



**ISABEL CRISTINA COSTA LEITE**

**ANÁLISE DE COMPONENTES  
INDEPENDENTES APLICADA A AVALIAÇÃO  
DE IMAGEM RADIOGRÁFICA DE SEMENTES**

**LAVRAS - MG**

**2013**

**ISABEL CRISTINA COSTA LEITE**

**ANÁLISE DE COMPONENTES INDEPENDENTES APLICADA A  
AVALIAÇÃO DE IMAGEM RADIOGRÁFICA DE SEMENTES**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientadora  
Dra. Thelma Sáfydi  
Coorientadora  
Dra. Maria Laene de Moreira Carvalho

**LAVRAS - MG  
2013**

**Ficha Catalográfica Elaborada pela Divisão de Processos Técnicos da  
Biblioteca da UFLA**

Leite, Isabel Cristina Costa.

Análise de componentes independentes aplicada a avaliação de  
imagem radiográfica de sementes / Isabel Cristina Costa Leite. –  
Lavras : UFLA, 2013.

123 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2013.

Orientador: Thelma Sáfadi.

Bibliografia.

1. Análise de imagens. 2. Análise discriminante. 3. Raios X. 4.  
Qualidade de sementes. I. Universidade Federal de Lavras. II.  
Título.

CDD – 519.535

**ISABEL CRISTINA COSTA LEITE**

**ANÁLISE DE COMPONENTES INDEPENDENTES APLICADA A  
AVALIAÇÃO DE IMAGEM RADIOGRÁFICA DE SEMENTES**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 13 de junho de 2013.

Dra. Lúcia Pereira Barroso	USP
Dr. Danton Diego Ferreira	UFLA
Dr. Agostinho Roberto de Abreu	UFLA
Dr. Daniel Furtado Ferreira	UFLA

Dra. Thelma Sáfydi  
Orientadora

**LAVRAS - MG  
2013**

A minha eternamente querida mãe, Maria Isabel (*in memoriam*), que em sua  
breve vida não pôde acompanhar os nossos passos e conquistas;  
A vó Guió (*in memoriam*), grande mulher, que nos seus 107 anos bem vividos foi  
e continua sendo minha referência de vida;  
A Julian, companheiro de cada momento.

DEDICO

## AGRADECIMENTOS

À Universidade Federal de Lavras (UFLA) e ao Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), pela realização do Doutorado Interinstitucional, que tornou possível a concretização deste projeto de estudo;

À Capes, pela viabilização de recursos ao DINTER e concessão de bolsa;

Aos professores do Departamento de Ciências Exatas (DEX), particularmente aos diretamente envolvidos com o DINTER, que com competência, disponibilidade e grandeza humana contribuíram para que estes anos de estudo fossem uma experiência enriquecedora e prazerosa;

À professora Thelma, pela confiança e contínua presença, orientando com competência, simplicidade e amizade;

À professora Laene, pela coorientação e pela oportunidade de me fazer conhecer um pouco o ‘mundo’ das sementes, tão distante das experiências vividas até então na área das Ciências Exatas;

Estendo o agradecimento a todos do Laboratório de Análise de Sementes da UFLA que colaboraram nos testes para a obtenção dos dados desta pesquisa, em especial a Nayara, pela disponibilidade e paciência;

Aos professores membros da banca, pelas valiosas contribuições a este trabalho;

Aos colegas e amigos de doutorado, Norma, Azly, Cleide, Edmary, Nelson, Vasquez, Walter, Regilson, Jaime, Jailson, Marcinho e Otaviano (*in memoriam*), por compartilharmos horas de estudo e momentos de confraternizações, que nos fortaleceram nas dificuldades. Em especial a Ângela e Tânia, pelos cuidados, apoio e amizade na convivência diária, no período que passamos em Lavras;

Ao professor Lurimar que, representando o IFBA no projeto DINTER, sempre colaborou com eficiência e presteza;

A todos os colegas da Coordenação de Matemática (IFBA), pelo incentivo e apoio;

À minha família, meu pai Aristóteles e irmãs Rosana e Vanessa, por serem uma base sólida na minha vida, garantia de amor e apoio nos meus projetos de vida;

Ao meu querido Julian, pela contínua presença e pelo amor traduzido em apoio incondicional e concreto em todos os aspectos dessa caminhada;

A todos que direta ou indiretamente contribuíram na realização deste trabalho.

## RESUMO

A análise de imagens radiográficas de sementes é um método efetivo cada vez mais utilizado na avaliação de lotes de sementes. Com este trabalho, propõe-se a utilização da análise de componentes independentes (ICA) e da análise discriminante com o objetivo de classificar imagens radiográficas de sementes em níveis de qualidade física, diferenciando sementes cheias de sementes com algum tipo de dano ou deformação. A ICA foi aplicada a um conjunto de imagens de sementes de girassol gerando uma base de imagens estatisticamente independentes entre si. As coordenadas de cada imagem de semente nesta base são os parâmetros de entrada para a análise discriminante. O método foi testado num conjunto de imagens de sementes geradas por simulação e num conjunto de imagens reais de sementes. A classificação obteve um acerto global de até 97% para as imagens simuladas e 82% para as imagens reais. A partir dos resultados observados na classificação dos radiogramas procurou-se estabelecer relações com o potencial fisiológico das sementes. Os resultados mostraram que a metodologia proposta pode contribuir para uma avaliação rápida e menos subjetiva de imagens radiográficas de sementes.

Palavras-chave: Análise de componentes independentes. Análise de imagens. Análise discriminante. Raios X. Qualidade de sementes.

## ABSTRACT

The analysis of seed radiographic images is an effective and increasingly used method for the evaluation of seed lots. This work proposes the use of the independent component analysis (ICA) and of the discriminant analysis with the objective of classifying seed radiographic images in physical quality levels, differentiating full seeds from deformed or damaged seeds. The ICA was applied to a set of sunflower seed images generating a base of images statistically independent from each other. The coordinates of each seed image in this base are the entry parameters for the discriminant analysis. The method was tested on a set of seed images generated by simulation and on a set of real seed images. The classification presented a global hit of up to 97% for the simulated images and 82% for the real images. From the results observed in the radiogram classifications, we aimed at establishing relations with the physiological potential of the seeds. The results show that the methodology proposed may contribute to a fast and less subjective evaluation of seed radiographic images.

Keywords: Discriminant analysis. Image analysis. Independent component analysis. Seed quality. X-rays.

## LISTA DE FIGURAS

### PRIMEIRA PARTE

Figura 1	Modelo de mistura e de separação da análise de componentes independentes.....	24
Figura 2	Distribuição conjunta de duas variáveis gaussianas.....	28
Figura 3	(A) Mistura de duas variáveis com distribuição uniforme (B) Decomposição das misturas por análise de componentes principais (PCA) e por análise de componentes independentes (ICA).....	32
Figura 4	Funções $G_1$ e $G_2$ aproximações de negentropia e função $G_3$ aproximação de curtose.....	41
Figura 5	Imagem de semente como combinação linear de imagens-base....	67
Figura 6	Imagem de 2 componentes independentes estimados por ICA ....	72
Figura 7	Imagem de 30 componentes independentes estimados por ICA...	72
Figura 6	Reconstrução de imagens de sementes de girassol com diferentes níveis de qualidade física, a partir de bases de componentes independentes de diferentes dimensões. (A) imagem original; (B) reconstruída com 2 IC's; (C) reconstruída com 20 IC's; (D) reconstruída com 30 IC's; (E) reconstruída com 45 IC's; (F) reconstruída com 60 IC's .....	73

### SEGUNDA PARTE – ARTIGOS

#### ARTIGO 1

Figura 1	Imagens radiográficas de sementes de girassol cultivar Hélio-250.....	90
Figura 2	Imagens radiográficas de sementes de diferentes níveis de	

	qualidade: (a) cheia; (b) com dano leve, mas com preservação do eixo embrionário; (c) com dano grave e (d) deformada, com danos afetando o eixo embrionário.....	91
Figura 3	Imagem de semente decomposta em uma combinação linear de $k$ imagens-base.....	93
Figura 4	Imagem de 60 componentes independentes estimados por ICA ..	97
Figura 5	Imagens radiográficas originais de sementes de girassol com diferentes níveis de qualidade física (primeira coluna) e imagens simuladas, a partir de bases de componentes independentes de diferentes dimensões ( $k$ ) com variância no erro aleatório de $10^{-6}$ (colunas A) e $3,6 \cdot 10^{-7}$ (colunas B).....	98
Figura 6	Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância $10^{-6}$ a partir de amostras escolhidas pelo analista .....	99
Figura 7	Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância $3,6 \cdot 10^{-7}$ a partir de amostras escolhidas pelo analista .....	100
Figura 8	Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância $10^{-6}$ a partir de amostras escolhidas aleatoriamente.....	101
Figura 9	Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância $3,6 \cdot 10^{-7}$ a partir de amostras escolhidas aleatoriamente.....	101

## ARTIGO 2

Figure 1	Radiographic images of sunflower seeds with different quality levels: A, full; B, slightly injured, although preserving
----------	---

	characteristics of the embryonic axis; and C, deformed, with severe injury affecting the embryonic axis.....	110
Figure 2	Radiographic images of sunflower seeds. (a) Image of seed decomposed into a linear combination of $k$ basis-images. (b) Image of 30 independent components estimated by ICA.....	112
Figure 3	Proportion of sunflower seeds correctly classified depending on the number of estimated independent components (ICs) from a new sample composition.....	115
Figure 4	Proportion of sunflower seeds correctly classified depending on the number of estimated independent components (ICs) based on the new visual classification of images.....	116
Figure 5	Proportion of sunflower seeds misclassified in each group using quadratic discriminant function for different data sizes ( $k$ ).....	117
Figure 6	Germination proportions of 445 sunflower seeds separated by levels of physical quality.....	118
Figure 7	Germination results of misclassified sunflower seeds for data size $k = 30$ .....	118

## LISTA DE TABELAS

### PRIMEIRA PARTE

Tabela 1	Distribuição dos indivíduos segundo as populações de origem e classificação, em que $n_{il}$ , $i \neq l$ , representa o número de elementos de $\Pi_i$ classificados incorretamente em $\Pi_l$ e $n_{ii}$ o número de elementos classificados corretamente em $\Pi_i$ .....	57
Tabela 2	Percentual de variabilidade explicada a partir do número $k$ de componentes independentes (IC's) estimados.....	71
Tabela 3	Probabilidade global de acerto de classificação das funções discriminantes de Fisher e quadrática, usando dados de diferentes dimensões.....	74
Tabela 4	Quantidade de sementes classificadas em cada categoria usando a função discriminante quadrática para dados de dimensão 20.....	75
Tabela 5	Quantidade de sementes classificadas em cada categoria usando a função discriminante quadrática para dados de dimensão 30.....	75

### SEGUNDA PARTE – ARTIGOS

#### ARTIGO 2

Table 1	Results of seed classification by quadratic discriminant function for data size $k = 30$ .....	106
---------	--	-----

## SUMÁRIO

	<b>PRIMEIRA PARTE</b>	
<b>1</b>	<b>INTRODUÇÃO</b> .....	15
<b>1.1</b>	<b>Objetivos</b> .....	17
<b>1.2</b>	<b>Revisão de literatura</b> .....	18
<b>1.3</b>	<b>Organização do trabalho</b> .....	21
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> .....	23
<b>2.1</b>	<b>Análise de componentes independentes</b> .....	23
<b>2.1.1</b>	<b>Definição</b> .....	23
<b>2.1.2</b>	<b>Pressuposições</b> .....	25
<b>2.1.2.1</b>	<b>Independência estatística</b> .....	25
<b>2.1.2.2</b>	<b>Variáveis não gaussianas</b> .....	26
<b>2.1.3</b>	<b>Ambiguidades</b> .....	28
<b>2.1.4</b>	<b>Pré-processamento para ICA</b> .....	29
<b>2.1.4.1</b>	<b>Centralização</b> .....	29
<b>2.1.4.2</b>	<b>Branqueamento</b> .....	30
<b>2.1.5</b>	<b>Exemplo de ICA</b> .....	31
<b>2.1.6</b>	<b>Estimação ICA pela maximização da não gaussianidade</b> .....	33
<b>2.1.6.1</b>	<b>Curtose</b> .....	35
<b>2.1.6.2</b>	<b>Algoritmo de ponto fixo usando curtose</b> .....	36
<b>2.1.6.3</b>	<b>Negentropia</b> .....	38
<b>2.1.6.4</b>	<b>Aproximações da negentropia</b> .....	39
<b>2.1.6.5</b>	<b>Algoritmo de ponto fixo usando negentropia</b> .....	41
<b>2.1.6.6</b>	<b>Algoritmo FastICA</b> .....	43
<b>2.1.6.7</b>	<b>Algoritmo FastICA para vários componentes independentes</b> ..	45
<b>2.1.7</b>	<b>Estimação ICA pela maximização da verossimilhança</b> .....	47
<b>2.1.8</b>	<b>Outros princípios de estimação ICA</b> .....	50
<b>2.2</b>	<b>Análise discriminante</b> .....	51
<b>2.2.1</b>	<b>Regras de classificação</b> .....	52
<b>2.2.2</b>	<b>Estimação de classificações incorretas</b> .....	56
<b>2.2.2.1</b>	<b>Método da ressubstituição</b> .....	58
<b>2.2.2.2</b>	<b>Método de Lachenbruch</b> .....	59
<b>2.3</b>	<b>Análise de sementes</b> .....	59
<b>2.3.1</b>	<b>Teste de raios X</b> .....	61
<b>2.3.2</b>	<b>Teste de germinação</b> .....	63
<b>3</b>	<b>MATERIAL E MÉTODOS</b> .....	66
<b>3.1</b>	<b>Obtenção e processamento de dados</b> .....	66
<b>3.2</b>	<b>Aplicação de ICA na extração de características</b> .....	67
<b>3.3</b>	<b>Classificação por análise discriminante</b> .....	69

4	<b>RESULTADOS E DISCUSSÃO</b> .....	71
5	<b>CONSIDERAÇÕES GERAIS</b> .....	77
	<b>REFERÊNCIAS</b> .....	78
	<b>SEGUNDA PARTE – ARTIGOS</b> .....	85
	<b>ARTIGO 1 Análise de componentes independentes na avaliação de imagens radiográficas de sementes: um estudo de simulação</b> .....	85
1	<b>INTRODUÇÃO</b> .....	87
2	<b>MATERIAL E MÉTODOS</b> .....	89
2.1	<b>Obtenção e processamento de dados</b> .....	89
2.2	<b>Análise de componentes independentes</b> .....	91
2.3	<b>Aplicação de ICA ao conjunto de dados</b> .....	93
2.4	<b>Análise discriminante</b> .....	94
2.5	<b>Simulação de amostras de testes</b> .....	95
3	<b>RESULTADOS E DISCUSSÃO</b> .....	96
4	<b>CONCLUSÃO</b> .....	102
	<b>REFERÊNCIAS</b> .....	103
	<b>ARTIGO 2 Evaluation of seed radiographic images by independent component analysis and discriminant analysis</b> ....	106
	<b>CONSIDERAÇÕES FINAIS</b> .....	122

## **PRIMEIRA PARTE**

### **1 INTRODUÇÃO**

A análise de sementes é de fundamental importância na determinação do valor de lotes de sementes para a semeadura. A qualidade é avaliada a partir dos atributos físicos e fisiológicos da semente relacionados à capacidade desta desempenhar suas funções vitais, ser portadora de vida e produzir uma plântula normal. Por este motivo a maioria dos testes para avaliação da qualidade de sementes examina o nível de deterioração, a capacidade germinativa e o vigor das mesmas.

A análise de imagens radiográficas de sementes tem se apresentado como uma alternativa aos testes tradicionais realizados em laboratório, pois possui a vantagem de ser uma forma de avaliação rápida e não destrutiva. Consiste em analisar a imagem radiográfica da semente, diferenciando sementes bem formadas das sementes vazias, com danos mecânicos ou com ataque de insetos. As sementes analisadas podem posteriormente ser utilizadas na semeadura, permitindo estabelecer relações entre os resultados observados nos radiogramas e o potencial físico e fisiológico da semente (KOBORI; CICERO; MEDINA, 2012). Desta forma é possível avaliar também a eficiência do teste.

Geralmente a imagem radiográfica é examinada visualmente por um analista de sementes e por este motivo os resultados estão sujeitos à subjetividade do analista na interpretação das imagens. Essa subjetividade pode ser minimizada com o uso de técnicas de processamento automático de imagens, nas quais as imagens radiográficas das sementes são analisadas por softwares.

O processamento de imagens é um estágio necessário para novos processamentos de dados tais como aprendizagem de máquina ou reconhecimento de

padrões. Construção de histogramas, aplicação de filtros, detecção de borda, uso da transformada de Fourier são alguns dos inúmeros recursos usados no processamento de imagens.

Desenvolvida inicialmente como uma técnica de processamento de sinais, a análise de componentes independentes (ICA - *Independent Component Analysis*) é um método estatístico e computacional, cujo objetivo é encontrar uma representação linear de dados não-gaussianos, de modo que os componentes sejam estatisticamente independentes ou tenham dependência estatística minimizada. É muitas vezes referida como técnica de separação cega de fontes (BSS - *Blind Source Separation*) que consiste na extração de sinais de informação de dados que se encontram misturados e/ou corrompidos. É aplicada no processamento de imagens, sinais de audio, eletromagnéticos, sensores químicos, redes de telecomunicações, satélites, sinais biomédicos obtidos por eletroencefalogramas (EEG), ecocardiogramas e outros. Os dados analisados pela ICA podem ser oriundos de várias áreas de aplicação incluindo base de dados de documentos, bem como indicadores econômicos e medidas psicométricas.

A ICA também é largamente aplicada na extração de características, que é uma forma especial de redução da dimensão de um conjunto de dados observado. Se as características extraídas forem cuidadosamente escolhidas espera-se que esse conjunto represente a parte relevante da informação para se executar a tarefa desejada ao invés de se usar os dados de entrada na íntegra. A seleção de características é uma etapa essencial que precede a análise de dados e muitas vezes é utilizada com o objetivo de reduzir o esforço computacional e alcançar melhor desempenho na classificação dos objetos em estudo.

Entre as diversas técnicas de classificação utilizadas na análise multivariada de dados, destacam-se a análise discriminante, a análise de agrupamentos, o

método do vizinho mais próximo e a sua variação para os  $k$  vizinhos mais próximos (kNN - *k Nearest Neighbor*), redes neurais artificiais, máquina de vetor de suporte, árvores de classificação e regressão (método CART - *Classification and Regression Trees*) e classificadores Bayesianos.

Usando a abordagem de extração de características e classificação, o presente estudo visa aplicar a análise de componentes independentes no processamento de imagens radiográficas de sementes e a análise discriminante para classificar estas imagens, propondo uma nova metodologia para avaliação da qualidade de sementes.

## 1.1 Objetivos

O objetivo geral com este trabalho é propor um novo método de avaliação de imagens radiográficas de sementes, aplicando a análise de componentes independentes na extração de características destas imagens. Dentre os objetivos específicos destacam-se:

- Automatizar e diminuir a imprecisão da análise de imagens radiográficas de sementes realizada visualmente.
- Classificar as sementes quanto à qualidade física, usando a análise discriminante como técnica de classificação. As características obtidas por ICA são os parâmetros de entrada do classificador.
- Avaliar o potencial germinativo das sementes associando os resultados da classificação realizada com resultados da germinação destas mesmas sementes.

## 1.2 Revisão de literatura

A análise de componentes independentes teve os seus trabalhos seminais formulados por Herault e Jutten (1986) e Herault, Jutten e Ans (1985) na década de 80. São estudos na área de neurofisiologia relacionados à codificação empregada pelo sistema nervoso central para a ativação muscular. Eles propuseram um modelo e um algoritmo com o objetivo de revelar separadamente informações de posição e velocidade angular do movimento a partir da observação de sinais sensoriais de contração muscular. A técnica foi denominada análise de componentes independentes pelas suas similaridades com a análise de componentes principais (PCA - *Principal Component Analysis*), considerando que a diferença fundamental entre as duas técnicas é que a PCA obtém componentes não correlacionados, enquanto que a ICA deseja obter componentes estatisticamente independentes. Inicialmente o termo ICA estava intimamente associado à separação cega de fontes (BSS - *Blind Source Separation*), devido à sua principal aplicabilidade na separação de sinais misturados, cujas fontes são desconhecidas.

Neste período a divulgação da nova técnica ficou restrita a pesquisadores franceses. Em 1989, com o primeiro workshop internacional, estudos foram divulgados na área de análise espectral de ordens superiores com trabalhos em ICA de Cardoso (1989) e Comon (1994). Cardoso (1989) usou métodos algébricos, especialmente com tensores cumulantes de quarta ordem que conduziram à criação do algoritmo JADE (*Joint Approximate Diagonalization of Eigenmatrices*) (CARDOSO; SOULOUMIAC, 1993). No trabalho de Comon (1994) se delineou uma estrutura matemática mais bem definida e aplicável para a ICA, demonstrando como a independência estatística se insere no problema de separação de fontes. Essa contribuição teve papel fundamental no desenvolvimento de novos métodos de BSS.

Nos anos 90, Bell e Sejnowski (1995) e Cichocki e Unbenhauen (1996) propuseram os algoritmos ICA mais populares. O algoritmo de Bell e Sejnowski (1995) é baseado no princípio Infomax (Information Maximization), princípio desenvolvido por Linsker ainda nos anos 80, que formulou a idéia de máxima preservação de informação numa rede neural (HAYKIN, 1999). Este algoritmo despertou grande interesse na comunidade científica internacional, conduzindo a posteriores reformulações ou refinamentos, como o proposto por Amari, Cichocki e Yang (1996) com o uso do gradiente natural conectado ao princípio de estimação de máxima verossimilhança. Em 1996 publicou-se o primeiro artigo científico dedicado à aplicação do algoritmo de Bell e Sejnowski em sinais eletroencefalográficos (MAKEIG et al., 1996).

Um grupo de cientistas finlandeses trouxe grandes contribuições no desenvolvimento da ICA. Estudos de Karhunen, Pajunen e Oja (1998) permitiram interpretar a ICA como uma extensão não-linear da análise de componentes principais. Tal abordagem teve um papel fundamental no entendimento da ICA como um tema relevante em análise de dados multivariados. Hyvärinen contribuiu para o desenvolvimento de critérios baseados na maximização da não-gaussianidade, nos quais se baseia o algoritmo FastICA (*Fast Independent Component Analysis*) (HYVÄRINEN; OJA, 2000).

Com os sucessivos desenvolvimentos da técnica ICA as suas aplicações foram se diversificando em inúmeras áreas e aos poucos se dissociando da BSS. Em Econometria, trabalhos aplicando ICA na previsão de séries temporais financeiras podem ser lidos em Back e Weigend (1997), García-Ferrer, González-Prieto e Peña (2008), Kiviluoto e Oja (1998), Lu (2010) e Lu, Lee e Chiu (2009).

Trabalhos que utilizam ICA na extração de características na maioria das vezes a associam à aplicação de alguma técnica de classificação dos dados, como

nos estudos citados a seguir.

A utilização de ICA em imagens iniciou-se com trabalhos de decomposição de imagens naturais em uma base de imagens independentes entre si, visando aplicações em redução de ruído, compressão, extração de características, entre outros domínios do processamento de imagens (HYVÄRINEN; HOYER, 2000).

Em trabalho de reconhecimento de faces (DÉNIZ; CASTRILLÓN; HERNÁNDEZ, 2003) comparou-se a utilização de ICA ou PCA combinados com o classificador máquina de vetor de suporte (SVM - *Support Vector Machine*). A combinação ICA/SVM apresentou melhor desempenho, apesar de PCA/SVM também apresentar bons resultados.

Muitas aplicações utilizando imagens encontram-se na área da engenharia biomédica. Em Prasad, Sowmya e Koch (2004) foram realizados estudos para a classificação de enfisemas em imagens de tomografias computadorizadas de alta resolução, usando ICA na seleção de um subconjunto de características a serem usadas por três diferentes classificadores: Naive Bayes, Seeded K-Means e C4.5. A classificação realizada utilizando Naive Bayes foi a que obteve melhores resultados. Utilizando imagens provenientes de raios X, Christoyianni et al. (2002) foram os primeiros a utilizarem ICA na extração de características de mamogramas, usando previamente PCA na redução de dimensão dos dados, e uma rede neural artificial para a classificação dos tecidos da mama entre normais e anormais, benignos e malignos. Compararam o desempenho de ICA com outros dois métodos de análise de imagens (Gray level histograms moments e Spacial gray level dependence matrix) e a ICA obteve melhor performance. Nesta mesma linha de pesquisa Campos, Barros e Silva (2007) também utilizaram ICA e uma rede neural multicamadas perceptron para a classificação de regiões de mamogramas. A metodologia obteve sucesso de 97,8%. Em Costa et al. (2007), utilizando

ICA como extrator de características, foi comparada a eficiência dos classificadores análise discriminante linear (LDA - *Linear Discriminant Analysis*) e SVM para distinguir imagens de mamogramas entre nódulos ou não nódulos e os tecidos lesionados entre benignos ou malignos. Obteve-se até 99,6% de acurácia na classificação com estes métodos, sendo os resultados do uso de SVM superiores aos obtidos com LDA. Tang, Wang e Chen (2012) usaram ICA para extrair vasos coronarianos das angiografias digitais das coronárias com melhor qualidade que os extraídos com as tradicionais técnicas de subtração.

