testing

- The *Kac-Rice test* provides exact type 1 error controls for a class of regularized regression problems of the following form:

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2} ||y - X\beta||_2^2 + \lambda \cdot \mathcal{P}(\beta) \tag{2}$$

  with an outcome $y \in \mathbb{R}^p$, a predictor matrix $X \in \mathbb{R}^{N \times p}$ and a penalty term $\mathcal{P}(\cdot)$ with a regularization parameter $\lambda \geq 0$.
- Assuming that the outcome $y \in \mathbb{R}^N$ is generated from $y \sim$ Normal$(X\beta, \Sigma)$, the *Kac-Rice test* provides a method for testing

$$\mathbb{H}_0 : \mathcal{P}(\beta) = 0 \tag{3}$$

  that yields exact type 1 error controls under the assumption that the penalty function $\mathcal{P}$ is a support function of a convex set $\mathcal{C} \subseteq \mathbb{R}^p$, that is,

$$\mathcal{P}(\beta) = \max_{u \in \mathcal{C}} u^\top \beta.$$

- When applied to a matrix denoising problem of a popular form, (3) becomes a global null hypothesis:

$$\mathbb{H}_0 : \Lambda_1 = 0 \equiv \mathsf{rank}(B) = 0 \equiv B = 0_{N \times p}, \tag{4}$$

where $\Lambda_1 \geq \cdots \geq \Lambda_p \geq 0$ denote the singular values (SV) of B.

Details

- For an observed data matrix $Y \in \mathbb{R}^{N \times p}$, a widely used method to recover the signal matrix B in (1) is to solve the following criterion:

$$\widehat{B} \in \underset{B \in \mathbb{R}^{N \times p}}{\arg\min} \frac{1}{2}||Y - B||_F^2 + \lambda ||B||_* \qquad \text{where } \lambda > 0, \tag{5}$$

where $||\cdot||_F$ denotes a Frobenius norm, and $||\cdot||_*$ denotes a nuclear norm. [1]

[1] The nuclear norm term plays an analogous role as an ℓ_1 penalty term in lasso regression (Tibshirani, 1996).

- The objective function (5) falls into the class of regression problems described in (2), with the predictor matrix X and the penalty function $\mathcal{P}(\cdot)$ respectively being

$$X = I_N \otimes I_p, \quad \mathcal{P}(B) = ||B||_* = \max_{u \in \mathcal{C}} \langle u, B \rangle$$

  with $\mathcal{C} = \{A : ||A||_{op} \leq 1\}$ where $|| \cdot ||_{op}$ denotes a spectral norm.

- Therefore we can directly apply the *Kac-Rice test* with the resulting test statistic as follows, under the assumed model discussed in (1):

$$\mathbb{S}_{1,0} = \frac{\int_{d_1}^{\infty} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) dz}{\int_{d_2}^{\infty} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) dz}, \tag{6}$$

  where $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ denote the observed singular values of Y.

- The test statistic $\mathbb{S}_{1,0}$ in (6) is uniformly distributed under the global null hypothesis (4), and provides exact type 1 error controls for testing the global null hypothesis.

More on the Kac-Rice Test Statistic in (6)

- The value $\mathbb{S}_{1,0}$ represents the probability of observing more extreme values than d_1 under the null hypothesis, which coincides with the traditional notion of p-value.
- Another view: The test statistic $\mathbb{S}_{1,0}$ corresponds to a conditional survival probability of the largest observed singular value d_1 conditioned on all the other observed singular values $d_2, \cdots, d_p$.
    - The integrand of $\mathbb{S}_{1,0}$ coincides with the conditional distribution of the largest eigenvalue of a central Wishart matrix (James, 1964) via a change of variables.
    - The denominator of $\mathbb{S}_{1,0}$ acts as a normalizing constant since the domain of the largest singular value d_1 conditioning on all the other singular values becomes (d_2, ∞).
- A small magnitude of $\mathbb{S}_{1,0}$ implies large d_1 compared to d_2, and thus supports $\mathbb{H}_1 : \lambda > 0$.

Review of the Kac-Rice Test: Confidence Intervals for the Largest Signal

- Along with the *Kac-Rice test* mentioned above, a procedure constructing an exact confidence interval for the leading signal in adaptive regression is proposed in Taylor, Loftus and Tibshirani (2016).
- By applying the result of Taylor, Loftus and Tibshirani (2016) to the matrix denoising setting, we can generate an exact confidence interval for

$$\tilde{\Lambda}_1 = \langle U_1 V_1^\top, B \rangle,$$

  where
  - $Y = UDV^\top$ is a singular value decomposition (SVD) of Y with $D = \text{diag}(d_1, \cdots, d_p)$ for $d_1 \geq \cdots \geq d_p$
  - U_1 and V_1 are the first column vectors of U and V, respectively.
- It is desirable to directly find the confidence interval for Λ_1 (a leading SV of B). However, as B is unobservable, $U_1 V_1^\top$ is the "best guess" of the unit vector associated with Λ_1 in its direction.