Muito pouco se conhece da aplicação de ICA na área agrícola. Zhu et al. (2007) propuseram um método usando ICA na seleção de comprimentos de onda ótimos em imagens de fluorescência hiperespectral para discriminação entre fragmentos de casca ou fruto de nozes. O método dos  $k$  vizinhos mais próximos (kNN) foi utilizado como técnica de classificação. Até o presente momento não foram encontrados estudos voltados à avaliação de sementes. A proposta de uma aplicação de ICA em análise de imagens radiográficas de sementes daria uma contribuição significativa para o avanço dos métodos de avaliação da qualidade dos lotes de sementes.

### **1.3 Organização do trabalho**

O texto está organizado em formato de coletânea de artigos sobre aplicação de ICA na avaliação de imagens radiográficas de sementes, conforme as normas da Universidade Federal de Lavras - UFLA (2010).

Nesta primeira parte são apresentados motivação, contextualização, objetivos do trabalho, revisão de literatura e referencial teórico. Alguns dos primeiros resultados da aplicação de ICA e da análise discriminante na classificação de imagens radiográficas de sementes também são apresentados como base para os

estudos posteriores desenvolvidos nos artigos.

A segunda parte é constituída de dois artigos. Objetivou-se no primeiro artigo validar o uso de ICA na extração de características de imagens de sementes utilizando simulação. Simulando novas imagens a partir das características obtidas com a aplicação de ICA, avaliou-se a classificação destas imagens realizada por análise discriminante.

No artigo 2 foram avaliadas as imagens de raios X de um novo conjunto de sementes. Os resultados obtidos da aplicação de ICA a um conjunto de 300 imagens serviu como base de dados para a classificação deste novo conjunto. Procurou-se estabelecer uma correlação entre os resultados obtidos com a classificação quanto ao nível de qualidade física da semente e os resultados do teste de germinação.

Nas considerações finais são resumidos os aspectos principais deste trabalho e apresentadas perspectivas de trabalhos futuros.

## 2 REFERENCIAL TEÓRICO

### 2.1 Análise de componentes independentes

A análise de componentes independentes é uma técnica estatística e computacional que revela componentes ou fatores subjacentes a um conjunto de variáveis aleatórias, medições, ou sinais observados multivariados. Difere das técnicas tradicionais utilizadas nesta área, porque tais componentes são estatisticamente independentes (ou têm dependência estatística minimizada) e são não gaussianos (HYVÄRINEN; KARHUNEN; OJA, 2001).

#### 2.1.1 Definição

Seja o vetor aleatório  $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$ , cujos  $n$  elementos são gerados pela mistura de  $n$  componentes estatisticamente independentes entre si de um vetor aleatório  $\mathbf{S} = [S_1, S_2, \dots, S_n]^T$ . O modelo ICA expressa cada  $X_i$  como uma combinação linear de componentes independentes, dada por

$$X_i = a_{i1}S_1 + a_{i2}S_2 + \dots + a_{in}S_n, \text{ para todo } i = 1, 2, \dots, n, \quad (2.1)$$

em que  $a_{ij}, j = 1, 2, \dots, n$ , são coeficientes reais.

Usando notação matricial o modelo pode ser escrito como

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}, \quad (2.2)$$

em que  $\mathbf{A}$  é a matriz dos coeficientes  $a_{ij}$  das combinações lineares.

Sendo  $a_{ij}$  um coeficiente que pondera a mistura dos componentes independentes (sinais ou fontes originais), a matriz  $\mathbf{A}$  é denominada matriz de mistura.

Tanto os coeficientes  $a_{ij}$  como os componentes independentes  $S_i$  são desconhecidos e devem ser estimados a partir da observação dos sinais misturados  $X_i$ .

Este é um modelo generativo, pois descreve como os dados observados são gerados a partir de um processo de mistura dos componentes  $S_i$ .

Alternativamente pode-se definir ICA como o problema de determinar uma transformação linear dada pela matriz  $\mathbf{W}$ ,

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}, \quad (2.3)$$

em que  $\mathbf{Y}$  é o vetor aleatório de componentes  $Y_1, Y_2, \dots, Y_n$  que são estimativas dos componentes independentes e  $\mathbf{W}$  é a matriz inversa de  $\mathbf{A}$ , denominada matriz de separação.

A equação (2.3) evidencia o objetivo do método que é estimar os componentes independentes. Na Figura 1 é descrito o caminho de estimação a partir da análise de componentes independentes.

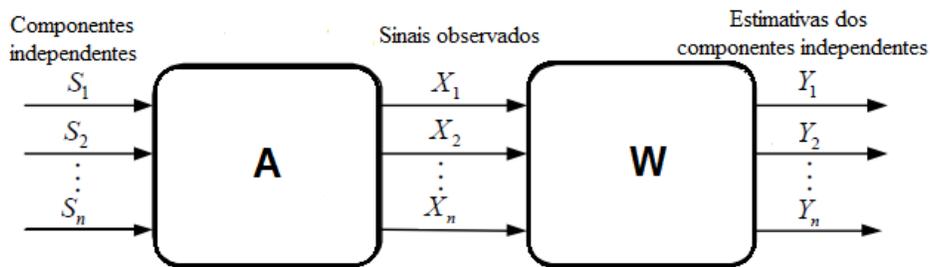


Figura 1 Modelo de mistura e de separação da análise de componentes independentes

Segundo Hyvärinen, Karhunen e Oja (2001), esta definição de modelo é uma das mais básicas pois omite a existência de ruído e considera a mistura linear. Estas considerações podem ser ampliadas definindo modelos ICA mais comple-

xos. Também assume-se que o número de componentes independentes seja igual ao número de sinais observados com o objetivo de se ter uma matriz de mistura quadrada e não singular, simplificando a estimação da matriz  $\mathbf{W}$  e dos componentes independentes obtidos em (2.3). Nos casos em que o número de sinais observados é maior que o número de componentes independentes a estimação pode ser realizada com o uso da inversa generalizada (pseudoinversa) da matriz de mistura, considerando-se que  $\mathbf{A}$  seja uma matriz de posto coluna completo.

### 2.1.2 Pressuposições

Para que o modelo ICA esteja bem definido é necessário supor que os componentes  $S_i$  a serem estimados sejam estatisticamente independentes entre si e apresentem distribuição de probabilidade não-gaussiana.

#### 2.1.2.1 Independência estatística

Duas variáveis aleatórias  $Y_i$  e  $Y_j$  são ditas *independentes* se a informação contida na variável  $Y_i$  não fornece nenhuma informação sobre a probabilidade de ocorrência da variável  $Y_j$ , para  $i \neq j$ .

Em Hyvärinen (1999) encontram-se as seguintes definições e considerações a respeito da condição de independência estatística exigida pelo modelo ICA.

Sejam  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias com função de densidade conjunta  $f(y_1, y_2, \dots, y_n)$ . Diz-se que as variáveis  $Y_i$  são estatisticamente independentes se a função de densidade conjunta pode ser fatorada na forma

$$f(y_1, y_2, \dots, y_n) = f_1(y_1)f_2(y_2) \dots f_n(y_n), \quad (2.4)$$

em que  $f_i(y_i)$  denota a densidade marginal de  $Y_i$ , para todo  $i = 1, 2, \dots, n$ .

A partir do conceito de esperança pode-se demonstrar que

$$E[g(Y_i)h(Y_j)] = E[g(Y_i)] \cdot E[h(Y_j)], \quad \text{para } i \neq j, \quad (2.5)$$

em que  $g(Y_i)$  e  $h(Y_j)$  são quaisquer funções integráveis de  $Y_i$  e  $Y_j$ .

Por outro lado, diz-se que duas variáveis  $Y_i$  e  $Y_j$  são *não correlacionadas* quando a covariância entre elas é nula

$$\text{Cov}(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i) \cdot E(Y_j) = 0, \quad (2.6)$$

o que equivale a

$$E(Y_i Y_j) = E(Y_i) \cdot E(Y_j), \quad \text{para } i \neq j. \quad (2.7)$$

Para variáveis aleatórias de média nula, com as quais se trabalha mais frequentemente, tem-se que

$$E(Y_i Y_j) = 0 \quad (2.8)$$

é também uma condição de ortogonalidade.

Comparando as equações (2.5) e (2.7) conclui-se que independência estatística é uma propriedade muito mais forte do que não correlação. De fato, a equação (2.7) que define não correlação pode ser vista como um caso particular da propriedade de independência (2.5), em que  $g(Y_i)$  e  $h(Y_j)$  são funções lineares.

### 2.1.2.2 Variáveis não gaussianas

A restrição de variáveis gaussianas para os componentes independentes é uma condição fundamental no modelo ICA, como será observado a seguir.

Considerando que  $S_1$  e  $S_2$  sejam dois componentes independentes gaus-

sianos com média zero, com variância unitária e descorrelacionados, tem-se que a função de densidade de probabilidade conjunta é dada por

$$f(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{s_1^2 + s_2^2}{2}\right) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right). \quad (2.9)$$

Assumindo que a matriz  $\mathbf{A}$  seja ortogonal (como será exposto na seção 2.1.4), tem-se que  $\mathbf{A}^{-1} = \mathbf{A}^T$ . Usando a fórmula clássica do método jacobiano para transformação de funções de densidade de probabilidade (*fdp*) tem-se que a densidade conjunta das misturas  $X_1$  e  $X_2$ , obtidas pela transformação (2.2), é dada por

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{A}^T \cdot \mathbf{x}\|^2}{2} |\det \mathbf{A}^T|\right). \quad (2.10)$$

Se  $\mathbf{A}$  é ortogonal, então  $\mathbf{A}^T$  também é ortogonal e segue-se que  $\|\mathbf{A}^T \cdot \mathbf{x}\|^2 = \|\mathbf{x}\|^2$  e  $|\det \mathbf{A}| = 1$ .

Logo,

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right), \quad (2.11)$$

donde se conclui que uma matriz de mistura ortogonal não altera a *fdp* conjunta de misturas gaussianas, pois sequer aparece na *fdp* da mistura.

Ao observar o gráfico da distribuição conjunta de duas fontes gaussianas (Figura 2), verifica-se que esta densidade é completamente simétrica e não se modifica quando submetida a uma rotação. A *fdp* e a sua representação gráfica não contêm nenhuma informação sobre as direções das colunas da matriz de mistura, explicando porque não há como inferir a matriz  $\mathbf{A}$  para mistura provenientes de variáveis gaussianas.

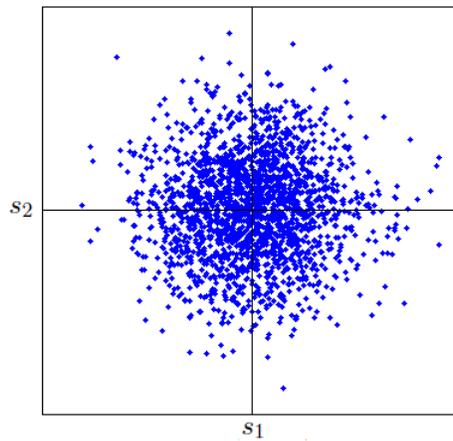


Figura 2 Distribuição conjunta de duas variáveis gaussianas

Variáveis gaussianas possuem a particularidade de que as propriedades de não correlação e independência são equivalentes. Estas estão completamente definidas a partir da média e covariância. Seus cumulantes de ordem mais alta são nulos, mas as informações contidas nessas medidas são essenciais na estimativa do modelo ICA. Apesar desta restrição para variáveis gaussianas, é importante notar que o modelo não exige qualquer prévio conhecimento sobre as distribuições de probabilidade dos componentes independentes.

### 2.1.3 Ambiguidades

A partir da definição do modelo ICA (2.2) é fácil identificar algumas indeterminações que persistem.

- **Não é possível determinar a variância dos componentes independentes**

Isto se deve ao fato de que tanto  $S$  como  $A$  sendo desconhecidos, qualquer escalar  $\alpha_i$  multiplicado a algum dos sinais originais  $S_i$  pode ser sempre cancelado

dividindo-se pelo mesmo  $\alpha_i$  a correspondente coluna  $\mathbf{a}_i$  de  $\mathbf{A}$ :

$$\mathbf{X} = \sum_{i=1}^n \left( \frac{1}{\alpha_i} \mathbf{a}_i \right) (S_i \alpha_i). \quad (2.12)$$

Este problema pode ser contornado assumindo-se que os componentes independentes tenham variância unitária,  $E(S_i^2) = 1$ . Porém permanece a ambiguidade do sinal, pois sempre se pode multiplicar um componente independente por -1 sem afetar o modelo. Essa ambiguidade é irrelevante na maioria das aplicações.

- **Não é possível determinar a ordem dos componentes independentes**

Aplicando uma matriz de permutação  $\mathbf{P}$  e a sua inversa ao modelo, este resulta em

$$\mathbf{X} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{S}, \quad (2.13)$$

em que os elementos de  $\mathbf{P}\mathbf{S}$  são componentes independentes em outra ordem, assim como  $\mathbf{A}\mathbf{P}^{-1}$  é uma outra matriz de mistura a ser estimada.

#### 2.1.4 Pré-processamento para ICA

Com o objetivo de tornar mais simples a estimação dos componentes independentes, os dados observados são submetidos a alguns procedimentos antes da aplicação de algum tipo de algoritmo ICA.

##### 2.1.4.1 Centralização

Um procedimento básico e necessário como pré-processamento para a ICA é subtrair a média dos dados observados, ou seja, tornar os dados com média nula. Este processo é chamado de centralização.

Seja  $\tilde{X}$  o vetor aleatório das variáveis observadas. O vetor submetido ao algoritmo ICA será

$$X = \tilde{X} - E(\tilde{X}). \quad (2.14)$$

Consequentemente a média dos componentes independentes estimados será também nula, como segue-se da equação (2.3)

$$E(Y) = W \cdot E(X). \quad (2.15)$$

Este processamento não afeta a estimação da matriz de mistura que permanece a mesma independentemente da média dos dados observados e dos componentes independentes.

Ao fim da estimação os componentes independentes dos dados observados originais deverão ter a média restituída como se segue

$$\tilde{Y} = Y + W \cdot E(\tilde{X}). \quad (2.16)$$

Sendo a distribuição de probabilidade do vetor aleatório geralmente não conhecida, na prática a esperança é computada pela média amostral das realizações dos vetores aleatórios.

#### 2.1.4.2 Branqueamento

Branqueamento é o processo de tornar as variáveis observadas não correlacionadas e com variância unitária, ou seja, a matriz de covariância dos dados branqueados é a matriz identidade. O processo será aplicado ao vetor  $X$  dos dados observados previamente centralizado.

O vetor de dados branqueados  $Z$  é obtido a partir de uma decomposição

ortogonal do vetor  $\mathbf{X}$

$$\mathbf{Z} = \mathbf{V} \cdot \mathbf{X}, \quad (2.17)$$

em que

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T, \quad (2.18)$$

sendo  $\mathbf{E}$  e  $\mathbf{D}$ , respectivamente, a matriz ortogonal dos autovetores e a matriz diagonal dos autovalores da matriz de covariância amostral de  $\mathbf{X}$ .

Convém observar que o modelo ICA aplicado aos dados branqueados é muito mais simples, pois segue-se das equações (2.2) e (2.17) que

$$\begin{aligned} \mathbf{Z} &= \mathbf{V} \cdot \mathbf{A} \cdot \mathbf{S}, \\ \mathbf{Z} &= \tilde{\mathbf{A}} \cdot \mathbf{S}. \end{aligned} \quad (2.19)$$

A busca pelos componentes independentes passa a ser feita a partir do vetor  $\mathbf{Z}$ . A nova matriz de mistura  $\tilde{\mathbf{A}}$  é uma matriz ortogonal, o que reduz a quantidade de parâmetros a serem estimados de  $n^2$  a  $n(n-1)/2$ . Por este motivo pode-se dizer que branqueamento resolve metade do problemas de ICA (HYVÄRINEN; OJA, 2000) e os algoritmos ICA aplicados ao vetor  $\mathbf{Z}$  terão melhor desempenho.

Ao ser feita a decorrelação dos dados pode ser útil aproveitar também para reduzir a dimensão do problema. Como usualmente é feito na técnica de componentes principais, seleciona-se na matriz  $\mathbf{E}$  os autovalores que proporcionalmente mais explicam da variabilidade total dos dados e descarta-se os menores.

### 2.1.5 Exemplo de ICA

Na Figura 3A é apresentado um exemplo de como a transformação linear age sobre a distribuição conjunta de duas variáveis com distribuição uniforme. O

paralelogramo representa as variáveis misturadas pela transformação na direção dos vetores  $\mathbf{a}_1$  e  $\mathbf{a}_2$ .

Na Figura 3B são observadas as duas misturas decompostas em variáveis branqueadas após aplicação de PCA. Estas variáveis ainda não foram separadas, não são os componentes independentes. O passo seguinte da ICA será o de estimar a transformação ortogonal que será feita após a decorrelação.

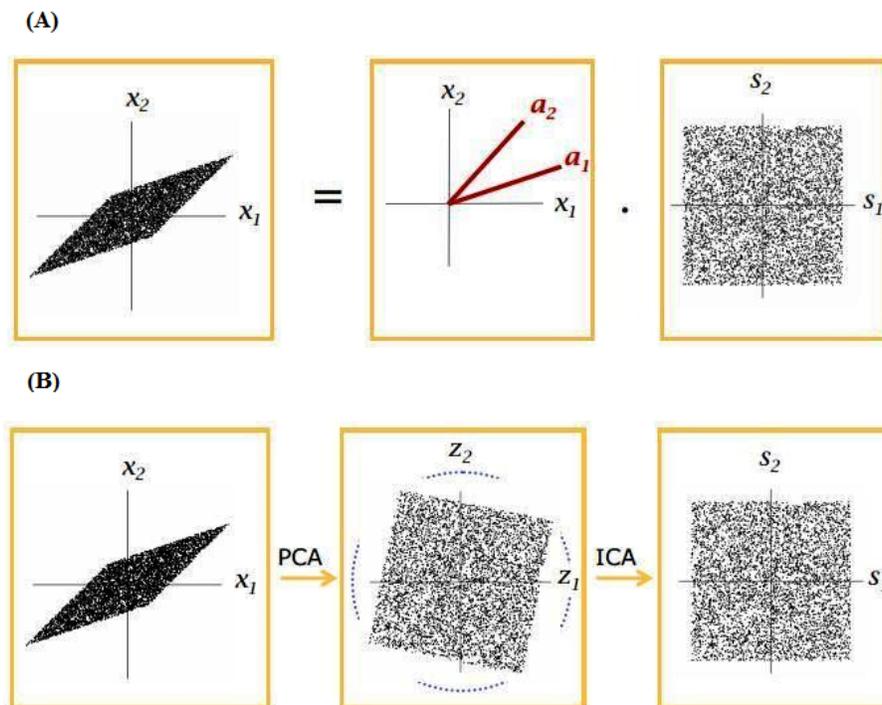


Figura 3 (A) Mistura de duas variáveis com distribuição uniforme (B) Decomposição das misturas por análise de componentes principais (PCA) e por análise de componentes independentes (ICA)

Fonte: Journée (2008)

### 2.1.6 Estimação ICA pela maximização da não gaussianidade

Os vários métodos de estimação dos componentes independentes se baseiam na formulação de uma função objetivo que será maximizada ou minimizada por algoritmos de otimização. Tais funções estão relacionadas com a independência das variáveis e diferem entre si a partir do princípio no qual se baseiam.

A motivação para o uso de medidas de não gaussianidade na estimação de ICA é o Teorema do Limite Central (TLC) (MAGALHÃES, 2006). Este afirma que uma soma de variáveis aleatórias independentes e identicamente distribuídas tende a possuir uma distribuição gaussiana, sob certas condições.

Portanto, adotando o modelo generativo que define o modelo ICA (2.2) e, por simplicidade, assumindo que todos os componentes independentes  $S_i$  sejam identicamente distribuídos, pelo TLC pode-se afirmar que os dados observados no vetor  $X$  têm distribuição mais próxima da distribuição gaussiana que qualquer uma das variáveis aleatórias originais componentes do vetor  $S$ .

A busca pelos componentes independentes se faz a partir da estimação da transformação linear do modelo ICA na sua forma invertida (2.3). Para facilitar a notação, seja  $Y$  simplesmente algum dos componentes  $Y_i$  do vetor  $Y$  dado pela combinação linear dos  $X_i$ ,

$$Y = \mathbf{w}^T \cdot \mathbf{X}. \quad (2.20)$$

Todo o problema consiste na determinação de  $\mathbf{w}$ , um vetor apropriado que reconstrói o componente independente.

Substituindo (2.2) em (2.20), tem-se que

$$Y = \mathbf{w}^T \cdot \mathbf{A} \cdot \mathbf{S}. \quad (2.21)$$

Seja  $\mathbf{q}^T = \mathbf{w}^T \cdot \mathbf{A}$ . Note-se que se  $\mathbf{w}^T$  for uma das linhas da inversa de  $\mathbf{A}$ ,  $\mathbf{q}$  será um vetor com um elemento 1 e os outros nulos e  $Y$  corresponderá a um dos componentes independentes. Na prática não é possível determinar  $\mathbf{w}$  exatamente desta maneira, pois não se tem nenhuma informação sobre  $\mathbf{A}$ , mas pode-se chegar a um estimador para este vetor com boa aproximação.

Sendo

$$Y = \mathbf{q}^T \cdot \mathbf{S}, \quad (2.22)$$

o TLC garante que  $Y$  é mais gaussiano que qualquer  $S_i$  e se torna menos gaussiano quando se iguala a algum  $S_i$ . Portanto, teoricamente, espera-se que variando os coeficientes em  $\mathbf{q}$  e observando como a distribuição de  $Y$  muda, possa-se chegar ao componente independente.

Efetivamente esta busca não é feita sobre o vetor  $\mathbf{q}$ , mas sobre o vetor  $\mathbf{w}$ , uma vez que de (2.20) e (2.22), segue-se que  $\mathbf{q}^T \cdot \mathbf{S} = \mathbf{w}^T \cdot \mathbf{X}$ . No caso, variando os coeficientes de  $\mathbf{w}$  e olhando a distribuição de  $\mathbf{w}^T \cdot \mathbf{X}$ , procura-se que este vetor seja o mais distante possível de uma distribuição gaussiana.

A maximização da não gaussianidade do componente  $Y$  no espaço  $n$ -dimensional dos vetores  $\mathbf{w}$  permite encontrar  $2n$  máximos locais, dois para cada componente independente, correspondendo a  $-S_i$  e  $S_i$ , no que decorre que os componentes independentes não podem ser univocamente determinados. Para estimar os outros componentes independentes é necessário encontrar todos esses máximos locais. Uma vez que os componentes independentes são não correlacionados, a busca é feita sob a restrição de que as novas estimativas sejam não correlacionadas com as anteriores, o que corresponde à ortogonalização num espaço de vetores branqueados.

### 2.1.6.1 Curtose

A curtose é uma estatística de quarta ordem que pode ser usada para medir a não gaussianidade de uma variável aleatória  $Y$ .

O coeficiente de curtose de uma variável aleatória  $Y$  com média  $\mu$  mede a intensidade dos picos da sua distribuição de probabilidade e é definido como

$$\kappa_4 = \frac{E[(Y - \mu)^4]}{\{E[(Y - \mu)^2]\}^2}, \quad (2.23)$$

supondo-se a existência do quarto momento de  $Y$  (MAGALHÃES, 2006).

A curtose de distribuições gaussianas com média zero e variância unitária é igual a 3. É comum usar o coeficiente de curtose da distribuição gaussiana para compará-lo com outras distribuições de probabilidade. Alternativamente, alguns autores já definem este coeficiente subtraído de 3 (HYVÄRINEN; OJA, 2000; MOOD; GRAYBILL; BOES, 1974), sendo também conhecido como excesso de curtose.

Considerando que a estimação do modelo ICA é feita a partir de variáveis centralizadas e branqueadas, o coeficiente de excesso de curtose para estas variáveis é dado por

$$kurt(Y) = E(Y^4) - 3. \quad (2.24)$$

Logo, o excesso de curtose é nulo para variáveis gaussianas e não nulo para praticamente a grande maioria das variáveis aleatórias não gaussianas. Variáveis aleatórias com curtose negativa são denominadas variáveis subgaussianas ou com distribuição platicúrtica e aquelas com curtose positiva supergaussianas ou com distribuição leptocúrtica.

A curtose, ou ainda o seu valor absoluto, tem sido largamente utilizada para quantificar o grau de não gaussianidade de uma variável aleatória  $Y$  em ICA.

É uma medida que apresenta simplicidade computacional pela sua fácil estimação a partir do quarto momento de uma amostra de dados.

### 2.1.6.2 Algoritmo de ponto fixo usando curtose

Como dito no início desta seção, a estimação de um componente independente  $Y$  é feita buscando o vetor  $\mathbf{w}$  que torna a distribuição de  $\mathbf{w}^T \cdot \mathbf{X}$  o mais distante possível de uma distribuição gaussiana. Considerando que o vetor  $\mathbf{X}$  tornou-se o vetor  $\mathbf{Z}$  após o pré-processamento de branqueamento e assumindo que a medida de não gaussianidade usada é a curtose, o problema de otimização consiste em buscar uma direção de projeção  $\mathbf{w}$  que satisfaça o critério

$$\mathbf{w} = \arg \max |kurt(\mathbf{w}^T \cdot \mathbf{Z})|. \quad (2.25)$$

Na prática, para maximizar o valor absoluto da curtose, inicia-se com algum vetor de pesos  $\mathbf{w}$  arbitrário, calcula-se a direção na qual o valor absoluto da curtose de  $\mathbf{w}^T \cdot \mathbf{Z}$  cresce mais fortemente, que é a direção do seu gradiente, e move-se  $\mathbf{w}$  nesta direção.

Um algoritmo de iteração de ponto fixo é uma variação do método do gradiente que torna a estimação mais rápida. Não usa taxa de aprendizado, nem depende das condições iniciais do algoritmo que pode tornar a convergência mais lenta a depender da escolha destes valores.

O gradiente do valor absoluto da curtose de  $\mathbf{w}^T \cdot \mathbf{Z}$  pode ser calculado como

$$\frac{\partial |kurt(\mathbf{w}^T \cdot \mathbf{Z})|}{\partial \mathbf{w}} = 4 \text{sign}(kurt(\mathbf{w}^T \cdot \mathbf{Z})) \{E[\mathbf{Z}(\mathbf{w}^T \cdot \mathbf{Z})^3] - 3\mathbf{w}E[(\mathbf{w}^T \cdot \mathbf{Z})^2]\}. \quad (2.26)$$

Para dados branqueados  $E[(\mathbf{w}^T \cdot \mathbf{Z})^2] = \|\mathbf{w}\|^2$ , no que decorre que

$$\frac{\partial |kurt(\mathbf{w}^T \cdot \mathbf{Z})|}{\partial \mathbf{w}} = 4 \cdot \text{sign}(kurt(\mathbf{w}^T \cdot \mathbf{Z})) \{E[\mathbf{Z}(\mathbf{w}^T \cdot \mathbf{Z})^3] - 3\mathbf{w} \cdot \|\mathbf{w}\|^2\}. \quad (2.27)$$

Para derivar uma iteração de ponto fixo eficiente, no ponto estável do algoritmo o gradiente (2.27) deve ser igual a  $\mathbf{w}$  multiplicado por uma constante escalar. O problema de otimização está sujeito à restrição de manter a norma de  $\mathbf{w}$  unitária. Desse modo, adicionar o gradiente a  $\mathbf{w}$  não muda a sua direção e a convergência é atingida. Assim temos que

$$\mathbf{w} \propto E[\mathbf{Z} \cdot (\mathbf{w}^T \cdot \mathbf{Z})^3] - 3 \cdot \mathbf{w} \cdot \|\mathbf{w}\|^2. \quad (2.28)$$

O novo valor de  $\mathbf{w}$  é atualizado pelo cálculo da expressão do lado direito da equação

$$\mathbf{w} \leftarrow E[\mathbf{Z} \cdot (\mathbf{w}^T \cdot \mathbf{Z})^3] - 3 \cdot \mathbf{w}. \quad (2.29)$$

Após cada iteração o vetor  $\mathbf{w}$  é dividido pela sua norma para permanecer no domínio da restrição  $\|\mathbf{w}\| = 1$ . O vetor final  $\mathbf{w}$  fornece um dos componentes independentes a partir da combinação linear  $\mathbf{w}^T \cdot \mathbf{Z}$ . Na prática toda a computação é feita a partir de uma amostra disponível dos dados branqueados  $\mathbf{Z}(1), \mathbf{Z}(2), \dots, \mathbf{Z}(T)$  e a esperança em (2.29) deve ser substituída pela média amostral.

Atingir a convergência na iteração de ponto fixo significa que os valores de  $\mathbf{w}$ , o novo e o anterior, apontam na mesma direção. Não é necessário que o vetor convirja a um único ponto, uma vez que  $\mathbf{w}$  e  $-\mathbf{w}$  definem a mesma direção. Como visto na seção 2.1.3 os componentes independentes não podem ser determinados de forma única, pois sempre permanece uma ambiguidade no sinal.

Este algoritmo é uma das versões do algoritmo FastICA que será apresen-

tado mais detalhadamente na seção 2.1.6.6.

Um dos problemas que pode ocorrer na estimação é devido à sensibilidade da curtose a *outliers*, ou valores extremos. Por ser um termo de quarta potência pode gerar valores bastante elevados afetando fortemente o resultado final. Isto significa que, apesar de ser uma medida de fácil computação, não é um estimador robusto de não gaussianidade.

### 2.1.6.3 Negentropia

Outra forma de quantificar o grau de não gaussianidade é usando negentropia. A negentropia é baseada na quantidade de informação teórica da entropia diferencial.

Entropia é um conceito básico da teoria de informação e é uma medida da incerteza média associada à observação de uma variável aleatória (COVER; THOMAS, 1991). A entropia será maior o quanto mais imprevisível for a variável. A entropia diferencial  $H$  de uma variável aleatória  $Y$  com densidade  $f_Y(\cdot)$  é definida como

$$H(Y) = - \int_y f_Y(y) \ln f_Y(y) dy. \quad (2.30)$$

Uma propriedade fundamental da teoria da informação é que uma variável gaussiana tem maior entropia que qualquer outra variável aleatória de mesma variância. Neste sentido, definindo negentropia como

$$J(Y) = H(Y_{gauss}) - H(Y), \quad (2.31)$$

obtém-se uma medida sempre não negativa que pode ser usada para medir o grau de não gaussianidade de uma variável aleatória  $Y$  ao ser comparada com uma variável gaussiana  $Y_{gauss}$ . A maximização da negentropia conduz à estimação de

um componente independente, procedendo de forma análoga ao uso de curtose.

Negentropia é um ótimo estimador de não gaussianidade por estar solidamente fundamentado na teoria estatística. O problema do seu uso é o difícil cálculo da entropia que requer uma estimação (possivelmente não paramétrica) da densidade da variável aleatória envolvida e pode exigir grande esforço computacional.

#### 2.1.6.4 Aproximações da negentropia

Algumas funções de aproximação da negentropia são usadas na implementação dos algoritmos para contornar a dificuldade de estimar a entropia usando a definição.

O método clássico para estimar a negentropia é a expansão em densidade polinomial, usando cumulantes de ordem mais alta, cuja aproximação é dada por (HYVÄRINEN; OJA, 2000; JONES; SIBSON, 1987)

$$J(Y) \approx \frac{1}{12}[E(Y^3)]^2 + \frac{1}{48}[kurt(Y)]^2, \quad (2.32)$$

em que  $Y$  é uma variável aleatória de média zero e variância unitária. No entanto, por fazer uso da curtose no seu cálculo, esta aproximação apresenta o mesmo problema de não ser um estimador robusto.

Outras abordagens usam funções não quadráticas  $G_i$  propondo a aproximação da negentropia baseada nas esperanças  $E[G_i(Y)]$

$$J(Y) \approx k_1\{E[G_1(Y)]\}^2 + k_2\{E[G_2(Y)] - E[G_2(Y_{gauss})]\}^2, \quad (2.33)$$

em que  $k_1$  e  $k_2$  são constantes positivas,  $Y_{gauss}$  é uma variável gaussiana de média zero e variância unitária e  $Y$  é também uma variável de média zero e variância unitária. Mesmo nos caso em que esta aproximação não seja muito acurada (2.33)

pode ser usada como uma medida de não gaussianidade consistente, pois é sempre não negativa e é igual a zero no caso de  $Y$  ter distribuição gaussiana.

Quando é usada apenas uma função não quadrática  $G$ , a aproximação torna-se

$$J(Y) \propto \{E[G(Y)] - E[G(Y_{gauss})]\}^2. \quad (2.34)$$

Com a escolha correta da função  $G$  obtém-se aproximações da negentropy muito melhores que a dada pela equação (2.32). Em particular, escolhendo uma  $G$  que não cresça muito rapidamente obtém-se um estimador mais robusto, computacionalmente mais simples e com propriedades estatísticas interessantes.

Exemplos de funções  $G$  normalmente usadas com as características descritas acima são

$$G_1(y) = \frac{1}{a_1} \log \cosh(a_1 y) \quad (2.35)$$

$$G_2(y) = -\exp\left(-\frac{y^2}{2}\right), \quad (2.36)$$

em que  $1 \leq a_1 \leq 2$  é uma constante, considerando-se frequentemente  $a_1 = 1$ .

Os gráficos destas funções estão ilustrados na Figura 4, junto com o gráfico da função  $G_3(y) = \frac{y^4}{4}$ , normalmente utilizado como aproximação da curtose. Analisando as três curvas, pode-se observar a velocidade de crescimento de cada uma delas.

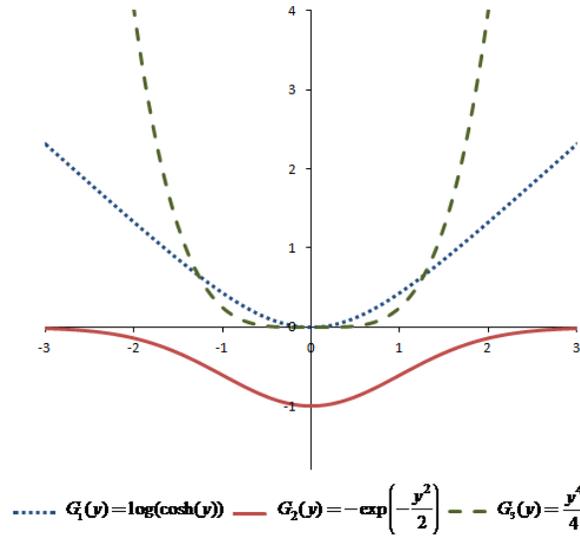


Figura 4 Funções  $G_1$  e  $G_2$  aproximações de negentropia e função  $G_3$  aproximação de curtose

### 2.1.6.5 Algoritmo de ponto fixo usando negentropia

A aproximação de negentropia dada em (2.34) fornece uma função objetivo para a construção de um algoritmo de iteração de ponto fixo para a estimação ICA. Para se encontrar um componente independente  $Y = \mathbf{w}^T \cdot \mathbf{Z}$ , a partir da maximização da negentropia, deve-se maximizar a função  $J_G$  dada por

$$J_G(\mathbf{w}) = \{E[G(\mathbf{w}^T \cdot \mathbf{Z})] - E[G(Z_{gauss})]\}^2, \quad (2.37)$$

em que  $\mathbf{w}$  é um vetor de pesos sob a restrição de que  $E[(\mathbf{w}^T \cdot \mathbf{Z})^2] = \|\mathbf{w}\|^2 = 1$ .

Os máximos da aproximação da negentropia ocorrem para valores ótimos de  $E[G(\mathbf{w}^T \cdot \mathbf{Z})]$ . Utilizando o método de Lagrange, estes máximos sujeitos à restrição  $\|\mathbf{w}\|^2 = 1$  são obtidos nos pontos em que o gradiente do Lagrangiano é

nulo:

$$E[Zg(\mathbf{w}^T \cdot \mathbf{Z})] + \beta \mathbf{w} = 0, \quad (2.38)$$

sendo  $g$  a função derivada de  $G$  e  $\beta$  uma constante que pode ser calculada por  $\beta = E[\mathbf{w}_0^T \mathbf{Z} g(\mathbf{w}_0^T \mathbf{Z})]$  para  $\mathbf{w}_0$  um valor ótimo de  $\mathbf{w}$ .

Denotando a função do lado esquerdo da equação (2.38) por  $F$  e resolvendo (2.38) pelo método de Newton, obtém-se a matriz Jacobiana  $JF(\mathbf{w})$  (equivalente à segunda derivada do Lagrangiano) dada por

$$JF(\mathbf{w}) = E[ZZ^T g'(\mathbf{w}^T \cdot \mathbf{Z})] + \beta \mathbf{I}. \quad (2.39)$$

Fazendo-se a aproximação  $E[ZZ^T g'(\mathbf{w}^T \cdot \mathbf{Z})] \approx E[ZZ^T] \cdot E[g'(\mathbf{w}^T \cdot \mathbf{Z})] = E[g'(\mathbf{w}^T \cdot \mathbf{Z})]$ , a matriz  $JF(\mathbf{w})$  torna-se diagonal, ficando facilmente inversível. A constante  $\beta$  também pode ser aproximada usando o atual valor de  $\mathbf{w}$  no lugar de  $\mathbf{w}_0$ , no que deriva a seguinte iteração

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{E[Zg(\mathbf{w}^T \cdot \mathbf{Z})] + \beta \mathbf{w}}{E[g'(\mathbf{w}^T \cdot \mathbf{Z})] + \beta}. \quad (2.40)$$

O algoritmo pode ainda ser simplificado multiplicando ambos os lados de (2.40) por  $E[g'(\mathbf{w}^T \cdot \mathbf{Z})] + \beta$ , ficando

$$\mathbf{w} \leftarrow E[Zg(\mathbf{w}^T \cdot \mathbf{Z})] - E[g'(\mathbf{w}^T \cdot \mathbf{Z})] \cdot \mathbf{w}. \quad (2.41)$$

Após cada iteração o vetor  $\mathbf{w}$  é normalizado até atingir a convergência como no algoritmo que maximiza a curtose.

### 2.1.6.6 Algoritmo FastICA

Os procedimentos descritos nas seções 2.1.6.2 e 2.1.6.5 são diferentes formulações do algoritmo FastICA (Fast Independent Component Analysis), um método com desempenho computacional altamente eficiente na estimação de ICA, proposto por Hyvärinen e Oja (1997). A formulação geral do algoritmo será brevemente descrita a seguir.

O primeiro passo é a escolha de uma função não linear  $g$ , derivada de funções  $G$  não quadráticas, por exemplo as funções dadas em (2.35) e (2.36) que dão aproximações robustas de negentropia ou a derivada correspondente à função de quarta potência aproximação da curtose. As funções derivadas são basicamente

$$g_1(y) = \tanh(a_1 y) \quad (2.42)$$

$$g_2(y) = y \exp\left(-\frac{y^2}{2}\right) \quad (2.43)$$

$$g_3(y) = y^3, \quad (2.44)$$

em que  $1 \leq a_1 \leq 2$  é uma constante, considerando-se frequentemente  $a_1 = 1$ .

O passo seguinte é a iteração propriamente dita, dada em (2.41) seguida da normalização do vetor atualizado, até que se atinja a convergência.

As derivadas de  $g$  são dadas por

$$g'_1(y) = a_1(1 - \tanh^2(a_1 y)) \quad (2.45)$$

$$g'_2(y) = (1 - y^2) \exp\left(-\frac{y^2}{2}\right) \quad (2.46)$$

$$g'_3(y) = 3y^2. \quad (2.47)$$

Como existe a restrição  $E(y^2) = 1$ , a derivada  $g'_3$  é reduzida à constante

3.

De forma mais detalhada seguem-se as seguintes etapas a partir de uma amostra de dados disponível  $\mathbf{X}(1), \mathbf{X}(2), \dots, \mathbf{X}(T)$ , sendo as esperanças das funções dadas substituídas pela média amostral:

1. Centralizar e branquear os dados observados  $\mathbf{X}$ , obtendo o vetor  $\mathbf{Z}$ .
2. Escolher aleatoriamente valores iniciais para o vetor  $\mathbf{w}$  com norma unitária.
3. Calcular

$$\mathbf{w} \leftarrow E[\mathbf{Z}g(\mathbf{w}^T \cdot \mathbf{Z})] - E[g'(\mathbf{w}^T \cdot \mathbf{Z})] \cdot \mathbf{w},$$

com  $g$  escolhida entre as equações definidas (2.42) - (2.44) e  $g'$  a sua derivada (2.45) - (2.47).

4. Normalizar o novo  $\mathbf{w}$  obtido no passo 3

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}.$$

5. Caso não convirja, voltar para o passo 3.

Atingir a convergência significa que os valores antigo e novo de  $\mathbf{w}$  apontam na mesma direção, não precisam ser necessariamente os mesmos.

Como já discutido na descrição do algoritmo para curtose, FastICA é um algoritmo iterativo de ponto-fixo muito mais rápido que os convencionais algoritmos de gradiente descendente, pois a sua convergência é no mínimo quadrática. Não há necessidade de escolha de parâmetros para a execução do algoritmo, tais como taxa de aprendizado, o que o torna mais confiável e fácil de usar. Em Hyvärinen (1999) encontram-se mais detalhes sobre o comportamento das funções objetivo escolhidas e demonstrações de como os estimadores são consistentes, robustos, com variância assintótica e de rápida convergência.

### 2.1.6.7 Algoritmo FastICA para vários componentes independentes

O algoritmo descrito estima apenas um componente independente. Portanto, para se estimar diversos componentes, é necessário executar o algoritmo várias vezes. Para garantir que um mesmo componente independente não seja novamente estimado recorre-se a métodos de ortogonalização que diferem entre si por ortogonalizarem os componentes estimados um de cada vez (ortogonalização deflacionária) ou em paralelo (ortogonalização simétrica).

- **Ortogonalização deflacionária**

Na ortogonalização deflacionária utiliza-se o processo de ortogonalização de Gram-Schmidt, em que os vetores de saída são ortogonalizados um de cada vez e a ortogonalização do componente seguinte depende dos anteriores. Após a estimação do  $p$ -ésimo vetor  $\mathbf{w}_p$ , subtrai-se de  $\mathbf{w}_p$  as projeções dele sobre os demais  $p - 1$  vetores que já foram estimados e então se renormaliza o vetor  $\mathbf{w}_p$ . Mais precisamente, seguem-se os seguintes passos:

1. Escolher  $k$ , número de componentes independentes a ser estimado. Definir  $p \leftarrow 1$ .
2. Escolher aleatoriamente valores iniciais para o vetor  $\mathbf{w}_p$  com norma unitária.
3. Executar uma iteração do algoritmo FastICA para estimar uma unidade  $\mathbf{w}_p$ .
4. Ortogonalizar  $\mathbf{w}_p$  da seguinte forma:

$$\mathbf{w}_p \leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_j) \mathbf{w}_j.$$

5. Normalizar  $\mathbf{w}_p$  obtido no passo 4, dividindo-o por sua norma.

6. Caso não convirja, voltar para o passo 3.
7. Definir  $p \leftarrow p + 1$ . Se  $p \leq k$  voltar ao passo 2.

A escolha do número de componentes independentes a serem estimados é feita na fase de branqueamento dos dados. Ao estimar a matriz de covariância dos dados observados, o FastICA aplica PCA ordenando os componentes em ordem decrescente de energia. Escolhe-se então um número de componentes independentes  $k < n$  que preserve a maior quantidade de energia, ou seja, que explique um bom percentual de variação dos dados.

- **Ortogonalização simétrica**

Na orthogonalização simétrica todos os componentes são calculados em paralelo e a orthogonalização é aplicada através de métodos simétricos na matriz resultante composta por todos os vetores estimados após uma execução de cada componente.

A orthogonalização simétrica da matriz  $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k)^T$  pode ser realizada pelo método clássico

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}. \quad (2.48)$$

A raiz quadrada da inversa de  $\mathbf{W}\mathbf{W}^T$  pode ser obtida da decomposição em autovalores de  $\mathbf{W}\mathbf{W}^T = \mathbf{E} \cdot \text{diag}(d_1, \dots, d_m) \cdot \mathbf{E}^T$  dada por

$$(\mathbf{W}\mathbf{W}^T)^{-1/2} = \mathbf{E} \cdot \text{diag}(d_1^{-1/2}, \dots, d_m^{-1/2}) \cdot \mathbf{E}^T. \quad (2.49)$$

Uma simples alternativa de orthogonalização de  $\mathbf{W}$  pode ser utilizando o algoritmos iterativo:

1. Seja  $\mathbf{W} \leftarrow \mathbf{W}/\|\mathbf{W}\|$ .
2. Seja  $\mathbf{W} \leftarrow \frac{3}{2}\mathbf{W} - \frac{1}{2}\mathbf{W}\mathbf{W}^T\mathbf{W}$ .
3. Se  $\mathbf{W}\mathbf{W}^T$  não é aproximadamente a matriz identidade, repetir o passo 2.

Utilizando um destes métodos de ortogonalização, o processo completo do FastICA segue os seguintes passos:

1. Escolher  $k$ , número de componentes independentes a ser estimado.
2. Escolher aleatoriamente valores iniciais para os vetores  $\mathbf{w}_i, i = 1, 2, \dots, k$ , com norma unitária. Ortogonalizar a matriz  $\mathbf{W}$  como na etapa 4 abaixo.
3. Executar uma iteração do algoritmo FastICA para estimar cada unidade  $\mathbf{w}_i$  em paralelo.
4. Ortogonalizar a matriz  $\mathbf{W}$  como descrito em (2.49) ou utilizando a ortogonalização por processo iterativo, como descrito nesta seção.
5. Caso não convirja, voltar para o passo 3.

### 2.1.7 Estimação ICA pela maximização da verossimilhança

A estimação por máxima verossimilhança é uma técnica fundamental na teoria estatística e também pode ser aplicada na estimação ICA.

A função de verossimilhança para o modelo ICA é obtida usando a fórmula clássica do método jacobiano para transformação de funções de densidade de probabilidade (fdp). Considerando o modelo ICA, definido em (2.2), a densidade  $p_X$  do vetor de misturas é dada por

$$p_X(\mathbf{x}) = |\det\mathbf{W}|p_S(\mathbf{s}) = |\det\mathbf{W}|\prod_i p_i(s_i), \quad (2.50)$$

em que  $\mathbf{W} = \mathbf{A}^{-1}$  e  $p_i$  denota as densidades dos componentes independentes. Expressando  $p_{\mathbf{X}}(\mathbf{x})$  como função de  $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n)^T$  e  $\mathbf{x}$ , tem-se

$$p_{\mathbf{X}}(\mathbf{x}) = |\det \mathbf{W}| \prod_i p_i(\mathbf{w}_i^T \mathbf{x}). \quad (2.51)$$

Assumindo-se que existam  $T$  observações  $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(T)$ , a função de verossimilhança  $L(\mathbf{W})$  pode ser obtida como o produto da densidade calculada nos  $T$  pontos

$$L(\mathbf{W}) = \prod_{t=1}^T \prod_{i=1}^n p_i(\mathbf{w}_i^T \mathbf{x}(t)) |\det \mathbf{W}| \quad (2.52)$$

e a função de log-verossimilhança é dada por

$$\log L(\mathbf{W}) = \sum_{t=1}^T \sum_{i=1}^n \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{W}|. \quad (2.53)$$

Para simplificar a notação pode-se substituir a soma sobre as amostras  $\mathbf{x}(t)$  pelo operador esperança e dividir a expressão por  $T$ , obtendo-se

$$\frac{1}{T} \log L(\mathbf{W}) = E \left[ \sum_{i=1}^n \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) \right] + \log |\det \mathbf{W}|, \quad (2.54)$$

lembrando que na prática a esperança usada nesta expressão é uma média calculada usando os valores observados na amostra.

O problema maior nesta abordagem é que geralmente as densidades dos componentes independentes não são conhecidas. Uma forma de contornar a dificuldade em estimar estas densidades é aproximá-las por uma família de densidades que são especificadas por um número limitado de parâmetros. Em Hyvärinen, Karhunen e Oja (2001) é provado que é possível fazer a estimativa da distribuição dos componentes independentes a partir de apenas duas aproximações. Para cada

componente independente, basta determinar qual das duas aproximações é mais adequada (subgaussiana ou supergaussiana) através do cálculo de momentos não polinomiais e escolher aquela que melhor satisfaz um critério de estabilidade adotado. Como exemplo pode-se adotar as seguintes aproximações para os logaritmos de densidades supergaussianas e subgaussianas, respectivamente,

$$\log \tilde{p}_i^+(s) = \alpha_1 - 2 \log \cosh(s) \quad (2.55)$$

$$\log \tilde{p}_i^-(s) = \alpha_2 - [s^2/2 - \log \cosh(s)], \quad (2.56)$$

em que  $\alpha_1$  e  $\alpha_2$  são constantes positivas escolhidas de modo a tornar estas duas funções logaritmos de densidades de probabilidade.