- In the matrix denoising problem of (5), the result from Taylor, Loftus and Tibshirani (2016) yields an exact conditional survival probability of d_1 as follows:

$$\mathbb{S}_{1,\tilde{\Lambda}_1} = \frac{\int_{d_1}^{\infty} e^{-\frac{(z-\tilde{\Lambda}_1)^2}{2\sigma^2}} \, z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) \, dz}{\int_{d_2}^{\infty} e^{-\frac{(z-\tilde{\Lambda}_1)^2}{2\sigma^2}} \, z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) \, dz} \sim \text{Unif}(0,1). \tag{7}$$

- A special case: When data is generated under $\Lambda_1 = 0$, $\tilde{\Lambda}_1 = 0$ holds. Then, $\mathbb{S}_{1,\tilde{\Lambda}_1}$ in (7) is the same as $\mathbb{S}_{1,0}$ for testing $\mathbb{H}_0 : \lambda_1 = 0$, and yields exact type 1 error controls due to its uniformity.

- To construct the level α confidence interval, let

$$\mathsf{CI} = \left\{\delta : \text{min}\left(\mathbb{S}_{1,\delta}, 1 - \mathbb{S}_{1,\delta}\right) > \frac{\alpha}{2}\right\}. \tag{8}$$

Since $\mathbb{S}_{1,\tilde{\Lambda}_1}$ in (7) is uniformly distributed, we observe that

$$\mathbb{P}\left(\tilde{\Lambda}_1 \in \mathsf{CI}\right) = 1 - \alpha,$$

and thus CI in (8) generates an exact level α confidence interval.

General Hypothesis Testing

- Choi, Taylor and Tibshirani (2017) extend the test for the global null in (4) to a genaral test which investigates whether there exists the kth largest signal in B.
- Consider the following hypothesis:

$$\begin{aligned} & \mathbb{H}_{0,k} : \Lambda_k = 0 \text{ vs } \mathbb{H}_{1,k} : \Lambda_k > 0 \\ \Longleftrightarrow \quad & \mathbb{H}_{0,k} : \text{rank}(B) < k \text{ vs } \mathbb{H}_{1,k} : \text{rank}(B) \geq k, \end{aligned} \tag{9}$$

  for $k = 1, \cdots, p-1$.
  - For $k = 1$, the null hypothesis in (9) corresponds to a global null as in (4).
- We assume $\text{rank}(B) < p$ and do not consider the case of $k = p$.
  - At $k = p$, the signal matrix B is full rank under the alternative hypothesis with $\Lambda_p \neq 0$.
  - In this scenario, the problem of $\text{rank}(B) = p$ with the noise level of σ^2 becomes unidentifiable with the problem of $\text{rank}(B) = p - 1$ with the noise level of $\sigma^2 + \Lambda_p^2$.

- One of the most straightforward approaches for extending the global test (6) to testing (9) for $k = 1, \cdots, p-1$ would be to apply it sequentially, with removing the first observed $k-1$ singular values at the kth step, and start at the kth observed singukar value.
- That is, at the kth step, we can ignore the first $k-1$ observed singular values of Y, and apply the test with plugging in the kth value to the place of the 1st value, the $(k+1)$th to the place of the 2nd value, and so on.
- This approach of ignoring the existence of the first $k-1$ singular values is analogous to other methods dealing with essentially the same hypothesis testing (Muirhead, 1982; Kritchman and Nadler, 2008).

How well does the “sequential Kac-Rice test” work?

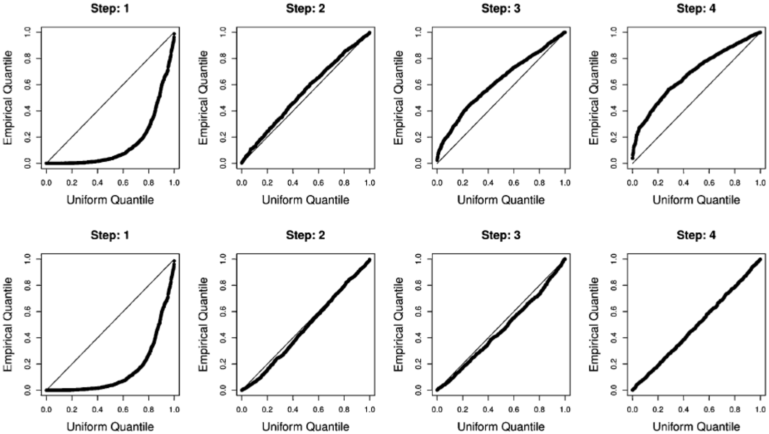


FIG. 2. *Quantile-quantile plots of the observed quantiles of p-values versus the uniform quantiles. With $N = 20$ and $p = 10$, the true rank of B is* $\mathrm{rank}(B) = 1$. *The top panels are from the sequential* Kac–Rice test *and the bottom panels are from the* CSV. *The kth column represents quantile-quantile plots of the p-values for testing $H_{0,k} : \mathrm{rank}(B) \leq k - 1$ for steps $k = 1, 2, 3, 4$.*