Uma vez definidas as densidades, para realizar a estimação de máxima verossimilhança são necessários algoritmos que maximizem a função de verossimilhança. Entre os métodos existentes na literatura o algoritmo de Bell-Sejnowski é o mais popular, inicialmente proposto em Bell e Sejnowski (1995). O método consiste basicamente em calcular o gradiente da função de log-verossimilhança (2.54) dado por

$$\frac{1}{T} \frac{\partial \log L}{\partial \mathbf{W}} = (\mathbf{W}^T)^{-1} + E[\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T], \quad (2.57)$$

em que  $\mathbf{g}(\mathbf{y}) = (g_1(y_1), \dots, g_n(y_n))$  é um vetor contendo as derivadas das aproximações das distribuições dos componentes independentes  $g_i(y_i) = (\log p_i)'$ .

Assim, a iteração para o algoritmo de estimação por máxima verossimilhança é dada por

$$\Delta \mathbf{W} \propto (\mathbf{W}^T)^{-1} + E[\mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^T]. \quad (2.58)$$

Porém, este algoritmo converge lentamente em alguns casos devido à necessidade de inversão da matriz  $\mathbf{W}$  a cada passo. Nos anos seguintes novas formulações desta ideia resultaram em algoritmos com melhor velocidade de convergên-

cia, por exemplo, com a utilização do gradiente natural na maximização da função de verossimilhança. Mais detalhes e outras abordagens para este tipo de estimação podem ser encontrados em Amari (1998), Bell e Sejnowski (1995) e Hyvärinen, Karhunen e Oja (2001).

### 2.1.8 Outros princípios de estimação ICA

Outros princípios para a estimação ICA não serão desenvolvidos neste trabalho, mas serão brevemente citados nesta seção.

- **ICA pela minimização da informação mútua.**

É uma abordagem inspirada na teoria da informação que usa o conceito de informação mútua. A informação mútua é uma medida de aferição da distância entre a densidade conjunta e a densidade das marginais de variáveis aleatórias, podendo ser usada como uma medida natural de dependência entre os componentes independentes. Pode-se definir ICA, no seu modelo invertido (2.3), como uma decomposição linear que minimiza a mútua informação entre os componentes  $S_i$ . Pode-se dizer que é um processo de certa forma equivalente a encontrar direções nas quais a negentropia é maximizada.

- **ICA por métodos tensoriais.**

Consiste no uso de tensores cumulantes de ordens superiores. Tensores podem ser vistos como uma generalização de matrizes ou operadores lineares; por exemplo, são uma generalização da matriz de covariância que é um tensor cumulante de segunda ordem, enquanto que um tensor de quarta ordem pode ser considerado como uma matriz de dimensão quatro cujos elementos são cumulantes de quarta ordem,  $\text{cum}(X_i, X_j, X_k, X_l)$ . Assim como diagonalizar a matriz de

covariância produz componentes não correlacionados, a idéia é diagonalizar tensores cumulantes de ordem quatro para se obter componentes com independência estatística nesta ordem. Joint Approximate Diagonalization of Eigenmatrices - JADE (CARDOSO; SOULOUMIAC, 1993) é o algoritmo mais conhecido utilizando esta abordagem e possui ótimo desempenho quando aplicado a dados de baixa dimensão.

## 2.2 Análise discriminante

A análise discriminante é uma técnica estatística empregada para diferenciar populações ou classificar novos indivíduos em uma das várias populações, considerando a estrutura multidimensional dos dados observados.

Considerando-se que existam  $m$  grupos ou populações  $\Pi_1, \Pi_2, \dots, \Pi_m$ , conhecidos *a priori*, e amostras aleatórias para estes grupos de tamanhos  $n_1, n_2, \dots, n_m$ , cada observação aleatória  $\mathbf{X}_j$  é um vetor  $p$ -dimensional dado por  $\mathbf{X}_j = [X_{j1}, X_{j2}, \dots, X_{jp}]^T$ ,  $j = 1, 2, \dots, n_i$ . Segundo Ferreira (2008), conhecer os grupos significa que o número de grupos  $m$  e os indivíduos de cada grupo são perfeitamente identificados antes da análise ser aplicada. O objetivo é determinar a qual dos grupos ou populações irá pertencer um novo indivíduo ou um novo conjunto de indivíduos, construindo para este fim regras de discriminação.

De modo geral, conceitualmente estas regras procuram encontrar a separação máxima entre as populações considerando a maximização da diferença entre as médias das populações relativamente aos desvios-padrão no interior de cada população, sem perder a estrutura de covariância das variáveis observadas.

### 2.2.1 Regras de classificação

Segundo Manly (2008), a abordagem mais simples para elaborar regras de classificação é a proposta de Fisher (1936) de tomar uma combinação linear das variáveis  $X_1, X_2, \dots, X_p$ ,

$$Z = a_1X_1 + a_2X_2 + \dots + a_pX_p. \quad (2.59)$$

Se o valor médio da variável  $Z$  muda consideravelmente de grupo para grupo, com os valores dentro do grupo sendo aproximadamente constantes, os grupos podem ser separados.

Uma maneira de determinar os coeficientes  $a_1, a_2, \dots, a_p$  é maximizar a razão entre a variabilidade entre as populações e a variabilidade comum dentro das populações, que consiste num problema de maximização de uma razão de formas quadráticas. Quando esta abordagem é usada é possível determinar várias combinações lineares para separar grupos. O número de soluções  $s$  disponíveis é o mínimo entre  $p$  e  $m - 1$ . Estas combinações lineares são denominadas funções discriminantes canônicas ou eixos discriminantes de Fisher.

Quando as densidades das populações não são conhecidas, as suas características são estimadas a partir das correspondentes amostras aleatórias, comumente chamadas de amostras de treinamento. Por motivo de praticidade nas aplicações deste trabalho, as funções discriminantes serão aqui definidas diretamente a partir dos estimadores amostrais. Considerando  $\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{in_i}$  uma amostra aleatória da  $i$ -ésima população, sejam

$$\bar{\mathbf{X}}_i = \frac{\sum_{j=1}^{n_i} \mathbf{X}_{ij}}{n_i} \quad (2.60)$$

e

$$\bar{\mathbf{X}} = \frac{\sum_{i=1}^m n_i \bar{\mathbf{X}}_i}{\sum_{i=1}^m n_i} \quad (2.61)$$

os estimadores da média amostral da  $i$ -ésima população e da média global de todas as populações, respectivamente, em que  $\mathbf{X}_{ij}$  representa a observação da  $j$ -ésima unidade amostral aleatória da  $i$ -ésima população.

Para a determinação da regra discriminante inicialmente são calculadas a matriz de somas de quadrados e produtos cruzados entre as médias das  $m$  populações,  $\mathbf{B}$ , e a matriz de soma de quadrados e produtos cruzados dentro da amostra,  $\mathbf{W}$ , dadas por

$$\mathbf{B} = \sum_{i=1}^m n_i (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^T \quad (2.62)$$

e

$$\mathbf{W} = \sum_{i=1}^m \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T. \quad (2.63)$$

Em seguida determina-se os autovalores e autovetores da matriz  $\mathbf{W}^{-1}\mathbf{B}$ . Se os autovalores são  $\lambda_1 > \lambda_2 > \dots > \lambda_s$ , então  $\lambda_i$  é a razão da soma dos quadrados entre os grupos e da soma de quadrados dentro dos grupos para a  $i$ -ésima combinação linear,  $Z_i$ , enquanto que os elementos do correspondente autovetor,  $\mathbf{a}_i^T = (a_{i1}, a_{i2}, \dots, a_{ip})$ , são os coeficientes das variáveis  $X$  para este índice.

Os eixos discriminantes  $Z_1, Z_2, \dots, Z_s$  são combinações lineares das variáveis originais escolhidos de tal maneira que  $Z_1$  captura as diferenças de grupos tanto quanto possível,  $Z_2$  captura tanto quanto possível as diferenças de grupos não representadas por  $Z_1$ ,  $Z_3$  captura tanto quanto possível as diferenças de grupos não representadas por  $Z_1$  e  $Z_2$  e assim sucessivamente. O que se deseja é que os primeiros poucos eixos discriminantes consigam refletir quase todas as possí-

veis e importantes diferenças entre os grupos.

Dado um vetor  $\mathbf{x}$ , realização de uma observação aleatória  $p$ -dimensional  $\mathbf{X}$ , é possível estabelecer uma regra de classificação deste novo indivíduo em uma única das  $m$  populações, a partir dos eixos discriminantes. Esta regra, dita critério de Fisher, é baseada nos  $r \leq s$  primeiros eixos discriminantes e estabelece que devemos alocar  $\mathbf{x}$  na população  $\Pi_i$  se

$$\sum_{j=1}^r [\mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{X}}_i)]^2 \leq \sum_{j=1}^r [\mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{X}}_l)]^2, \quad (2.64)$$

para todo  $l \neq i, l = 1, \dots, m$ .

Uma das principais vantagens ao se adotar o critério de Fisher (1936) na discriminação entre populações é o de não ser necessário o conhecimento das densidades populacionais, nem assumir que estas sejam gaussianas, apesar de se ter que considerar homogeneidade das matrizes de covariância das diferentes populações.

Outra opção de regra discriminante para classificar uma observação  $p$ -variada  $\mathbf{x}$  na população  $\Pi_i$ , considera o caso em que as matrizes de covariância das populações são diferentes entre si, mas pressupõe que as densidades populacionais sejam conhecidas; em particular, que estas tenham distribuição gaussiana  $p$ -variada. Esta regra está baseada na minimização da probabilidade total de classificação incorreta (FERREIRA, 2008).

Como no caso anterior, não se conhecendo os parâmetros populacionais, são usados os estimadores correspondentes obtidos na amostra de treinamento. A função discriminante quadrática ou escore quadrático de discriminação para a  $i$ -ésima população, representado por  $Q_i(\mathbf{x})$ , é dado por

$$Q_i(\mathbf{x}) = -\frac{1}{2} \ln(|\mathbf{C}_i|) - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{X}}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i) + \ln(p_i), \quad (2.65)$$

para  $i = 1, 2, \dots, m$ , em que  $p_i$  é a probabilidade *a priori* da amostra pertencer à população  $i$  e  $\bar{\mathbf{X}}_i$  e  $\mathbf{C}_i$  são os estimadores amostrais da média e covariância da  $i$ -ésima população, respectivamente, dados por (2.60) e por

$$\mathbf{C}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T. \quad (2.66)$$

A partir da função  $Q_i$  considera-se classificar a observação  $\mathbf{x}$  na população  $\Pi_i$  se

$$Q_i(\mathbf{x}) \geq Q_l(\mathbf{x}) \quad (2.67)$$

para todo  $l \neq i, l = 1, \dots, m$ .

Interessante observar que ao se usar a função discriminante quadrática assumindo igualdade das matrizes de covariância, verifica-se uma importante relação entre as funções discriminantes linear de Fisher (1936) e quadrática.

Supondo as matrizes de covariâncias das populações iguais a  $\Sigma$ , um estimador não viesado da matriz de covariância comum é dado por

$$\mathbf{C}_p = \frac{\sum_{i=1}^m (n_i - 1) \mathbf{C}_i}{\sum_{i=1}^m (n_i - 1)}. \quad (2.68)$$

Para este caso particular, a função discriminante quadrática (2.65) fica simplificada por

$$Q_i(\mathbf{x}) = -\frac{1}{2} \ln(|\mathbf{C}_p|) - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{X}}_i)^T \mathbf{C}_p^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i) + \ln(p_i), \quad (2.69)$$

para  $i = 1, 2, \dots, m$ .

Por outro lado, desenvolvendo o termo a ser minimizado no critério de

Fisher (2.64) para as  $s$  funções discriminantes canônicas, pode-se provar que

$$\sum_{j=1}^s [\mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{X}}_i)]^2 = (\mathbf{x} - \bar{\mathbf{X}}_i)^T \mathbf{C}_p^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i), \quad (2.70)$$

donde se observa que o critério de Fisher (1936) minimiza a distância de Mahalanobis entre a observação  $\mathbf{x}$  e o vetor de médias da  $i$ -ésima população.

Comparando (2.69) e (2.70), tem-se

$$\sum_{j=1}^s [\mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{X}}_i)]^2 = -2Q_i(\mathbf{x}) - \ln(|C_p|) + 2\ln(p_i), \quad (2.71)$$

concluindo-se que os critérios de discriminação de Fisher e quadráticos são equivalentes, a menos da parcela  $2\ln(p_i)$ . Portanto, se os valores  $p_i$  forem iguais para todo  $i = 1, 2, \dots, m$ , pode-se dizer que minimizar a distância de Mahalanobis no critério de Fisher, em que não é necessário pressupor densidade gaussiana, é equivalente a maximizar a quantidade  $Q_i$  para o caso particular de populações com homogeneidade de matrizes de covariância. Mais detalhes podem ser vistos em Ferreira (2008).

### 2.2.2 Estimação de classificações incorretas

As regras de discriminação, por utilizarem amostras de treinamento na estimação dos parâmetros populacionais ou por terem alguma das suas suposições violadas, estão sujeitas a erros de classificação. A qualidade da função discriminante é avaliada pela estimativa da probabilidade global de acerto. O valor desta probabilidade associado às estimativas de probabilidades de erros de classificação estabelecem um indicador frequentista da acurácia da regra de classificação.

A probabilidade de erro de um indivíduo ser classificado na população  $l$

pela regra de classificação utilizada, sendo que a sua população de origem seja a população  $i$ , para  $l$  e  $i = 1, 2, \dots, m$  e  $l \neq i$ , é dada por

$$P(l|i) = \frac{n_{il}}{n_i}, \quad (2.72)$$

em que  $n_{il}$  representa o número de elementos provenientes de  $\Pi_i$  classificados incorretamente em  $\Pi_l$  e  $n_i$  é o número de elementos da população  $\Pi_i$ .

Pode-se construir uma matriz, denominada *matriz de confusão*, em que são apresentados o número de observações de cada população classificadas correta ou incorretamente nas demais populações. Esta matriz está descrita na Tabela 1.

Tabela 1 Distribuição dos indivíduos segundo as populações de origem e classificação, em que  $n_{il}$ ,  $i \neq l$ , representa o número de elementos de  $\Pi_i$  classificados incorretamente em  $\Pi_l$  e  $n_{ii}$  o número de elementos classificados corretamente em  $\Pi_i$ .

População de origem	População classificada				Total
	$\Pi_1$	$\Pi_2$	$\dots$	$\Pi_m$	
$\Pi_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1k}$	$n_1$
$\Pi_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2k}$	$n_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\Pi_m$	$n_{m1}$	$n_{m2}$	$\dots$	$n_{mm}$	$n_m$
<b>Total</b>	$\hat{n}_1$	$\hat{n}_2$	$\dots$	$\hat{n}_m$	$n$

A estimativa da probabilidade global de acerto é dada por

$$P(\text{acerto global}) = \frac{\sum_{i=1}^m n_{ii}}{n}, \quad (2.73)$$

enquanto que a estimativa da probabilidade global de erro, complementar da probabilidade global de acerto, é também denominada taxa de erro aparente total (TEA) e pode ser descrita por

$$P(\text{erro global}) = TEA = \frac{\sum_{i=1}^m \sum_{\substack{l=1 \\ l \neq i}}^m n_{il}}{n}. \quad (2.74)$$

Existem alguns métodos utilizados para se estimar as probabilidades de classificação errônea, baseados nas frequências de ocorrência.

### 2.2.2.1 Método da ressubstituição

A partir de todos os elementos das amostras das  $m$  populações a função discriminante é estimada e usada para reclassificar os mesmos indivíduos destas amostras, cuja classificação é conhecida a priori. Por isto é possível avaliar quais elementos foram erroneamente classificados. Uma porcentagem alta de acerto na classificação dos elementos amostrais em relação à população a que de fato pertencem indica a boa qualidade da função discriminante.

Segundo Ferreira (2008) pelo fato de os mesmos indivíduos participarem da construção da função discriminante e da estimação dos erros de classificação, este método é considerado viciado e tende a subestimar a probabilidade total de classificação incorreta.

Uma forma de contornar este viés nas estimativas das probabilidades de erro é repartir o conjunto total de elementos  $n_1 + n_2 + \dots + n_m$  em duas partes com números diferentes de indivíduos. Uma fração maior é utilizada na construção da função discriminante (amostra de treinamento) e a outra, menor, para a estimação das probabilidades de erros de classificação (amostra de teste). Este

procedimento pode diminuir a acurácia do procedimento se as amostras não forem grandes, pois há uma redução do tamanho amostral original para a construção da regra de discriminação.

#### **2.2.2.2 Método de Lachenbruch**

Conhecido na literatura como método de validação cruzada ou pseudo *jackknife* de Lachenbruch e Mickey (1968), consiste nos passos descritos a seguir.

1. Retira-se um indivíduo do total de observados e utilizam-se as  $n_1 + n_2 + \dots + n_m - 1$  unidades amostrais restantes para construir a função discriminante.
2. Classifica-se o indivíduo retirado sob a regra construída, avaliando se a classificação foi correta ou não.
3. Retorna-se o indivíduo que foi retirado ao conjunto original e retira-se uma outra unidade amostral diferente da primeira, repetindo os passos 1 e 2.

Os três passos devem ser repetidos até que todos os  $n_1 + n_2 + \dots + n_m$  elementos amostrais sejam classificados e tenham as probabilidades de erro determinadas.

Segundo Mingotti (2005) as estimativas desse método são aproximadamente não viciadas e melhores que as do método da ressubstituição para populações normais e não normais.

### **2.3 Análise de sementes**

A análise de sementes de forma sistematizada surgiu no século XIX, quando problemas relacionados à qualidade das sementes e uso de práticas inescrupulosas

na comercialização destas estimularam a criação de Laboratórios de Análise de Sementes (LAS) para avaliar e definir padrões de qualidade de lotes de sementes.

No Brasil os LAS adotam as Regras para Análises de Sementes (RAS) (BRASIL, 2009), que contem métodos e procedimentos padrões para a realização das análises. As RAS são atualizadas de acordo com as regras prescritas pela Associação Internacional de Análise de Sementes (International Seed Testing Association - ISTA), incorporando a experiência e os avanços nacionais neste campo (NOVEMBRE, 2001).

Para se conhecer a qualidade de um lote de sementes é necessário avaliar resultados de diferentes tipos de análises, levando em consideração as peculiaridades de cada espécie. Pode-se considerar que o termo qualidade resulta da observação dos aspectos genéticos, físicos, fisiológicos e sanitários. Aspectos genéticos contêm as características específicas atribuídas a cada cultivar e têm influência dominante na produtividade da colheita. A qualidade física diz respeito tanto à composição dos lotes, como também à condição física da semente: tamanho, cor, teor de água, densidade, injúrias mecânicas e causadas por insetos, e uniformidade quanto a essas características. A qualidade fisiológica indica a capacidade da semente de desempenhar funções vitais, sendo caracterizada pelo poder germinativo, pelo vigor e pela longevidade. A qualidade sanitária indica a condição da semente quanto à presença e grau de ocorrência de insetos e microrganismos patogênicos, tais como fungos, bactérias, nematoides ou vírus, que podem afetar o desenvolvimento das plantas e a produtividade das culturas (EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA, 2005).

Os testes, ou o desenvolvimento de métodos para avaliação da qualidade de sementes são processos dinâmicos que devem ser revistos periodicamente para a inclusão de avanços na área de sementes.

Os testes realizados em laboratórios mais utilizados são os de germinação e pureza. Esses testes apresentam limitações, já que nem sempre refletem os resultados de campo. Em espécies que apresentam o pericarpo coriáceo, como no caso de aquênios de girassol, as estruturas internas podem estar deformadas, danificadas ou vazias, o que interfere nos resultados dos testes. A análise de imagens radiográficas de estruturas internas de sementes vem sendo cada vez mais utilizada e muitos estudos estão sendo desenvolvidos para validar as contribuições que este tipo de análise traz à tecnologia de sementes.

### **2.3.1 Teste de raios X**

O uso de imagens radiográficas na avaliação de sementes é um método recomendado pela International Seed Testing Association - ISTA (2011) que possibilita a visualização das estruturas internas do embrião, permitindo distinguir sementes bem formadas de sementes vazias, com danos mecânicos ou com ataques de insetos, além de criar um arquivo fotográfico das sementes analisadas (BRASIL, 2009).

É um método rápido, que não exige tratamentos prévios das sementes. Por ser usada radiação de baixo nível na aquisição das imagens, o teste não compromete a viabilidade das sementes (SIMAK; GUSTAFSSON, 1953), sendo considerado um método não destrutivo. Permite que as sementes radiografadas sejam também submetidas a testes fisiológicos, estabelecendo relações entre os danos mecânicos ou alterações observadas internamente nas sementes e os prejuízos causados para a germinação (CICERO et al., 1998). Desta forma pode-se correlacionar os dados de danos internos com o comportamento fisiológico das sementes.

O uso de raios X iniciou-se na Suécia no estudo de sementes de alguns tipos de coníferas e permitiu a identificação de anormalidades no embrião destas

sementes (SIMAK; GUSTAFSSON, 1953). O princípio da técnica consiste na absorção de raios X em diferentes quantidades pelos diferentes tecidos das sementes, o que depende da espessura, da densidade e da composição desses tecidos, além do comprimento de onda da radiação (BINO; AARTSE; BURG, 1993; ISTA, 2011). Neste sentido um dos parâmetros observado é o tamanho do embrião, que pode ser avaliado pelo seu grau de desenvolvimento e pelo espaço livre na cavidade interna da semente (MARCOS FILHO et al., 2010).

Em estudos com sementes de pimentão (*Capsicum annuum* L.), Dell' Aquila (2007) observou que sementes com área de espaços livres entre o embrião e o endosperma superior a 2,7% apresentaram redução progressiva da formação de plântulas normais.

Pinto et al. (2009) observaram que sementes de pinhão-manso (*Jatropha curcas* L.) que apresentaram nas imagens radiográficas manchas escuras em mais de 50% do endosperma e, ou no embrião, não germinaram ou originaram plântulas anormais.

Em espécies florestais de *Lauraceae* danos observados no embrião ou má formação do tecido cotiledonar resultaram em sementes mortas no teste de germinação (CARVALHO; CARVALHO; DAVIDE, 2009).

Também em estudos com mamona (*Ricinus communis* L.) tecidos que resultaram em imagens translúcidas, embriões deformados e tecidos com menos de 50% de endosperma ou manchados afetaram negativamente o potencial fisiológico dos lotes de sementes (CARVALHO; ALVES; OLIVEIRA, 2010).

Nas imagens radiográficas de sementes de berinjela (*Solanum melongena* L.) foi observado que danos mecânicos ou presença de tecidos deteriorados nas sementes afetaram negativamente o potencial fisiológico, porém a presença de maior área ocupada pelo embrião e endosperma não favoreceu a germinação (SILVA;

CICERO; BENNETT, 2012).

A utilização da análise de imagens geradas por raios X segundo Gonçalves (2012) e Luz et al. (2010) é eficiente para identificar danos internos em sementes de girassol que afetam negativamente sua qualidade fisiológica. Sementes de girassol classificadas como mal formadas, translúcidas e vazias tem a sua germinação, velocidade de emergência e emergência final prejudicadas.

Segundo Rocha (2012) as análises de imagens de raios X permitiram identificar danos mecânicos, má formação do embrião e tecidos deteriorados em sementes de girassol que podem ser relacionados com a presença de plântulas normais, anormais e sementes mortas no teste de primeira contagem de germinação.

Apesar dos inúmeros resultados já obtidos, uma preocupação atual quanto ao uso do teste de raios X é a automatização da análise e das medições visando estabelecer maior precisão e eliminar interpretações subjetivas. Esforços vêm sendo direcionados neste sentido e já existem softwares desenvolvidos para a análise automática de imagens de raios X digitalizadas. Entre os trabalhos citados o estudo de Dell'Aquila (2007) e Silva, Cicero e Bennett (2012) utilizaram o Image Pro Plus<sup>®</sup> (Media Cybernetics<sup>®</sup>, EUA), software que determina a área da semente, a área de espaço livre e a razão entre estas áreas.

A utilização do teste de raios X tem demonstrado ser uma técnica que contribui significativamente na análise de sementes, porém ainda se encontra restrita ao âmbito da pesquisa e não totalmente incorporada à rotina dos laboratórios envolvidos em programas de controle de qualidade (CICERO, 2010).

### **2.3.2 Teste de germinação**

O teste de germinação é o método mais utilizado para determinar o potencial máximo de germinação de um lote de sementes, servindo para comparar a

qualidade de diferentes lotes e estimar o valor para semeadura em campo. Normalmente é desenvolvido em laboratórios, onde é possível controlar condições de fatores externos, permitindo que a germinação se desenvolva de maneira mais regular, rápida e completa (BRASIL, 2009). Entre os principais fatores controlados estão disponibilidade de água, temperatura, oxigênio, luz (MARCOS FILHO, 2005) e substrato. Estas condições variam para cada espécie e são padronizadas para que os resultados dos testes de germinação possam ser reproduzidos e comparados, dentro de limites tolerados pelas RAS.

Avaliar a germinação num teste de laboratório significa verificar se houve a emergência e desenvolvimento de estruturas essenciais do embrião, que demonstram que este é capaz de produzir uma planta normal sob condições favoráveis de campo.

Na plântula de girassol o sistema apical consiste do hipocótilo alongado e dois cotilédones com broto terminal situado entre eles. Não existe epicótilo alongado dentro do período do teste; epicótilo e broto terminal não são perceptíveis. O sistema radicular consiste da raiz primária, geralmente com raízes fasciculadas, as quais podem ser bem desenvolvidas. Raízes secundárias podem ocasionalmente desenvolver durante o período do teste, mas elas não são consideradas na avaliação como plântula normal.

A partir da observação do desenvolvimento destas estruturas as plântulas são classificadas em normais ou anormais. As sementes que não germinam podem ser sementes duras, dormentes ou mortas (BRASIL, 2009).

O teste de germinação apresenta algumas limitações, como a demora na obtenção dos resultados e a detecção do nível de deterioração da semente em estádios mais avançados. Ainda, frequentemente, seus resultados não se correlacionam com a emergência em campo, onde as condições nem sempre são favoráveis.

É necessário, portanto, que as informações provenientes do teste de germinação sejam complementadas com outros testes a fim de melhor avaliar e informar sobre a qualidade das sementes.

### 3 MATERIAL E MÉTODOS

#### 3.1 Obtenção e processamento dos dados

O trabalho de obtenção dos dados foi desenvolvido no Laboratório de Análise de Sementes (LAS) do Departamento de Agricultura da Universidade Federal de Lavras. Os dados foram obtidos a partir de uma amostra de sementes de girassol (*Helianthus annuus* L.), cultivar Hélio-250, produzidas em Uberlândia - MG, safra 2010/2011.

Foram selecionadas aleatoriamente 600 sementes para serem radiografadas sem nenhum tipo de preparo especial. Presas com fita adesiva dupla face, subamostras de 25 sementes foram arranjadas todas numa mesma posição em lâminas transparentes e numeradas segundo a localização na lâmina (linha e coluna) de modo a poderem ser identificadas nas classificações e testes posteriores. As placas foram radiografadas com intensidade de 22 kV e tempo médio de 11 segundos de exposição em aparelho de raios-X Faxitron MX20, gerando imagens digitalizadas salvas no formato jpeg.

Em seguida as sementes radiografadas foram submetidas ao teste de germinação, seguindo os padrões estabelecidos pelas Regras para Análise de Sementes (BRASIL, 2009), na temperatura de 25 °C e umidade igual a duas vezes o peso do papel do substrato. No quarto e décimo dias após a semeadura foi efetuada a contagem e classificação das sementes que germinaram e não germinaram.

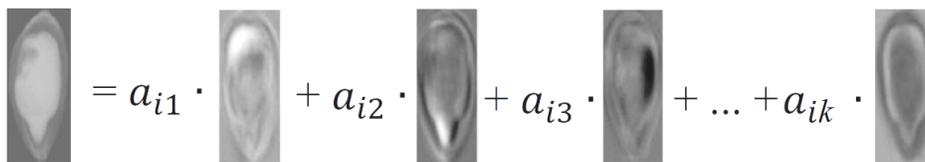
As imagens radiográficas foram analisadas visualmente por especialista da área de acordo com a morfologia interna e classificadas nos seguintes grupos: sementes cheias, com embrião opaco na análise radiográfica; sementes com danos leves, mas com preservação das características do eixo embrionário; sementes com danos graves ou deformadas, com danos afetando o eixo embrionário. As sementes

com danos graves, por serem em número reduzido, foram excluídas das análises iniciais.

Foram selecionadas manualmente 300 imagens retangulares de sementes individuais. O conjunto abrangia 100 sementes classificadas como cheias, 100 classificadas como sementes com danos leves e 100 como sementes deformadas. Por apresentarem tamanhos diferentes, as imagens foram padronizadas para o tamanho médio de todas as imagens de  $121 \times 268$  pixels. O processamento das imagens foi realizado com o Image Processing Toolbox do MATLAB<sup>®</sup>.

### 3.2 Aplicação de ICA na extração de características

A ideia básica que justifica a aplicação de ICA é considerar que uma imagem radiográfica de semente seja resultado da mistura de um conjunto de imagens-base, que são informações independentes comuns a todas as radiografias de sementes de diferentes níveis de qualidade. Cada imagem-base contribui com algum tipo de informação, tais como forma, grau de preenchimento e diferentes tipos de dano em regiões específicas da semente. Portanto, uma imagem analisada pode ser decomposta em uma combinação linear destas imagens-base (fontes ou componentes independentes) e a cada imagem corresponde um vetor cujos elementos são as coordenadas dos componentes independentes na mistura, como na Figura 5. Tais vetores são as características extraídas das imagens a partir da ICA.



The figure shows a mathematical equation where a grayscale image of a seed on the left is equal to a sum of several terms. Each term consists of a coefficient (a<sub>i1</sub>, a<sub>i2</sub>, a<sub>i3</sub>, ..., a<sub>ik</sub>) multiplied by a corresponding grayscale image of a seed. The images of the base seeds show different internal features, such as varying degrees of darkness and shape, representing independent components of the seed's appearance.

$$\text{Image} = a_{i1} \cdot \text{Image}_1 + a_{i2} \cdot \text{Image}_2 + a_{i3} \cdot \text{Image}_3 + \dots + a_{ik} \cdot \text{Image}_k$$

Figura 5 Imagem de semente como combinação linear de imagens-base

Para a estimação ICA foi utilizado o FastICA (HYVÄRINEN; OJA, 1997) com pacote implementado no MATLAB<sup>®</sup> (FASTICA..., 2005) e a entrada do algoritmo é a única informação disponível: as imagens radiográficas das sementes.

Cada imagem de semente é uma matriz de tamanho  $121 \times 268$ , cujas entradas são os valores numéricos que correspondem às intensidades de cinza de cada pixel. Divididas em 256 categorias, variando entre 0 e 255 tons de cinza, as imagens foram normalizadas para apresentarem valores entre 0 (preto) e 1 (branco). Não foi realizado nenhum outro tipo de operação morfológica na imagem (processos de segmentação) que poderia evidenciar algumas características desejadas (GONZALEZ; WOODS, 2003). Cada imagem retangular foi transformada num vetor linha de  $1 \times 32.428$  pixels, concatenando as colunas, que compôs a matriz de dados  $\mathbf{X}$ . Portanto, cada linha da matriz de dados é uma imagem de semente, a variável observada do modelo ICA. Sejam  $n$  o número de linhas da matriz (número de imagens) e  $p$  o número de colunas (realizações de cada imagem, representadas pelo valor de pixel normalizado). Neste caso, tem-se  $n = 300$  e  $p = 32.428$ .

Seguindo o modelo ICA (2.2) segue-se que

$$\mathbf{X}_{(n \times p)} = \mathbf{A}_{(n \times n)} \cdot \mathbf{S}_{(n \times p)} \quad (3.1)$$

O próprio algoritmo FastICA faz os pré-processamentos de centralizar e branquear os dados de entrada. Na fase de branqueamento é feita a redução de dimensão escolhendo-se um número de componentes principais, que será também o número de componentes independentes,  $k < n$  que preserve um bom percentual de variabilidade dos dados. Como nas equações (2.2) e (2.3), considerando a redução de dimensão tem-se

$$\mathbf{X}_{(n \times p)} \approx \mathbf{A}_{(n \times k)} \cdot \mathbf{S}_{(k \times p)} \quad (3.2)$$

e

$$\mathbf{Y}_{(k \times p)} \approx \mathbf{W}_{(k \times n)} \cdot \mathbf{X}_{(n \times p)}, \quad (3.3)$$

sendo  $\mathbf{W}_{(k \times n)}$  a matriz inversa generalizada de  $\mathbf{A}_{(n \times k)}$ .

O FastICA foi aplicado considerando  $k = 2$ ,  $k = 20$ ,  $k = 30$ ,  $k = 45$ ,  $k = 60$  e  $k = 90$ , quantidades de componentes independentes que explicam percentuais diferentes da variabilidade dos dados. O FastICA foi executado escolhendo-se a função de estimação cúbica (2.44), aproximação da curtose, e a ortogonalização deflacionária.

O algoritmo fornece como resultados a matriz de mistura estimada  $\hat{\mathbf{A}}$ , a matriz de separação estimada  $\hat{\mathbf{W}}$  e a estimativa dos componentes independentes  $\hat{\mathbf{Y}}$ .

Após a estimação dos componentes independentes foi feita a reconstrução das imagens das sementes  $\hat{\mathbf{X}}_{n \times p}$ , dada por

$$\hat{\mathbf{X}}_{(n \times p)} = \hat{\mathbf{A}}_{(n \times k)} \cdot \hat{\mathbf{Y}}_{(k \times p)}. \quad (3.4)$$

### 3.3 Classificação por análise discriminante

A técnica de análise discriminante é usada na classificação das imagens radiográficas das sementes segundo os níveis de qualidade física. Os dados de entrada da análise discriminante são as características das imagens extraídas por ICA; ou seja, cada imagem de semente passa a ser representada pelo vetor de pesos ou coeficientes da combinação linear dos componentes independentes, a correspondente linha da matriz de mistura estimada  $\hat{\mathbf{A}}_{(n \times k)}$ .

A partir deste conjunto de características, correspondente às 300 imagens previamente classificadas mediante análise visual, foram construídas as funções discriminantes de Fisher e quadrática, aplicando-se os métodos descritos na seção

2.2. Antes, porém, o conjunto de dados foi submetido aos testes de normalidade multivariada Shapiro-Wilk e teste Box's M que pressupõe homogeneidade das matrizes de covariância dos grupos (FERREIRA, 2008) para verificar se este atendia às pressuposições de uso destas funções discriminantes.

A técnica foi aplicada considerando diferentes dimensões dos dados extraídos pela ICA ( $k = 20, k = 30, k = 45, k = 60, k = 90$ ). A validação da técnica foi realizada segundo o método de Lachenbruch (validação cruzada) e foram estimadas as probabilidades totais de acerto e erro da classificação realizada pela análise discriminante.

Uma nova imagem radiográfica de semente poderá ser classificada em um dos grupos previamente definidos, usando as funções discriminantes. A partir da base de componentes independentes obtida com o conjunto das 300 imagens, as coordenadas da decomposição ICA desta nova imagem são estimadas fazendo

$$\hat{\mathbf{A}}_{(1 \times k)} = \mathbf{X}_{(1 \times p)} \cdot \hat{\mathbf{Y}}_{(k \times p)}^{-}, \quad (3.5)$$

em que  $\mathbf{X}_{(1 \times k)}$  é o vetor linearizado de pixels da imagem e  $\hat{\mathbf{Y}}_{(k \times p)}^{-}$  é a matriz inversa generalizada dos componentes independentes estimados.

#### 4 RESULTADOS E DISCUSSÃO

A apresentação destes primeiros resultados serve de base para os estudos que foram posteriormente desenvolvidos nos artigos. Aborda o problema da escolha do número de componentes independentes e avalia o desempenho das funções discriminantes de Fisher e quadrática, definindo a mais adequada para utilização com este tipo de dados.

A extração de parâmetros por ICA está associada à redução da dimensão dos dados com a aplicação de PCA. O número de componentes independentes foi variado de modo a se poder avaliar a qualidade da recomposição das imagens originais em função da quantidade de imagens-base escolhidas, assim como a qualidade da classificação a partir de dados de entrada com diferentes dimensões. A depender do número  $k$  de componentes escolhido preserva-se um percentual da variabilidade dos dados originais, conforme mostrado na Tabela 2.

Tabela 2 Percentual de variabilidade explicada a partir do número  $k$  de componentes independentes (IC's) estimados

Nº IC's ( $k$ )	Variabilidade explicada (%)
2	90,42
20	97,47
30	98,20
45	98,80
60	99,13
90	99,50

Convém observar que o primeiro autovalor da matriz de covariância dos dados já responde por 88,17% de variabilidade, o que justifica porque a escolha de apenas dois componentes independentes corresponde a 90,42% de explicação da variabilidade. Observando as duas imagens-base representativas destes compo-

nentes (Figura 6), vê-se que estas definem basicamente o formato da semente (com algumas variações na sua parte interna) e o fundo da imagem; mas este percentual elevado não é capaz de identificar diferenças relativas a danos e a deformações como se verifica ao se reconstruir a imagem a partir apenas destes dois componentes (Figura 8, coluna B). Na Figura 7 são mostradas as imagens-base geradas pela estimação de 30 componentes independentes. Percebe-se que cada imagem-base contribui com alguma modificação numa região diferente da semente.

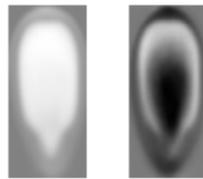


Figura 6 Imagem de 2 componentes independentes estimados por ICA

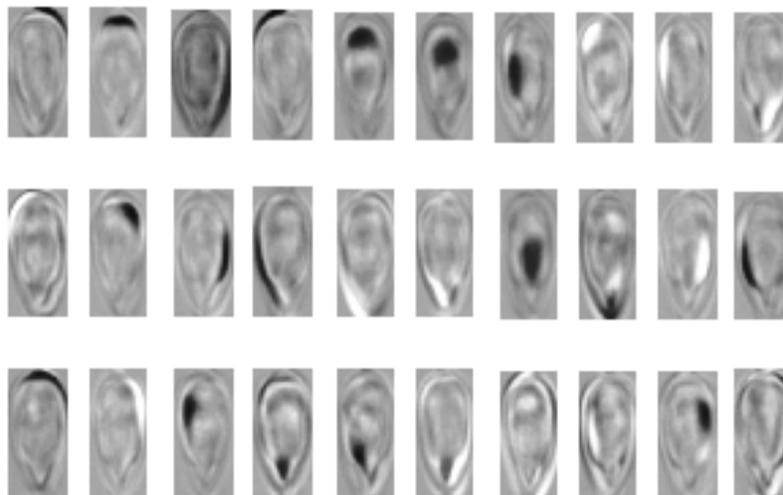


Figura 7 Imagem de 30 componentes independentes estimados por ICA

Na Figura 8 tem-se imagens originais de sementes de diferentes níveis de qualidade (coluna A) e nas colunas seguintes (B a F) a reconstrução destas imagens usando bases de componentes independentes de diferentes dimensões  $k$ , a saber, respectivamente,  $k = 2$ ,  $k = 20$ ,  $k = 30$ ,  $k = 45$  e  $k = 60$ . Nas linhas (de cima para baixo) as sementes seguem a seguinte classificação: cheia, com dano leve e deformada. Observa-se que a partir da estimação de 30 IC's já é possível distinguir mais detalhes nas imagens reconstruídas (coluna D), inclusive identificando o dano leve da imagem da segunda linha.

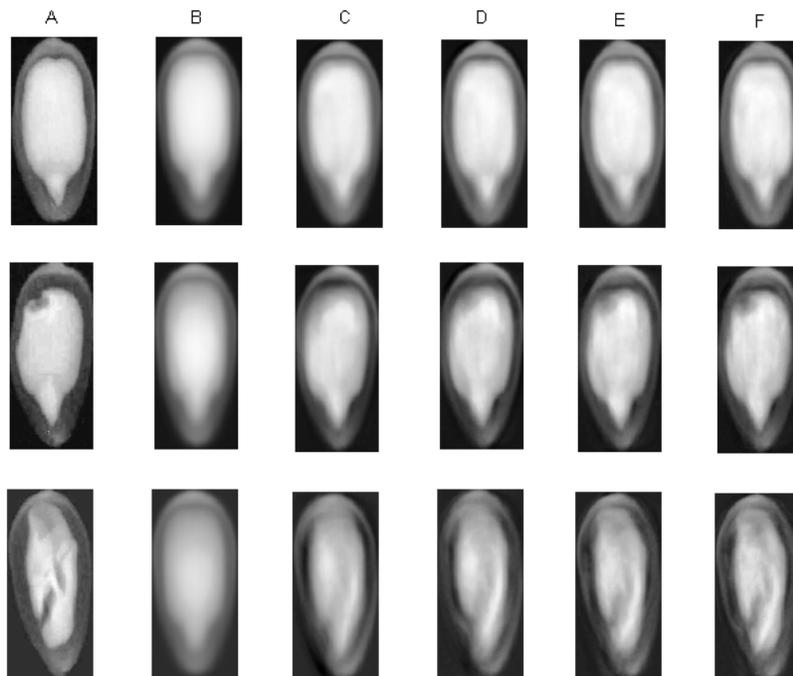


Figura 8 Reconstrução de imagens de sementes de girassol com diferentes níveis de qualidade física, a partir de bases de componentes independentes de diferentes dimensões. (A) imagem original; (B) reconstruída com 2 IC's; (C) reconstruída com 20 IC's; (D) reconstruída com 30 IC's; (E) reconstruída com 45 IC's; (F) reconstruída com 60 IC's

Os testes de aplicação da análise discriminante foram realizados a partir das características extraídas pela ICA do grupo das 300 imagens previamente classificadas, considerando os vetores de entrada com dimensão  $k = 20$ ,  $k = 30$ ,  $k = 45$ ,  $k = 60$  e  $k = 90$ . Para dados de dimensão  $k = 2$ , como se percebe pela reconstrução das imagens, não se consegue fazer a discriminação nos grupos.

O teste de normalidade multivariada Shapiro-Wilk foi realizado para cada grupo de classificação previamente definido ao nível de 5% de significância e obteve-se valor- $p$  nulo, indicando que nos diferentes grupos os dados não possuem distribuição gaussiana. Também obteve-se valor- $p$  nulo no teste Box's M indicando que os grupos não possuem matrizes de covariância iguais. Neste caso, tanto a função quadrática como a função discriminante linear de Fisher tiveram as suas pressuposições violadas e por isso ambas foram testadas para se avaliar a que apresentava melhor desempenho.

Na Tabela 3 são apresentadas as estimativas da probabilidade global de acerto (2.73) de classificação ao ser aplicado o método de Lachenbruch para avaliar o desempenho das funções discriminantes, conforme descrito na seção 2.2.2, usando vetores de entrada com diferentes dimensões.

Tabela 3 Probabilidade global de acerto de classificação das funções discriminantes de Fisher e quadrática, usando dados de diferentes dimensões

Função	Dimensão dos dados de entrada				
	$k = 20$	$k = 30$	$k = 45$	$k = 60$	$k = 90$
discriminante					
Fisher	0,6867	0,6733	0,6900	0,6767	0,6833
Quadrática	0,7433	0,7333	0,7267	0,6933	0,5533

Considerando as regras de classificação, o desempenho da função discrimi-

minante quadrática foi na grande maioria das vezes superior ao da função discriminante linear de Fisher. Em relação à dimensão dos dados, observa-se que os resultados globais ao se usar vetores com dimensão 20 e 30 foram um pouco superiores e são apresentados de forma mais detalhada nas Tabelas 4 e 5. Segundo Mingotti (2005, p. 240), “um equívoco muito comum é o de pensar que o aumento do número de variáveis-resposta aumenta a capacidade de discriminação”. Como se verifica na classificação realizada pela função discriminante quadrática, ocorre exatamente uma diminuição na probabilidade de acerto com o aumento na dimensão dos dados.

Tabela 4 Quantidade de sementes classificadas em cada categoria usando a função discriminante quadrática para dados de dimensão 20

População de origem	População classificada			Probabilidade de erro
	cheia	deformada	dano leve	
cheia	82	2	16	0,18
deformada	0	89	11	0,11
dano leve	33	15	52	0,48
Total	115	106	79	

Tabela 5 Quantidade de sementes classificadas em cada categoria usando a função discriminante quadrática para dados de dimensão 30

População de origem	População classificada			Probabilidade de erro
	cheia	deformada	dano leve	
cheia	75	2	23	0,25
deformada	0	87	13	0,13
dano leve	27	15	58	0,42
Total	102	104	94	

Apesar de a probabilidade global de acerto ter sido maior em dados de dimensão 20, vê-se que os erros de classificação das imagens com dimensão 30 estão melhor distribuídos. Com o aumento da dimensão de 20 para 30 os erros de classificação de sementes cheias aumentam, enquanto que diminuem os erros de classificação das sementes com danos leves. Observando-se os totais de imagens classificadas em cada categoria os resultados de dimensão 30 estão mais próximos dos totais originais de 100 sementes de cada categoria. Pode-se considerar que as sementes deformadas foram bem identificadas, o que é um resultado importante, pois este tipo de semente possui menor potencial de germinação.

## 5 CONSIDERAÇÕES GERAIS

Nesta primeira parte foram apresentadas a teoria da análise de componentes independentes e da análise discriminante, técnicas utilizadas neste trabalho. Também foram apresentados aspectos importantes na análise de sementes, em especial a utilização de imagens de raios X, objeto deste estudo.

A apresentação destes primeiros resultados objetivou abordar de forma detalhada aspectos que não foram explicitados nos artigos, justificando a escolha do número de componentes independentes e da função discriminante. Apesar de ainda não apresentarem números excelentes, estes primeiros resultados mostraram que a ICA é uma técnica adequada na extração de característica de raios X de sementes de girassol. Aplicada a um conjunto de 300 imagens de diferentes tipos de sementes, verificou-se que o número de 30 componentes independentes indica um tamanho de base suficiente à obtenção de uma classificação satisfatória. A análise discriminante também mostrou-se uma técnica viável para a classificação das sementes, sendo a função discriminante quadrática a mais adequada para utilização com este tipo de dados.

A partir destas informações a metodologia utilizada foi testada na classificação de um novo conjunto de sementes usando as mesmas bases de IC's e a função discriminante quadrática obtida com o conjunto das 300 imagens de sementes, como será mostrado nos artigos a seguir.

## REFERÊNCIAS

AMARI, S. I. Natural gradient works efficiently in learning. **Neural Computation**, Cambridge, v. 10, n. 2, p. 251-276, 1998.

AMARI, S. I.; CICHOCKI, A.; YANG, H. H. A new learning algorithm for blind signal separation. In: TOURETZKY, D., MOZER M; HASSELMO M. (Ed). **Advances in neural information processing systems**. Cambridge: MIT, 1996. p. 757-763.

BACK, A. D.; WEIGEND, A. S. A first application of Independent Component Analysis to extracting structure from stock returns. **International Journal of Neural Systems**, Singapore, v. 8, n. 4, p. 473-484, Aug. 1997.

BELL, A. J.; SEJNOWSKI, T. J. An information-maximization approach to blind separation and blind deconvolution. **Neural Computation**, Cambridge, v. 7, n. 6, p. 1129-1159, 1995.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Regras para análise de sementes**. Brasília, 2009. 399 p.

BINO, R. J.; AARTSE, J. W.; BURG, W. J. van der. Non-destructive X-ray analysis of Arabidopsis embryo mutants. **Seed Science Research**, Wallingford, v. 3, n. 3, p. 167-170, 1993.

CAMPOS, L. F. A.; BARROS, A. K.; SILVA, A. C. Independent component analysis and neural networks applied for classification of malignant, benign and normal tissues in digital mammography. **Methods of Information in Medicine**, Stuttgart, v. 46, n. 2, p. 212-215, 2007.

CARDOSO, J. F. Source separation using higher order moments. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 89., 1989, Glasgow. **Proceedings...** Glasgow: IEEE, 1089. p. 2109-2112. Disponível em: <<http://perso.telecom-paristech.fr/cardoso/Papers.PDF/icassp89.pdf>>. Acesso em: 15 maio 2012.

CARDOSO, J. F.; SOULOUMIAC, A. Blind beamforming for non Gaussian signals. **IEE Proceedings - Part F**, London, v. 140, n. 6, p. 362-370, Dec. 1993.

CARVALHO, L. R.; CARVALHO, M. L. M.; DAVIDE, A. C. Utilização do teste de raios X na avaliação da qualidade de sementes de espécies florestais de Lauraceae. **Revista Brasileira de Sementes**, Londrina, v.31, n. 4, p.57-66, 2009.

CARVALHO, M. L. M.; ALVES, R. A.; OLIVEIRA, L. M. Radiographic analysis in castor bean seeds (*Ricinus communis* L.). **Revista Brasileira de Sementes**, Londrina, v. 32, n. 1, p. 170-175, jan. 2010.

CHRISTOYIANNI, I. et al. Computer aided diagnosis of breast cancer in digitized mammograms. **Computerized Medical Imaging and Graphics**, New York, v. 26, n. 5, p. 309-319, Sept. 2002.

CICERO, S. M. Aplicação de imagens radiográficas no controle de qualidade de sementes. **Informativo ABRATES**, Londrina, v. 20, n. 3, p. 48-51, 2010.

Disponível em:

<<http://www.abrates.org.br/portal/images/stories/informativos/v20n3/minicurso02.pdf>>. Acesso em: 14 set. 2011.

CICERO, S. M. et al. Evaluation of mechanical damage in seeds of maize (*Zea mays* L.) by X-ray and digital imaging. **Seed Science and Technology**, Zürich, v. 26, n. 3, p.603-612, 1998.