- An example in Figure 2 shows one with $N = 20$, $p = 10$, and a rank one signal of moderate size.
- The top panels: Quantile-quantile plots of the p-values for the sequential *Kac-Rice test* versus the uniform distribution
  - The p-values are small when the alternative hypothesis $\mathbb{H}_{1,1}$ is true (step 1), and then fairly uniform for testing rank ≤ 1 versus > 1 (step 2) as described, which is the first case in which the null hypothesis is true.
  - However, the test becomes more and more conservative for higher steps, although the p-values are generated under the null distributions.
  - The conservativeness of these p-values can lead to potential loss of power. [2]

[2] One of the reasons is that, at at $k = 3$, for example, the test does not consider taht the two largest singular values have been removed. The test instead ignores the existence of the 1st and 2nd singular values, and treats the 3rd one as the 1st, plugging it into the place of the 1st singular value. As the 1st and 3rd largest singular values do not have the same distribution with the density of the 1st singular value having more weights on large values, the sequential *Kac-Rice test* at $k = 3$ is no longer uniformly distributed; the test results in conservative p-values.

- The bottom panels: Quantile-quantile plots of the p-values for the conditional singular value (CSV) test to be described below
  - It follows the uniform distribution quite well for all “null” steps.
  - For testing $\mathbb{H}_{0,k} : \Lambda_k = 0$ at the kth step, the CSV method takes it into account that our interest is the kth signal by conditioning on the first $k-1$ singular values when deriving its test statistic.
- The following sections discuss the CSV test procedure in detail, and the ICSV test, an integrated version of the CSV which has better power.

General Hypothesis Testing: The Conditional Singular Value Test

- In this section, we introduce a test in which the test statistic has an "almost exact" null distribution under $\mathbb{H}_{0,k}$ in (9) for $k \in \{1, \cdots, p-1\}$.
- Notations:
    - We write the singular value decomposition of a signal matrix B as $B = U_B D_B V_B^\top$, the singular value decomposition of an observed data matrix Y as $Y = UDV^\top$, and an $N \times r$ and an $N \times (p-r)$ column-wise submatrices of an $N \times p$ matrix M by $M_{[r]}$ and by $M_{[-r]}$ where $M = \left[M_{[r]}, M_{[-r]}\right]$.
    - Also, P_Q denotes a projection matrix onto a column-space of an $n_1 \times n_2$ matrix Q, and $P_Q^\perp$ denotes a projection matrix onto a kernel of P_Q, that is, assuming $n_1 \geq n_2$ and Q is of full-rank, we have

$$\begin{aligned} P_Q &= Q(Q^{\top Q})^{-1}Q^\top, \\ P_Q^\perp &= I_{n_1} - P_Q. \end{aligned}$$

- With these notations, the hypothesis in (9) can be rewritten as

$$\mathbb{H}_{0,k} : P^{\perp}_{U_{B[k-1]}} B P^{\perp}_{V_{B[k-1]}} = 0_{N\times p} \text{ vs } \mathbb{H}_{1,k} : P^{\perp}_{U_{B[k-1]}} B P^{\perp}_{V_{B[k-1]}} \neq 0_{N\times p} \qquad (10)$$

General Hypothesis Testing: The Integrated Conditional Singular Value Test

2.2.2

General Hypothesis Testing: Simulation Examples

2.2.3

Confidence Interval Construction

Confidence Interval Construction: Simulation Studies

2.3.1

Rank Estimation

Estimating the Noise Level

Additional Examples

Conclusions

Conclusions

- Choi, Taylor and Tibshirani (2017) have proposed distribution-based methods for choosing the number of principal components of a data matrix.
    - Both for hypothesis testing anf the construction of confidence intervals of the signals
- The methods have exact type 1 error control and show promising results in simulated examples.
- Choi, Taylor and Tibshirani (2017) have also introduced data-based methods for estimating the noise level.

Further Investigation

- The analysis of the power of the proposed tests
- The width of the constructed confidence interval
- Application to high-dimensional data using numerical approximations
- Multiple hypothesis testing corrections: Required to Study the dependence structure of the p-values between different steps
- Robustness to non-Gaussian noise: Bootstrap version of the proposed procedure
- Degrees of freedom of the spectral estimator of the signal matrix (?)
- Canonical correlation analysis (CCA)
- Linear discriminant analysis (LDA)