CICHOCKI, A.; UNBEHAUEN, R. Robust neural networks with on-line learning for blind identification and blind separation of sources. **IEEE Transactions on Circuits and Systems**, New York, v. 43, n. 11, p. 894-906, 1996.

COMON, P. Independent component analysis, a new concept? **Signal Processing**, Amsterdam, v. 36, n. 3, p. 287-314, 1994.

COSTA, D. D. et al. Independent component analysis in breast tissues mammograms images classification using LDA and SVM. In: INTERNATIONAL SPECIAL TOPIC CONFERENCE, 6., 2007, Boca Raton. **Proceedings...** Boca Raton: ITAB, 2007. p. 231-234.

COVER, T. M.; THOMAS, J. A. **Elements of information theory**. New York: J. Wiley, 1991. 542 p.

DELL'AQUILA, A. Pepper seed germination assessed by combined X-radiography and computer-aided imaging analysis. **Biologia Plantarum**, Prague, v. 51, n. 4, p. 777-781, Dec. 2007.

DÉNIZ, O.; CASTRILLÓN, M.; HERNÁNDEZ, M. Face recognition using independent component analysis and support vector machines. **Pattern Recognition Letters**, Amsterdam, v. 24, n. 13, p. 2153-2157, Sept. 2003.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Glossário**. Santo Antônio de Goiás, 2005. Disponível em:  
<<http://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Feijao/FeijaoIrrigadoNoroesteMG/glossario.htm>>. Acesso em: 9 maio 2013.

FASTICA for MATLAB 7.x and 6.x. Version 2.5. Chicago, 2005. Disponível em:  
<<http://research.ics.aalto.fi/ica/fastica/>>. Acesso em: 6 maio 2011.

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 662 p.

FISHER, R. A. The utilization of multiple measurements in taxonomic problems. **Annals of Eugenics**, London, v. 7, p. 179-188, 1936.

GARCÍA-FERRER, A.; GONZÁLEZ-PRIETO, E.; PEÑA, D. **A multivariate generalized independent factor GARCH model with an application to financial stock returns**. Madri: Universidad Carlos III, 2008. (Statistics and Econometrics Series, 28). Disponível em:  
<<http://ideas.repec.org/p/cte/wsrepe/ws087528.html>>. Acesso em: 25 mar. 2011.

GONÇALVES, N. R. **Qualidade de sementes de girassol no beneficiamento**. 2012. 66 p. Dissertação (Mestrado em Fitotecnia) - Universidade Federal de Lavras, Lavras, 2012. Disponível em:  
<<http://repositorio.ufla.br/jspui/bitstream/1/343/1/DISSERTA%C3%87%C3%83O%20Qualidade%20de%20sementes%20de%20girassol%20no%20beneficiamento.pdf>>. Acesso em: 8 maio 2013.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. São Paulo: E. Blücher, 2003. 509 p.

HAYKIN, S. **Neural networks: a comprehensive foundation**, 2<sup>nd</sup> ed. New Jersey: Prentice Hall, 1999. 842 p.

HERAULT, J.; JUTTEN, C. Space or time adaptive signal processing by neural network models: neural networks for computing. In: CONFERENCE OF THE AMERICAN INSTITUTE OF PHYSICS, 1., 1986, New York. **Proceedings...** New York: AIP, 1986. p. 206-211.

HERAULT, J.; JUTTEN, C.; ANS, B. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In: XÈME COLLOQUE GRETSI, 1., 1985, Paris. **Actes...** Paris: GRETSI, 1985. p. 1017-1022.

HYVÄRINEN, A. Survey on independent component analysis. **Neural Computing Surveys**, George, v. 2, n. 4, p. 94-128, 1999.

HYVÄRINEN, A.; HOYER, P. O. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. **Neural Computation**, Cambridge, v. 12, n. 7, p. 1705-1720, 2000.

HYVÄRINEN, A.; KARHUNEN J.; OJA E. **Independent component analysis**. New York: J. Wiley, 2001. 481 p.

HYVÄRINEN, A.; OJA, E. A fast fixed-point algorithm for independent component analysis. **Neural Computation**, Cambridge, v. 9, n.7, p. 1483-1492, 1997.

\_\_\_\_\_. Independent component analysis: algorithms and applications. **Neural Networks**, New York, v. 13, n. 4/5, p. 411-430, 2000.

INTERNATIONAL SEED TESTING ASSOCIATION. **International rules for seed testing Association**. Bassersdorf, 2011. 174 p.

JONES, M.; SIBSON, R. What is projection pursuit? **Journal of the Royal Statistical Society, Serie A**, London, v. 150, p. 1-36, 1987.

JOURNÉE, M. **Matlab project independent component analysis**. Liège: University of Liège, 2008. Disponível em <[http://www.inma.ucl.ac.be/~absil/Grenoble2008/Grenoble\\_Matlab\\_project.pdf](http://www.inma.ucl.ac.be/~absil/Grenoble2008/Grenoble_Matlab_project.pdf)> Acesso em: 18 ago. 2011.

KARHUNEN, J.; PAJUNEN, P.; OJA, E. The nonlinear PCA criterion in blind source separation: relations with other approaches. **Proceedings of Neurocomputing**, London, v. 22, n. 1, p. 5-20, Nov. 1998.

KIVILUOTO, K.; OJA, E. Independent component analysis for parallel financial time series. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION, 5., 1998, Kitakyushu. **Proceedings...** Kitakyushu: IOA, 1998. p. 895-898.

KOBORI, N. N.; CICERO, S. M.; MEDINA, P. F. Teste de raios X na avaliação da qualidade de sementes de mamona. **Revista Brasileira de Sementes**, Londrina, v. 34, n. 1, p. 125-133, 2012.

LACHENBRUCH, P.; MICKEY, M. Estimation of error rates in discriminant analysis. **Technometrics**, Washington, v. 27, n. 2, p. 189-198, Mar. 1968.

LU, C. J. Integrating independent component analysis-based denoising scheme with neural network for stock price prediction. **Expert Systems with Applications**, New York, v. 37, n. 10, p. 7056-7064, Oct. 2010.

LU, C. J.; LEE, T. S.; CHIU, C. C. Financial time series forecasting using independent component analysis and support vector regression. **Decision Support Systems**, Amsterdam, v. 47, n. 2, p. 115-125, May 2009.

LUZ, R. P. et al. Análise de imagens radiográficas na avaliação da qualidade de sementes de girassol. In: CONGRESSO DE PÓS-GRADUAÇÃO DA UFLA, 19., 2010, Lavras. **Anais...** Lavras: UFLA, 2010. Disponível em: <<http://www.sbpnet.org.br/livro/lavras/resumos/1146.pdf>>. Acesso em: 14 set.

2011.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. 2. ed. São Paulo: EDUSP, 2006. 428 p.

MAKEIG, S. et al. Independent component analysis of electroencephalographic data. In: TOURETZKY, D.; MOZER, M.; HASSELMO, M. (Ed.). **Advances in neural information processing systems**. Cambridge: MIT, 1996. p. 145-151.

MANLY, B. F. J. **Métodos estatísticos multivariados**: uma introdução. 3. ed. Porto Alegre: Bookman, 2008. 229 p.

MARCOS FILHO, J. **Fisiologia de sementes de plantas cultivadas**. Piracicaba: FEALQ, 2005. 495 p.

MARCOS FILHO, J. et al. Using tomato analyzer software to determine embryo size in X-rayed seeds. **Revista Brasileira de Sementes**, Londrina, v. 32, n. 2, p. 146-153, 2010.

MINGOTTI, S. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 297 p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3<sup>rd</sup> ed. New York: McGraw-Hill, 1974. 564 p.

NOVEMBRE, A. D. L. C., Avaliação da qualidade de sementes. **Revista SEED News**, Pelotas, v. 5, n. 3, 2001. Disponível em: <[http://www.seednews.inf.br/portugues/seed53/print\\_artigo53.html](http://www.seednews.inf.br/portugues/seed53/print_artigo53.html)>. Acesso em: 14 abr. 2011.

PINTO, T. L. F. et al. Avaliação da viabilidade de sementes de pinhão manso pelos testes de tetrazólio e de raios X. **Revista Brasileira de Sementes**, Londrina, v. 31, n. 2, p.195-201, 2009.

PRASAD, M. N.; SOWMYA, A.; KOCH, I. Feature subset selection using ICA

for classifying emphysema in HRCT images. In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 17., 2004, Cambridge. **Proceedings...** Cambridge: ICPR, 2004. p. 515-518.

ROCHA, C. R. M. **Avaliação da qualidade de sementes de girassol por meio de análise de imagens**. 2012. 68 p. Dissertação (Mestrado em Fitotecnia) - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 2012. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/11/11136/tde-09082012-083733/>>. Acesso em: 19 maio 2013.

SILVA, V. N.; CICERO, S. M.; BENNETT, M. Relationship between eggplant seed morphology and germination. **Revista Brasileira de Sementes**, Londrina, v. 34, n. 4, p. 597-604, 2012.

SIMAK, M.; GUSTAFSSON, A. X-ray photography and sensitivity in forest three species. **Hereditas**, Lund, v. 39, p. 458-468, 1953.

TANG, S.; WANG, Y.; CHEN, Y. W. Application of ICA to X-ray coronary digital subtraction angiography, **Neurocomputing**, New York, v. 79, n. 1, p. 168-172, Mar. 2012.

UNIVERSIDADE FEDERAL DE LAVRAS. Biblioteca da UFLA. **Manual de normalização e estrutura de trabalhos acadêmicos: TCC, monografias, dissertações e teses**. Lavras, 2010. Disponível em: <<http://www.biblioteca.ufla.br/site/index.php>>. Acesso em: 27 mar. 2011.

ZHU, B. et al. Walnut shell and meat differentiation using fluorescence hyperspectral imagery with ICA-kNN optimal wavelength selection. **Sensing and Instrumentation for Food Quality and Safety**, New York, v. 1, n. 3, p. 123-131, 2007.

## SEGUNDA PARTE - ARTIGOS

### **ARTIGO 1 Análise de componentes independentes na avaliação de imagens radiográficas de sementes: um estudo de simulação**

#### **RESUMO**

A análise de componentes independentes (ICA) é uma técnica estatística que decompõe um conjunto de dados multivariados numa base de componentes não gaussianos e estatisticamente independentes entre si, o máximo possível. Neste trabalho, a ICA foi aplicada a imagens de raios X de 300 sementes de girassol (*Helianthus annuus* L.). As imagens-base obtidas forneceram características usadas na simulação de novas imagens. As imagens radiográficas simuladas foram classificadas segundo diferentes níveis de qualidade física da semente, diferenciando sementes cheias de sementes com algum tipo de dano ou deformação. A classificação, realizada por análise discriminante, atingiu um acerto global de até 97,3%, mostrando ser ICA uma técnica apropriada para a análise de imagens radiográficas de sementes de girassol.

Palavras-chave: Análise de componentes independentes. Análise de imagens. Análise discriminante. Qualidade de sementes. Raios X. Simulação.

## ABSTRACT

The independent component analysis (ICA) is a statistical technique which decomposes a set of multivariate data on a base of non-gaussian components, statistically independent between each other, at the highest level possible. In this work, the ICA was applied to X-ray images of 300 sunflower (*Helianthus annuus* L.) seeds. The base images obtained provided characteristics used in the simulation of new images. The simulated radiographic images were classified according to different physical qualities of the seeds, differentiating full seeds from deformed or damaged seeds. The classification, performed by discriminant analysis, reached a global hit of up to 97.3%, showing that the ICA is an appropriate technique for the analysis of sunflower seed radiographic images.

Keywords: Discriminant analysis. Image analysis. Independent component analysis. Seed quality. Simulation. X-rays.

## 1 INTRODUÇÃO

No tratamento de dados multivariados é comum que se deseje encontrar uma representação que revele sua estrutura interna e fontes subjacentes que não podem ser medidas diretamente. Técnicas estatísticas clássicas, como a análise fatorial e a análise de componentes principais (PCA - *Principal Component Analysis*) são utilizadas com este objetivo e fazem uso de transformações lineares do conjunto original de dados.

A análise de componentes independentes (ICA - *Independent Component Analysis*) é uma técnica estatística e computacional, cujo objetivo é encontrar uma representação linear de dados multivariados, de modo que os componentes sejam não gaussianos e estatisticamente independentes ou com dependência estatística minimizada. Pode ser vista como uma extensão da PCA por obter componentes independentes, propriedade estatística mais forte do que a não correlação dos componentes extraídos com o emprego de PCA. Em relação à análise fatorial a novidade está na natureza não gaussiana dos componentes.

Segundo Comon (1994) e Jutten e Herault (1991) na década de 80 foram os primeiros a proporem um algoritmo adaptativo capaz de separar simultaneamente todas as fontes desconhecidas independentes a partir de observações neurofisiológicas, o que deu nome à técnica ICA. Foi desenvolvida como uma técnica de processamento de sinais, referida como técnica de separação cega de fontes (BSS - *Blind Source Separation*). A partir da década de 90 pesquisadores como Bell e Sejnowski (1995), Cardoso (1989), Cichocki e Unbehauen (1996), Comon (1994), Hyvärinen, Karhunen e Oja (2001) deram significativas contribuições ao desenvolvimento da técnica com formulações matemáticas, estatísticas mais consistentes e construção de novos algoritmos de estimação. Desta forma a ICA passou a ter um papel relevante na análise de dados multivariados e as aplicações em diferentes

áreas, não necessariamente ligadas ao problema de separação de fontes, fizeram com que gradativamente o nome ICA fosse se dissociando de BSS.

As aplicações são inúmeras no tratamento de sinais dos mais diversos tipos, sejam imagens, ondas sonoras, sinais elétricos, eletromagnéticos, sensores químicos, e sinais biomédicos provenientes de eletroencefalogramas, eletrocardiogramas, ressonância magnética, entre outros.

Um dos objetivos na aplicação da ICA é a extração de características de um conjunto de dados observado. Esta consiste na redução de informações redundantes, podendo-se aliar a uma consequente redução na dimensão dos dados.

O uso de imagens de raios X na análise de sementes teve início em 1953 com Simak e Gustafsson que o utilizaram na avaliação de sementes de espécies florestais. É recomendado pela Associação Internacional de Análise de Sementes (INTERNATIONAL SEED TESTING ASSOCIATION - ISTA, 2011) e é um método que permite uma avaliação rápida e não destrutiva, a partir da visualização das estruturas internas da semente, diferenciando sementes bem formadas de sementes vazias, com danos mecânicos ou com ataque de insetos. Diversos trabalhos (CARVALHO; ALVES; OLIVEIRA, 2010; DELL'AQUILA, 2009; KOBORI; CICERO; MEDINA, 2012) ratificam esta recomendação, concluindo que o teste de raios X é eficiente para avaliar a morfologia interna das sementes e permite prever seus reflexos no potencial fisiológico e selecionar sementes de alta qualidade.

Contudo, a análise das imagens radiográficas ainda está sujeita à avaliação subjetiva do analista. Esforços têm sido direcionados na automatização deste processo com o uso de softwares de processamento de imagens como forma de propiciar uma análise mais acurada e rápida (CARVALHO; ALVES; OLIVEIRA, 2010; DELL'AQUILA, 2009), minimizando a subjetividade pessoal do analista.

A aplicação de ICA em imagens de raios X aparece em trabalhos de di-

versas áreas, principalmente na área médica. Campos, Barros e Silva (2007) classificaram tecidos da mama em normais, com tumores benignos ou com tumores malignos, a partir do uso de ICA na extração de características de mamogramas. Chen et al. (2007) propuseram um método baseado em ICA para a remoção da dispersão dos raios-X em imagens de raios-X. Tiilikainen et al. (2007) propuseram um algoritmo genético usando ICA no ajuste de curvas de refletividade de raios X. Tang, Wang e Chen (2012) usaram ICA em angiografias digitais das coronárias, separando vasos coronarianos do fundo da imagem.

A proposta deste trabalho é validar a aplicação da ICA na análise de imagem de raios X de sementes utilizando simulação. Características obtidas da decomposição realizada por ICA foram usadas na simulação de novas imagens. As imagens simuladas foram submetidas à classificação pela técnica de análise discriminante.

## **2 MATERIAL E MÉTODOS**

### **2.1 Obtenção e processamento dos dados**

Foram selecionadas 300 imagens radiográficas de uma amostra de sementes de girassol (*Helianthus annuus* L.), cultivar Hélio-250, produzidas em Uberlândia, estado de Minas Gerais (MG), Brasil, safra 2010/2011. As sementes foram radiografadas no Laboratório de Análise de Sementes do Departamento de Agricultura da Universidade Federal de Lavras - MG, em aparelhos de raios-X Faxitron Modelo MX20 gerando imagens digitalizadas salvas no formato jpeg, como na Figura 1.

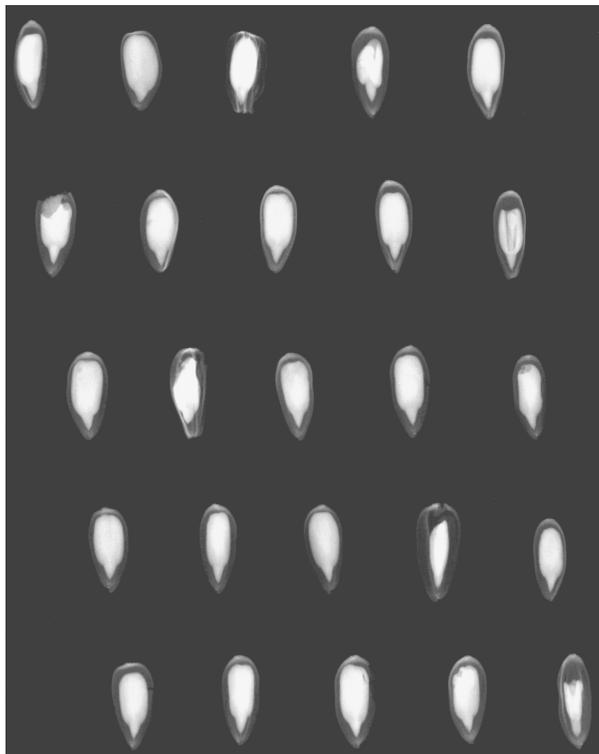


Figura 1 Imagens radiográficas de sementes de girassol cultivar Hélio-250

As imagens radiográficas foram analisadas visualmente de acordo com a morfologia interna e as sementes foram classificadas em três grupos: sementes cheias, com embrião opaco na análise radiográfica; sementes com dano leve, mas com preservação das características do eixo embrionário e sementes deformadas ou com dano grave, com danos afetando o eixo embrionário (Figura 2).

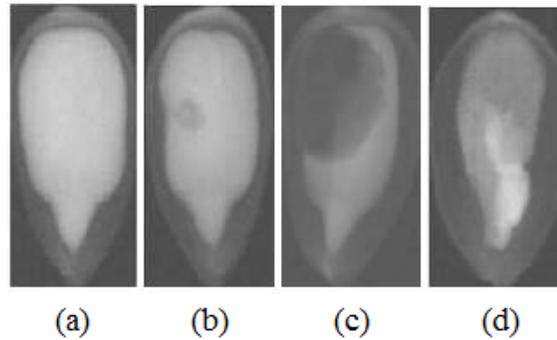


Figura 2 Imagens radiográficas de sementes de diferentes níveis de qualidade: (a) cheia; (b) com dano leve, mas com preservação das características do eixo embrionário; (c) com dano grave e (d) deformada, com danos afetando o eixo embrionário

Das imagens das lâminas com 25 sementes foram selecionadas manualmente 300 imagens retangulares de sementes individuais, separados em conjuntos de 100 imagens de cada grupo de classificação como descrito anteriormente. Por apresentarem diferentes dimensões, as imagens foram padronizadas para o tamanho médio de todas as imagens de  $121 \times 268$  pixels e redimensionadas num vetor linha de tamanho  $1 \times 32.428$ .

## 2.2 Análise de componentes independentes

Dado o vetor aleatório  $X = [X_1, X_2, \dots, X_n]^T$ , cujos  $n$  elementos são misturas de  $k$  componentes estatisticamente independentes entre si de um vetor aleatório  $S = [S_1, S_2, \dots, S_k]^T$ , o modelo ICA é expresso da forma

$$X = A \cdot S \quad (2.1)$$

em que  $A$  representa uma matriz de coeficientes  $a_{ij}$ , denominada matriz de mistura.

Neste modelo, apenas as observações  $X_i$  são conhecidas e a ICA propõe-se a determinar uma transformação linear dada pela matriz  $\mathbf{W}$  (inversa ou pseudo-inversa de  $\mathbf{A}$ ), denominada matriz de separação, de modo que o vetor  $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^T$  seja uma estimativa de  $\mathbf{S}$ .

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X} \quad (2.2)$$

Algumas restrições devem ser cumpridas para assegurar a estimação do modelo proposto (HYVÄRINEN, 1999).

1. Os componentes  $S_j$ , também ditos fontes, devem ser estatisticamente independentes entre si ou ter dependência estatística minimizada.
2. Todas as fontes devem possuir distribuição não gaussiana, com exceção de no máximo uma fonte.
3. O número de misturas observadas  $n$  deve ser no mínimo igual ao número de fontes  $k$ , ou seja,  $n \geq k$ .
4. A matriz de mistura  $\mathbf{A}$  deve possuir posto coluna completo.

Os diversos métodos de estimação de  $\mathbf{W}$  foram desenvolvidos explorando as restrições de independência estatística e de não gaussianidade das fontes. Os algoritmos diferem entre si na definição de uma função objetivo que será maximizada ou minimizada. Dentre os principais destacamos os baseados na maximização da não gaussianidade como o FastICA (HYVÄRINEN; OJA, 1997), o algoritmo JADE - Joint Approximate Diagonalization of Eigenmatrices (CARDOSO; SOULOUMIAC, 1993) que usa tensores cumulantes de quarta ordem e o Infomax (BELL; SEJNOWSKI, 1995) fortemente relacionado com a maximização da verossimilhança.

### 2.3 Aplicação de ICA ao conjunto de dados

A aplicação da ICA em imagens radiográficas de sementes parte do pressuposto de que cada imagem em particular seja resultado da mistura de um conjunto de imagens-base que são informações independentes comuns a todas as radiografias de sementes de diferentes níveis de qualidade. Cada imagem-base ou componente independente (IC) contribui com algum tipo de informação relativa à forma, grau de preenchimento e diferente tipo de dano numa região específica da semente.

Segundo o modelo ICA, definido na equação (2.1), cada imagem de semente é uma realização da variável aleatória  $X_i$ , denotada por  $x_i, i = 1, 2, \dots, n$ . Decompondo cada uma destas imagens numa combinação linear dos componentes independentes  $s_j, j = 1, 2, \dots, k$ , que a geram, temos

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \dots + a_{ik}s_k, \text{ para todo } i = 1, 2, \dots, n. \quad (2.3)$$

Com a aplicação da ICA, a imagem analisada  $x_i$  passa a ser representada por um vetor cujos elementos são os coeficientes  $a_{ij}$  de cada componente independente na mistura, como na Figura 3.

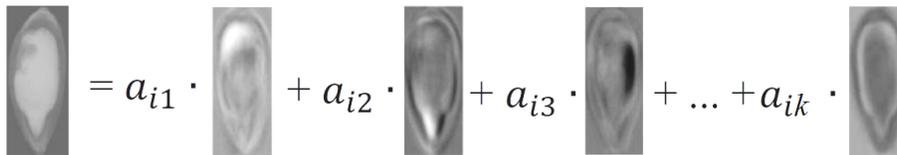


Figura 3 Imagem de semente decomposta em uma combinação linear de  $k$  imagens-base

Neste trabalho foi utilizado o algoritmo FastICA com pacote implemen-

tado no MATLAB<sup>®</sup> (FASTICA..., 2005), definido para maximizar a não gaussianidade dos componentes a partir da medida da curtose. O FastICA usa como dados de entrada a matriz  $\mathbf{X}_{(n \times p)}$  das 300 imagens radiográficas de sementes, cada uma disposta numa linha com  $p$  pixels, e estima a matriz de separação  $\widehat{\mathbf{W}}$ , a matriz de mistura  $\widehat{\mathbf{A}}$  e as estimativas dos componentes independentes  $\widehat{\mathbf{Y}}$ , conforme equações (2.1) e (2.2). Para facilitar a estimação o algoritmo realiza um pré-processamento nos dados observados, centralizando-os e tornando-os não correlacionados. Nesta fase é feita uma redução de dimensão, escolhendo-se o número de componentes independentes  $k < n$  que preserve um bom percentual de variação dos dados. Segundo Hyvärinen, Karhunen e Oja (2001) a decisão pelo número de componentes independentes a serem estimados é um problema frequente e esta escolha é feita geralmente por tentativa e erro. Desta forma o FastICA foi aplicado considerando  $k = 30$ ,  $k = 45$  e  $k = 60$ .

## 2.4 Análise discriminante

A análise discriminante (DA - *Discriminant Analysis*) é uma técnica estatística empregada para diferenciar populações ou classificar novos indivíduos em uma das várias populações, considerando a estrutura multidimensional dos dados observados.

Considerando-se que existam  $m$  grupos ou populações  $\Pi_1, \Pi_2, \dots, \Pi_m$ , conhecidos *a priori*, e amostras aleatórias para estes grupos de tamanhos  $n_1, n_2, \dots, n_m$ , o objetivo é construir regras de discriminação para determinar a qual dos grupos irá pertencer um novo indivíduo.

A função discriminante quadrática é dada por

$$Q_i(\mathbf{x}) = -\frac{1}{2} \ln(|\mathbf{C}_i|) - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{X}}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i) + \ln(p_i), \quad (2.4)$$

para  $i = 1, 2, \dots, m$ , em que  $p_i$  é a probabilidade *a priori* da observação multivariada  $\mathbf{x}$  pertencer ao grupo  $i$  e  $\bar{\mathbf{X}}_i$  e  $\mathbf{C}_i$  são os estimadores amostrais da média e covariância do  $i$ -ésimo grupo (FERREIRA, 2008).

As linhas da matriz  $\hat{\mathbf{A}}$ , estimada pelo FastICA, são os vetores das características extraídas das imagens de sementes usados na classificação por DA. Usando-os como amostras de treinamento dos três grupos de sementes, previamente classificadas quanto à morfologia interna, determina-se a função discriminante que caracteriza cada grupo. A partir de  $Q_i$  uma nova observação  $\mathbf{x}$  é alocada no grupo  $\Pi_i$  se  $Q_i(\mathbf{x}) \geq Q_l(\mathbf{x})$  para todo  $l \neq i, l = 1, 2, \dots, m$ .

## 2.5 Simulação de amostras de testes

A simulação de novas imagens visou testar a classificação das sementes realizada pela DA, verificando se esta funciona de forma eficaz a partir das características extraídas por ICA.

Foram selecionadas 15 imagens de sementes do conjunto de dados, distribuídas em cada uma das três categorias de classificação (5 sementes cheias, 5 com danos leves e 5 com danos graves). Cada vetor de imagem  $\mathbf{x}_i$  de dimensão  $p$  pixels passou a ser representado pelo seu respectivo vetor  $\mathbf{a}_i$ , constituído das coordenadas  $a_{ij}$  da base de  $k$  componentes independentes estimados. O vetor  $\mathbf{a}_i$  é uma linha da matriz de mistura  $\hat{\mathbf{A}}$ , estimada pelo FastICA, e é o vetor utilizado na simulação. Cada vetor  $\mathbf{a}_i$ , representativo de uma imagem selecionada  $\mathbf{x}_i$ , foi repetido 10 vezes e a cada uma dessas repetições foi adicionado um ruído aleatório com distribuição gaussiana de média zero e variância quase nula, gerando 150 novas imagens de sementes. A simulação foi realizada com dois valores de variância do ruído aleatório:  $10^{-6}$  e  $3,6 \cdot 10^{-7}$ . A mínima modificação nas características da imagem permite supor que a nova semente gerada continue fazendo parte do mesmo grupo

de classificação da imagem da semente original.

Os vetores correspondentes às 150 imagens de sementes simuladas constituíram uma amostra de teste que foi submetida à classificação pela função discriminante quadrática (2.4).

Foram utilizadas duas metodologias na escolha das imagens a serem usadas na simulação.

No primeiro caso as imagens foram escolhidas pelo analista, procurando-se aquelas que visualmente parecessem ser bem representativas de cada categoria de classificação. A simulação e classificação foram repetidas 10 vezes para cada um dos valores de variância e foram estimados os acertos e erros de classificação. Todo este procedimento foi repetido para outras três diferentes amostras de 15 imagens selecionadas visualmente.

A segunda metodologia consistiu na escolha das 15 imagens por sorteio, ainda considerando que fossem 5 de cada grupo de classificação. No total foram sorteadas 100 diferentes amostras de 15 sementes e para cada uma delas foi feita a simulação e classificação.

### **3 RESULTADOS E DISCUSSÃO**

Na Figura 4 tem-se as imagens-base geradas pela estimação de 60 componentes independentes nas quais são percebidas alterações em distintas regiões da semente.

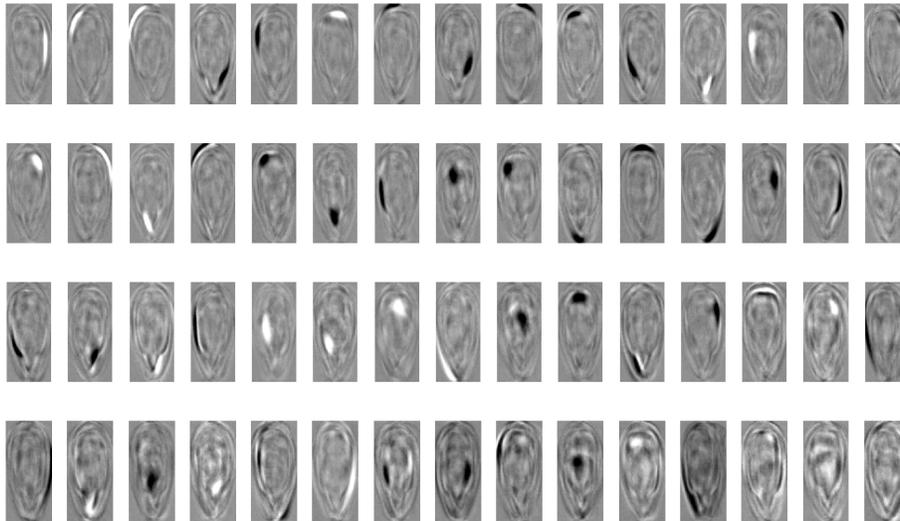


Figura 4 Imagens de 60 componentes independentes estimados por ICA

Na Figura 5 são mostradas imagens radiográficas originais de sementes de diferentes níveis de qualidade física (cheia, com dano leve e deformada, de cima para baixo) na primeira coluna; nas colunas seguintes tem-se imagens das sementes simuladas a partir das características estimadas por ICA, usando bases de diferentes dimensões ( $k$ ) e variância no erro aleatório de  $10^{-6}$  (imagens das colunas A) e de  $3,6 \cdot 10^{-7}$  (imagens das colunas B). Apesar de as imagens serem bem parecidas visualmente, uma variação em torno de 5% nas coordenadas dos vetores representativos das imagens produz diferentes resultados de classificação.

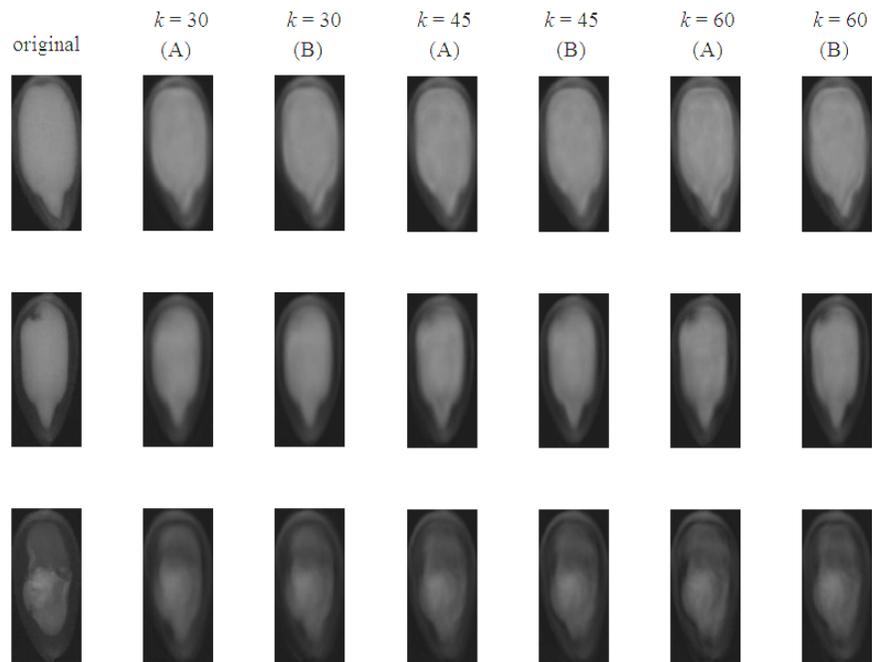


Figura 5 Imagens radiográficas originais de sementes de girassol com diferentes níveis de qualidade física (primeira coluna) e imagens simuladas, a partir de bases de componentes independentes de diferentes dimensões ( $k$ ) com variância no erro aleatório de  $10^{-6}$  (colunas A) e  $3,6 \cdot 10^{-7}$  (colunas B).

Nas Figuras 6 e 7 são apresentados os resultados das classificações das 150 imagens de sementes simuladas com variância no erro aleatório de  $10^{-6}$  e  $3,6 \cdot 10^{-7}$ , respectivamente, cujas amostras foram escolhidas visualmente. Estes resultados são a média dos resultados da classificação de todas as repetições na simulação das diferentes amostras. Observa-se que o grupo de sementes deformadas ou com danos graves foi o melhor classificado, chegando a atingir 100% de acerto na classificação. Melhores resultados de classificação são verificados quando a simulação aplicou menor variância no erro aleatório, garantindo que o

ruído adicionado ao vetor representativo da semente original não modificasse a imagem da semente ao ponto de transformá-la em representativa de outro grupo. Globalmente os melhores resultados foram acertos de 90,4% com uso de variância  $10^{-6}$  em dados de dimensão 30 e 97,3% com uso de variância  $3,6 \cdot 10^{-7}$  em dados de dimensão 45.

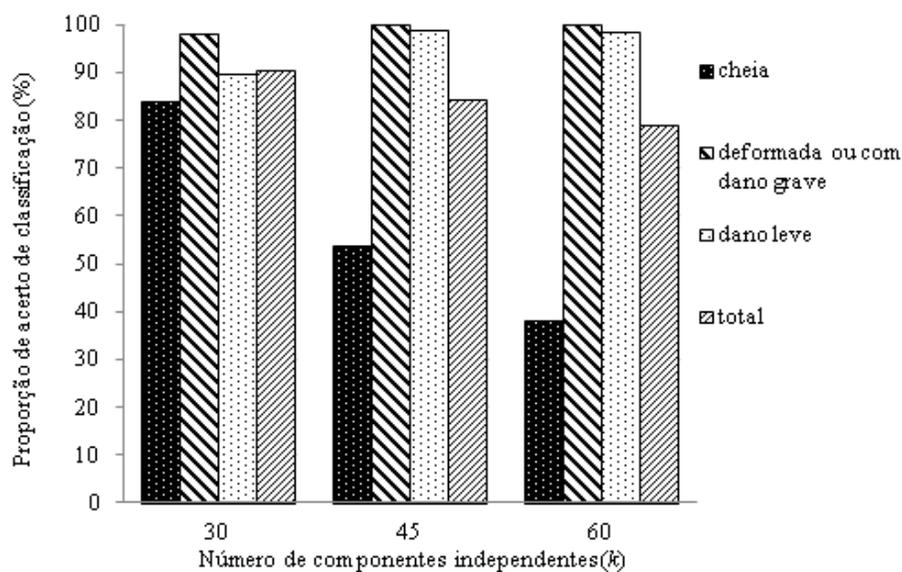


Figura 6 Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância  $10^{-6}$  a partir de amostras escolhidas pelo analista

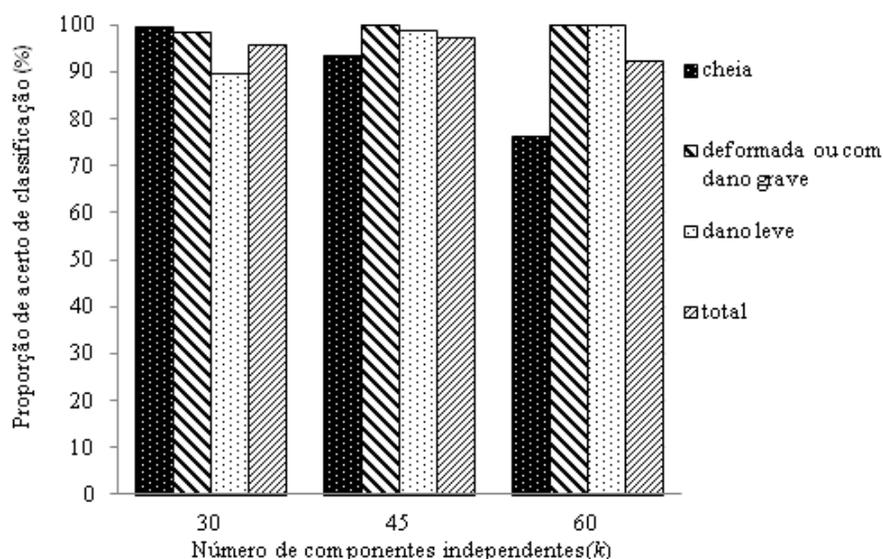


Figura 7 Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância  $3,6 \cdot 10^{-7}$  a partir de amostras escolhidas pelo analista

Nas Figuras 8 e 9 são apresentadas as médias dos resultados das 100 classificações das amostras escolhidas aleatoriamente e cujas sementes foram simuladas com variância  $10^{-6}$  e  $3,6 \cdot 10^{-7}$ , respectivamente. Estes resultados foram um pouco inferiores aos obtidos com amostras escolhidas visualmente pelo analista. Deve-se considerar que neste caso, além da aleatoriedade, não houve repetição nas simulações de uma mesma amostra como na metodologia anterior, o que influencia no resultado. Porém as características dos resultados anteriores se mantêm. O uso de uma variância menor na simulação resultou em melhor classificação. Para variância de  $10^{-6}$  os melhores resultados foram obtidos com o uso de 30 ICs com acerto global de 86% e para variância  $3,6 \cdot 10^{-7}$  a maior proporção de acerto global foi 94% com o uso de 45 ICs. As sementes deformadas ou com dano grave foram mais uma vez as que obtiveram melhores proporções de acerto de classificação.

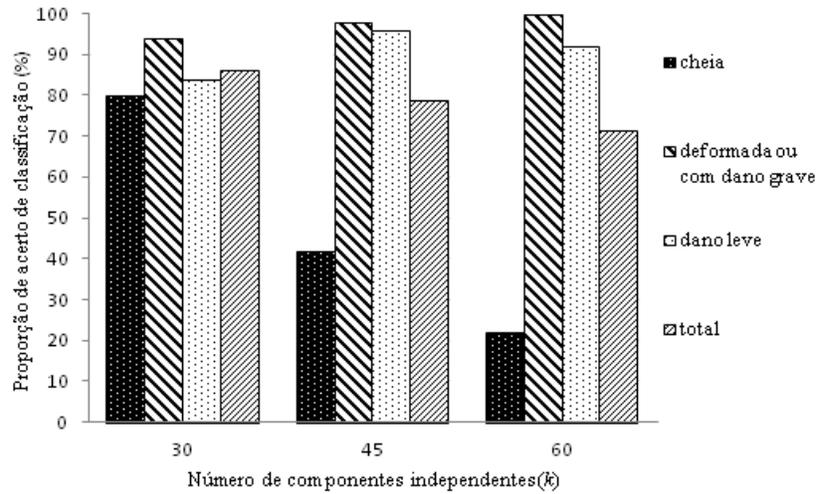


Figura 8 Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância  $10^{-6}$  a partir de amostras escolhidas aleatoriamente

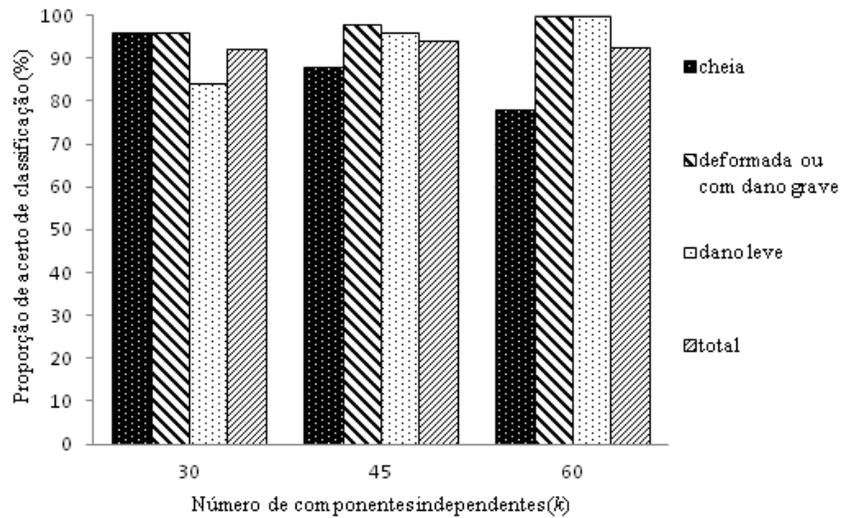


Figura 9 Resultados da classificação de 150 sementes simuladas com erro aleatório normal de variância  $3,6 \cdot 10^{-7}$  a partir de amostras escolhidas aleatoriamente

Em todos os testes observa-se que o aumento na dimensão do vetor representativo da imagem ( $k$ ) provoca uma inversão nos tipos de erro de classificação. Quanto maior o número de ICs utilizados, e conseqüente aumento no detalhamento da imagem, maior o erro em classificar sementes cheias como sementes portadoras de danos leves. Portanto, o critério para avaliar qual número de ICs resultou em melhores resultados não considerou apenas a proporção total de acerto, mas a homogeneidade dos percentuais de acertos entre os grupos.

#### **4 CONCLUSÃO**

A partir dos resultados obtidos pode-se considerar que a análise de componentes independentes é uma técnica apropriada para a extração de características de imagens radiográficas de sementes. Com a simulação de novas imagens a partir da aplicação de ICA foi possível testar a robustez da análise discriminante, assegurando que ambas as técnicas poderão ser utilizadas satisfatoriamente na classificação de novas imagens reais de sementes.

## REFERÊNCIAS

BELL, A. J.; SEJNOWSKI, T. J. An information-maximization approach to blind separation and blind deconvolution. **Neural Computation**, Cambridge, v. 7, n. 6, p. 1129-1159, 1995.

CAMPOS, L. F. A.; BARROS, A. K.; SILVA, A. C. Independent component analysis and neural networks applied for classification of malignant, benign and normal tissues in digital mammography. **Methods of Information in Medicine**, Stuttgart, v. 46, n. 2, p. 212-215, 2007.

CARDOSO, J. F. Source separation using higher order moments. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 89., 1989, Glasgow. **Proceedings...** Glasgow: IEEE, 1089. p. 2109-2112. Disponível em: <<http://perso.telecom-paristech.fr/cardoso/Papers.PDF/icassp89.pdf>>. Acesso em: 15 maio 2012.

CARDOSO, J. F.; SOULOUMIAC, A. Blind beamforming for non Gaussian signals. **IEE Proceedings - Part F**, London, v. 140, n. 6, p. 362-370, Dec. 1993.

CARVALHO, M. L. M.; ALVES, R. A.; OLIVEIRA, L. M. Radiographic analysis in castor bean seeds (*Ricinus communis* L.). **Revista Brasileira de Sementes**, Londrina, v. 32, n. 1, p. 170-175, jan. 2010.

CHEN, Y. W. et al. Independent component analysis for removing x-ray scatter in x-ray images. In: INSTRUMENTATION AND MEASUREMENT TECHNOLOGY CONFERENCE, 1., 2007, Warsaw. **Proceedings...** Warsaw: IMTC, 2007. p. 1-4.

CICHOCKI, A.; UNBEHAUEN, R. Robust neural networks with on-line learning for blind identification and blind separation of sources. **IEEE Transactions on Circuits and Systems**, New York, v. 43, n. 11, p. 894-906, 1996.

COMON, P. Independent component analysis, a new concept? **Signal Processing**, Amsterdam, v. 36, n. 3, p. 287-314, 1994.

DELL' AQUILA, A. Development of novel techniques in conditioning, testing and sorting seed physiological quality. **Seed Science and Technology**, Zurich, v. 37, n. 3, p. 608-624, 2009.

FASTICA for MATLAB 7.x and 6.x. Version 2.5. Chicago, 2005. Disponível em: <<http://research.ics.aalto.fi/ica/fastica/>>. Acesso em: 6 maio 2011.

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008, 662 p.

HYVÄRINEN, A. Survey on independent component analysis. **Neural Computing Surveys**, George, v. 2, n. 4, p. 94-128, 1999.

HYVÄRINEN, A.; OJA, E. A fast fixed-point algorithm for independent component analysis. **Neural Computation**, Cambridge, v. 9, n. 7, p. 1483-1492, 1997

HYVÄRINEN, A.; KARHUNEN J.; OJA E. **Independent component analysis**. New York: J. Wiley, 2001. 481 p.

INTERNATIONAL SEED TESTING ASSOCIATION. **International rules for seed testing Association**. Bassersdorf, 2011. 174 p.

JUTTEN, C.; HERAULT, J. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. **Signal Processing**, Amsterdam, v. 24, n. 1, p. 1-10, 1991.

KOBORI, N. N.; CICERO, S. M.; MEDINA, P. F. Teste de raios X na avaliação da qualidade de sementes de mamona. **Revista Brasileira de Sementes**, Londrina, v. 34, n. 1, p. 125-133, 2012.

SIMAK, M.; GUSTAFSSON, A. X-ray photography and sensitivity in forest three species. **Hereditas**, Lund, v. 39, p. 458-468, 1953.

TANG, S.; WANG, Y.; CHEN, Y. W. Application of ICA to X-ray coronary digital subtraction angiography. **Neurocomputing**, New York, v. 79, n. 1, p.

168-172, Mar. 2012.

TIILIKAINEN, J. et al. Genetic algorithm using independent component analysis in X-ray reflectivity curve fitting of periodic layer structures. **Journal of Physics D: Applied Physics**, London, v. 40, n. 19, p. 6000-6004, 2007.

**ARTIGO 2 Evaluation of seed radiographic images by independent component analysis and discriminant analysis**

## Evaluation of seed radiographic images by independent component analysis and discriminant analysis

I.C.C. LEITE<sup>1</sup>, T. SÁFADI<sup>2</sup> AND M.L.M. CARVALHO<sup>3</sup>

<sup>1</sup> Instituto Federal da Bahia, Departamento de Ciências Aplicadas - Rua Emídio dos Santos, s/n. Salvador, Bahia, Brasil, 40.301-015 (E-mail: isaleite@ifba.edu.br)

<sup>2</sup> Universidade Federal de Lavras, Departamento de Ciências Exatas, CP. 3037, Lavras, MG, Brasil, 37200-000 (E-mail: safadi@dex.ufla.br)

<sup>3</sup> Universidade Federal de Lavras, Departamento de Agricultura, Setor de Sementes, CP. 3037, Lavras, MG, Brasil, 37200-000 (E-mail: mlaenemc@dag.ufla.br)

### Summary

Although subjective, the use of X-ray images of seeds is an important tool for analysing seed lot quality. Here, we applied independent component analysis (ICA) for automatic processing of radiographic images of 600 sunflower seeds. The X-rayed seeds were also subjected to a germination test. The ICA technique was implemented with the FastICA algorithm, which decomposed X-ray images to independent basis images. Based on features extracted by ICA, we used discriminant analysis (DA) to classify seed quality. The classification achieved an overall accuracy of 82%. The results showed that ICA and DA were effective in X-ray analysis to associate seed morphology and seedling performance.

### Introduction

Analysis of seeds is essential for determining seed lot quality and hence sowing value. Laboratory tests are performed to determine physical and physiological potential of seeds with respect to germination capacity and seedling vigour; however, such tests are destructive and time-consuming. Results can also depend on the analyst's subjectivity. Hence, assessing seed quality with radiographic images has been used as alternative to standard laboratory testing.

X-ray images in seed testing began with Simak and Gustafsson (1953), who used the method for evaluating forest tree seeds. Recommended by the

International Seed Testing Association (ISTA, 2011), the method allows for quick and non-destructive assessment based on visualisation of seed internal structures, which distinguishes well-formed seeds from empty, mechanically damaged or insect-attacked seeds. However, evaluation of radiographic images still depends on the analyst's subjectivity. This may be reduced with techniques for automatic processing of images in which analysis by software helps analysts evaluate seeds. According to Carvalho *et al.* (2010) and Dell' Aquila (2006), one of the major requirements in developing machine vision systems for analysing and sorting seeds is the ability to analyse an image accurately and quickly.

Independent component analysis (ICA), a method that emerged in the mid-1980s with Herault and Jutten (1986), was subsequently developed as a signal processing technique. Independent component analysis of X-ray images has been used in several areas: Tang *et al.* (2012) used ICA to separate the coronary vessels and backgrounds in X-ray digital subtraction angiography; Tiilikainen *et al.* (2007) proposed a genetic algorithm using ICA in X-ray reflectivity curve fitting of periodic layer structures; Campos *et al.* (2007) proposed a method using features extracted by ICA to classify mammograms showing benign, malignant or normal tissues; and Saidi *et al.* (2004) used ICA of microarray data in the study of endometrial cancer and showed that ICA-generated patterns more clearly characterised malignant samples, compared with principal component analysis (PCA).

In our research ICA was used to reduce the dimension of observed data by getting a linear projection of the X-ray images of seeds on the basis of statistically independent reduced scale images, thereby providing a parsimonious representation with maximum information on original data (Fiori, 2003). Based on the image representation obtained by ICA, we then used discriminant analysis (DA) as a technique for classifying sunflower seeds under different levels of physical

quality.

### **Materials and methods**

Data collection was carried out at the Seed Analysis Laboratory of the Department of Agriculture, Federal University of Lavras, Minas Gerais, Brazil, from a seed sample of sunflower (*Helianthus annuus* L.) cultivar Hélio-250 produced in Uberlândia, state of Minas Gerais, Brazil, 2010/2011 harvest. A set of 600 seeds was randomly selected to be X-rayed without any special preparation. Using double-sided adhesive tape, subsamples of 25 seeds were arranged on transparent blades and numbered according to their location on the slide (row and column) to enable identification in subsequent tests and classifications. Seeds were radiographed (22 kV, average of 11 seconds exposure) in a Faxitron Model MX-20 to generate scanned images, which were saved in jpeg format.

Then, radiographed seeds were tested for germination, following the standards established by the Rules for Seed Analysis (Brasil, 2009), at 25°C with an amount of water that was twice the weight of the paper substrate. At four (first count) and 10 (final count) days after sowing, seeds were classified and counted. Seeds that had germinated by the time of the first count were classified as fast-germinating seeds, while those germinating by 10 days were classified as slow-germinating seeds. Seeds giving abnormal seedlings, dead and dormant seeds were classified as ungerminated seeds.

The radiographic images were analysed visually according to internal morphology and seeds classified into three groups: full seeds, with opaque embryo in X-ray analysis; slightly injured seeds, although preserving characteristics of embryonic axis; and deformed seeds, which showed damage affecting the embryonic axis (figure 1).

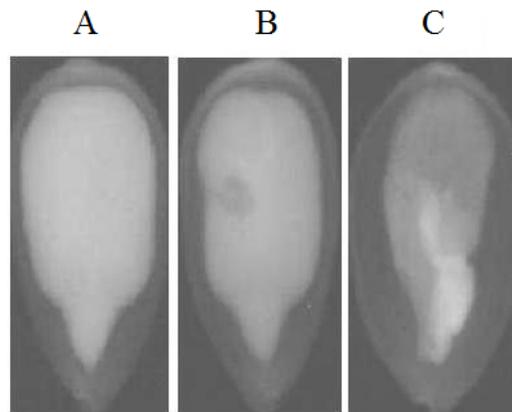


Figure 1 Radiographic images of sunflower seeds with different quality levels: A, full; B, slightly injured, although preserving characteristics of the embryonic axis; and C, deformed, with severe injury affecting the embryonic axis.

A set of 445 images of individual seeds was selected manually, comprising 175 seeds classified as full, 140 classified as deformed or severely injured, and 130 classified as slightly injured. As images had different dimensions, they were standardised to the average size ( $121 \times 268$  pixels).

Independent component analysis (ICA) and discriminant analysis techniques were used to analyse images. ICA is a statistical technique that reveals the internal structure of a set of multivariate data by decomposing it on a base of components as statistically independent as possible and non-Gaussian (Hyvärinen *et al.*, 2001). Thus, application of ICA for radiographic images of seeds assumes that each image results from the mixing of a set of basis images which are independent information that is common to all radiographs. Each basis image or independent component (IC) contributes information such as shape, degree of filling or type of damage in specific seed regions.

A matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$  whose rows are  $p$ -dimensional vectors comprised of pixels of each seed image, is generated by mixing  $n$  mutually independent components, such that the ICA model is given by

$$\mathbf{X}_{(n \times p)} = \mathbf{A}_{(n \times n)} \cdot \mathbf{S}_{(n \times p)},$$

where  $\mathbf{A}$  is the matrix of coefficients  $a_{ij}$  of the linear combination, the ‘mixing matrix’, while  $\mathbf{S}$  is the matrix of independent components  $\mathbf{s}_i$ .

For size reduction, a number  $k < n$  of ICs can be selected using principal component analysis (PCA) as pre-processing for ICA so that

$$\mathbf{X}_{(n \times p)} \approx \mathbf{A}_{(n \times k)} \cdot \mathbf{S}_{(k \times p)}.$$

Each  $\mathbf{x}_i$  image is decomposed into a linear combination of ICs (basis images) given by

$$\mathbf{x}_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 + \dots + a_{ik}\mathbf{s}_k, \text{ for every } i = 1, 2, \dots, n,$$

so that each image is represented by the coefficients of each independent component of the mixture, as shown in figure 2a.

The mixing matrix,  $\mathbf{A}$ , and the matrix of independent components,  $\mathbf{S}$ , are estimated by algorithms based on the independence of variables, using higher order statistics which maximises the non-Gaussianity of ICs. In this study we used the FastICA algorithm proposed by Hyvärinen and Oja (1997), which maximises the non-Gaussianity of data measured by kurtosis or negentropy.

The FastICA algorithm was applied to the set of images by varying the choice of number of ICs ( $k$ ). An X-ray image previously represented by a  $p$ -dimensional vector composed of pixels, is now represented by a  $k$ -dimensional

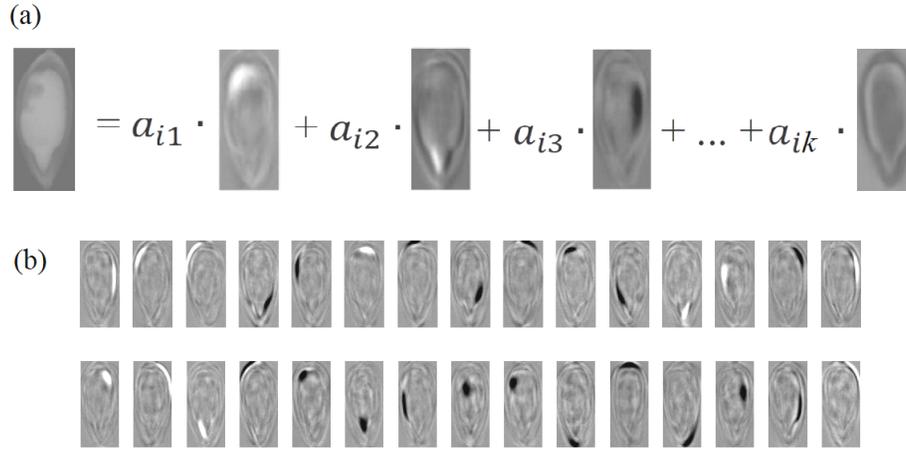


Figure 2 Radiographic images of sunflower seeds. (a) Image of seed decomposed into a linear combination of  $k$  basis-images. (b) Image of 30 independent components estimated by ICA.

vector of coefficients of the linear combination of independent components. ICA was performed estimating 20, 30, 45 and 60 ICs.

Discriminant analysis (DA) is a statistical technique used to differentiate populations or classify new individuals into one of the various populations by considering the multidimensional structure of observed data. DA was performed as a technique for seed image classification; the input vectors were the lines of matrix  $\mathbf{A}$  as estimated by FastICA, which correspond to weight vectors associated with each image.

A subset of the seed images was used as a training sample to create a discriminant rule that classifies a new image into one of the three groups defined previously. Based on the corresponding estimators obtained from samples of three groups, we determined the quadratic discriminant function for the  $i$ -th group

represented by  $Q_i(\mathbf{x})$  and given by

$$Q_i(\mathbf{x}) = -\frac{1}{2} \ln(|\mathbf{C}_i|) - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{X}}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i) + \ln(p_i),$$

in which  $p_i$  is a priori probability of sample belonging to group  $i$ , while  $\bar{\mathbf{X}}_i$  and  $\mathbf{C}_i$  are sample estimators of mean and covariance of the  $i$ -th group.

Based on function  $Q_i$ , classifying a new observation  $\mathbf{x}$  in group  $i$  is considered if  $Q_i(\mathbf{x}) \geq Q_l(\mathbf{x})$  for every  $l \neq i, l = 1, 2, 3$  (Johnson and Wichern, 2007).

Vectors corresponding to the remaining images in the data set were submitted to classification based on the given quadratic discriminant function. DA was performed with two data sets at different times.

We started with 400 images separating the set into training sample (300 images) and testing sample (100 images). The training sample was composed by 100 images of each group previously defined and was used to determine discriminant functions. In the testing sample there were 50 seeds classified as full, 20 classified as deformed, and 30 classified as slightly injured that were used for classification by the discriminant functions created.

A second analysis was performed by partitioning the same data set (400 images) differently, thereby varying the composition of training and testing samples and keeping sample totals, as described above.

In a third analysis the full set of images was reclassified by the analyst and the severely injured seeds, which had been excluded from the previous analysis, were included in the group of deformed seeds since they had similar patterns, thereby increasing the size of the data set. The 445 reclassified images were separated again into training (300 images) and testing (145 images) samples. The training sample had the same composition as the previous tests and the testing sample was composed of 75 seeds classified as full, 40 classified as deformed and 30 classified

as slightly injured.

Later we estimated accuracy and misclassification. Detailed errors of seed classification were calculated. A false-positive error occurs when an element is not classified in the right original group, whereas classifying elements in the wrong group is a false-negative error.

## **Results**

Basis images generated by estimation of 30 independent components with changes in different seed parts are shown in figure 2b.

The results of the first analysis were not satisfactory (data not shown). The overall proportion of seeds that were accurately identified was 45 and 46% using data extracted from the images with  $k = 30$  and  $k = 60$  dimensions, respectively. Although the differences between the results were minimal globally, using 60 ICs gave more correct classification of slightly injured seeds (67%) and deformed seeds (60%) and less correct classification of full seeds (28%) than using 30 ICs. The proportions of correctly classified seeds were 57% for slightly injured seeds, 50% for deformed seeds and 36% for full seeds using 30 ICs.

Better results were observed on the new evaluation (figure 3), when the same data set was partitioned differently in training and testing samples, than in the previous classification. Using 30 ICs, the proportions of correctly classified seeds in each group were similar, while the overall proportion of correct classification was higher than in the classification using 60 ICs. Nevertheless, deformed and slightly injured seeds were identified with better accuracy using 60 ICs.

Due to the large difference between test results shown by sample variation alone, visual classification of seed images could be ambiguous and the third analysis was performed with the reclassified data set.

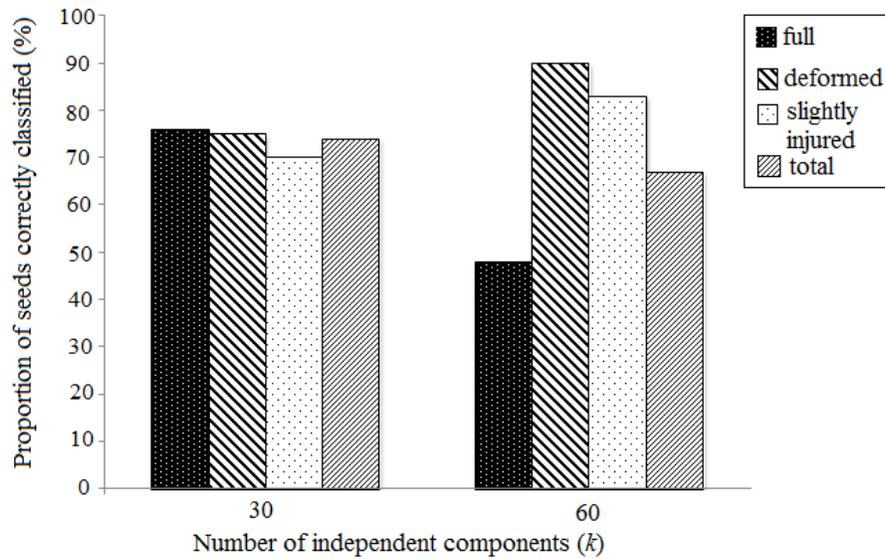


Figure 3 Proportion of sunflower seeds correctly classified depending on the number of estimated independent components (ICs) from a new sample composition

Proportions of accurate classification obtained in the third analysis were higher than observed previously (figure 4). The group of deformed or severely injured seeds maintained proportions of correct classification equal or superior to 90%, regardless of the size of input data.

As the number of ICs increased, more images previously classified as full were classified as slightly injured. Generally, the best classification results were found with 30 ICs (table 1).

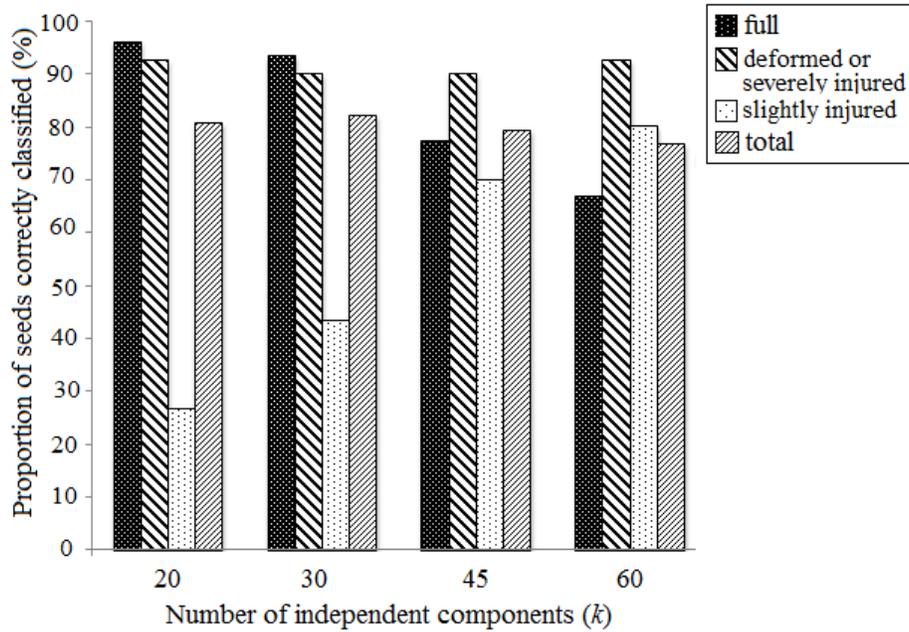


Figure 4 Proportion of sunflower seeds correctly classified regarding the number of estimated independent components (ICs) based on the new visual classification of images

Table 1 Results of seed classification by quadratic discriminant function for data size  $k = 30$ .

Original population	Classified population			Total in original population
	Full	Deformed or severely injured	Slightly injured	
Full	70	0	5	75
Deformed or severely injured	0	36	4	40
Slightly injured	15	2	13	30
Total in classified population	85	38	22	145

For  $k = 20$  and  $k = 30$  the highest proportion of seeds misclassified was for false-positive errors of slightly injured seeds, which were mostly considered full seeds (figure 5). As the number of ICs increased, an inversion in percentage of false-negative and false-positive errors occurred in the groups of full seeds and slightly injured seeds.

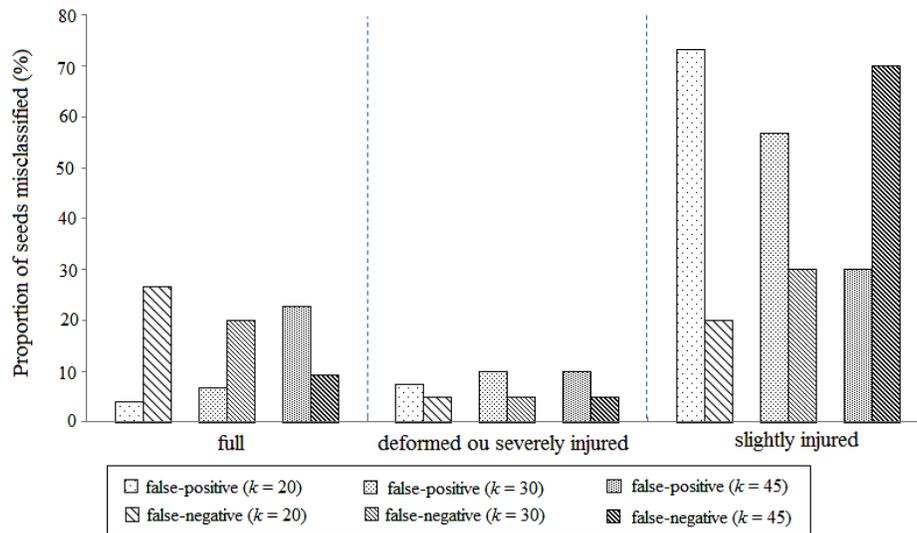


Figure 5 Proportion of sunflower seeds misclassified in each group using quadratic discriminant function for different data sizes ( $k$ )

Full seeds were the group that more resulted in fast germination, whereas slightly injured seeds resulted in equal proportions of ungerminated, slow- and fast-germination (figure 6). Most deformed or severely injured seeds did not germinate.

Figure 7 shows germination results of misclassified seeds (table 1). The greatest error was the number of slow-germinating seeds in the group of slightly injured seeds that were classified in the group of full seeds. The other errors were almost compensated if we consider the proportion of germination for each group

classified according to internal morphology.

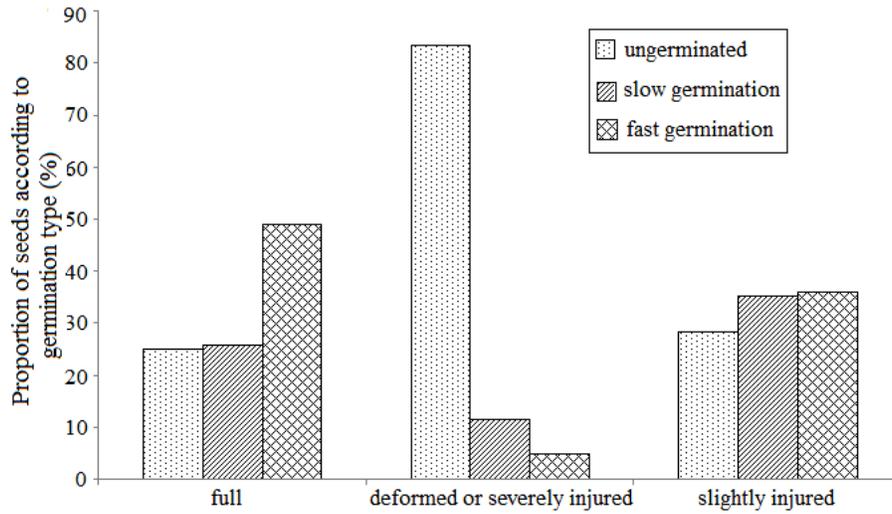


Figure 6 Germination proportions of 445 sunflower seeds separated by levels of physical quality

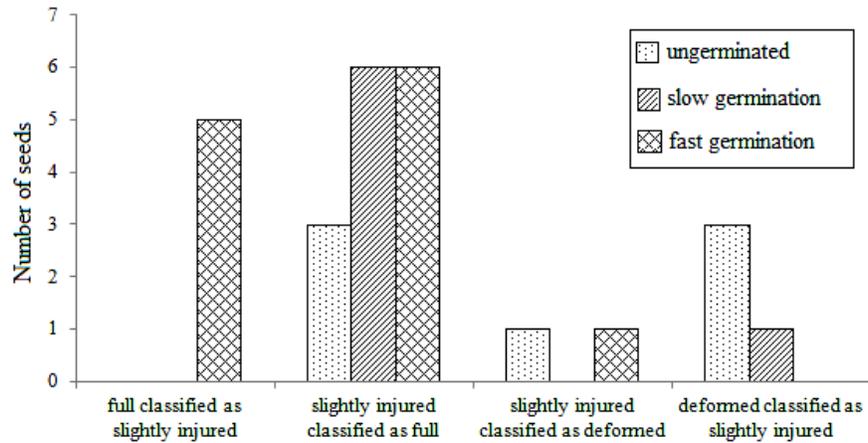


Figure 7 Germination results of misclassified sunflower seeds for data size  $k = 30$

## Discussion

Variation in data size with application of ICA interferes with the accuracy of automatic classification. According to Hyvärinen *et al.* (2001), a problem that often arises is how to determine the number of ICs to be estimated. Often, the dimension is actually chosen by trial and error with no theoretical guidelines. In this work, we observed that the high number of features extracted from images allows for small detail perception; however, it also increases the error rate by classifying full seeds as slightly injured seeds. Our choice was the test that obtained the best overall proportion of seeds that were accurately identified, using a data size of 30 ICs.

The classification performed by DA depends on correct prior classification of the X-ray images by the analyst. Different results in different tests confirm the subjectivity on the visual examination. Confusion between full and slightly injured seeds was already expected due to similarity in these groups and image accuracy. It should be borne in mind that these groups are actually very similar, and a more detailed image may show signs that can be confused with little seed damage. Classification of slightly injured seeds could lead to bad predictions regarding germination speed, although it could also indicate germination potential of seed lots.

There are several studies using X-ray analysis to associate seed morphology and seedling performance. According to Simak (1991), the degree of success of this analysis depends on the species. A relationship between morphology and vigour has been verified for tomato (*Solanum lycopersicum* L.) (Van der Burg *et al.*, 1994), maize (*Zea mays* L.) (Cícero *et al.*, 1998), pepper (*Capsicum annuum* L.) (Dell'Aquila, 2007), bell pepper (*Capsicum annuum* L.) (Gagliardi and Marcos-Filho, 2011), papaya (*Carica papaya* L.) (Santos *et al.*, 2009) and castor

bean (*Ricinus communis* L.) (Carvalho *et al.*, 2010) seeds. In most of these papers, the parameters for evaluation were the presence of internal seed structures, essential to germination, and the proportion of the area occupied by the embryo and endosperm in relation to the area of free space within the seed. In this work, these characteristics were considered initially by visual classification of X-ray images, forming a consistent data base for classification through DA, but the analysis was performed automatically with image features extracted by ICA.

In conclusion, ICA proved to be an appropriate technique for extracting features from radiographic images of seeds. In addition, discriminant analysis provided satisfactory classification of images based on extracted features. Much can still be improved to obtain images of higher quality and give faster processing to ensure an efficient data base to be used in the classification.

### **Acknowledgements**

To Capes - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional Científico e Tecnológico and FAPEMIG - Fundação de Amparo à Pesquisa do Estado de Minas Gerais for the financial support. The authors would like to thank the reviewers for their valuable comments to improve the quality of the paper.

### **References**

- Brasil. (2009). Teste de germinação [The germination test]. In *Regras para Análise de Sementes*, pp. 147-224, Ministério da Agricultura, Pecuária e Abastecimento/ACS, Brasília.
- Carvalho, M.L. M., Alves, R.A. and Oliveira, L.M. (2010). Radiographic analysis in castor bean seeds (*Ricinus communis* L.). *Revista Brasileira de Sementes*, **32**, 170-175.
- Campos, L.F.A., Barros, A.K. and Silva, A.C. (2007). Independent component analysis and neural networks applied for classification of malignant, benign and normal tissue in digital mammography. *Methods of Information in Medicine*, **46**, 212-215.

- Cícero, S.M., Van Der Heijden, G.W.A.M., Van Der Burg, W.J. and Bino, R.J. (1998). Evaluation of mechanical damage in seeds of maize (*Zea mays* L.) by X-ray and digital imaging. *Seed Science and Technology*, **26**, 603-612.
- Dell'Aquila, A. (2006). Computerised seed imaging: a new tool to evaluate germination quality. *Communications in Biometry and Crop Science*, **1**, 20-31.
- Dell'Aquila, A. (2007). Pepper seed germination assessed by combined X-radiography and computer-aided imaging analysis. *Biologia Plantarum* **51**, 777-781.
- Fiori, S. (2003). Overview of independent component analysis technique with an application to synthetic aperture radar (SAR) imagery processing. *Neural Networks*, **16**, 453-467.
- Gagliardi, B. and Marcos-Filho, J. (2011). Relationship between germination and bell pepper seed structure assessed by the X-ray test. *Scientia Agricola*, **68**, 411-416.
- Herault, J. and Jutten, C. (1986). Space or time adaptive signal processing by neural network models. In *Neural Networks for Computing. Proceedings of AIP Conference*, pp. 206-211, American Institute of Physics, New York.
- Hyvärinen, A., Karhunen J. and Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons Inc., New York.
- Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, **9**, 1483-1492.
- ISTA (2011). *International Rules For Seed Testing*, International Seed Testing Association, Bassersdorf, Switzerland.
- Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.
- Saidi, S.A., Holland, C.M., Kreil, D.P., Mackay, D.J., Charnock-Jones, D.S., Print, C.G. and Smith, S.K. (2004). Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, **23**, 6677-6683.
- Santos, S.A., Silva, R.F., Pereira, M.G., Machado, J.C., Borém, F.M., Gomes, V.M. and Tonetti, O.A.O. (2009). X-ray technique application in evaluating the quality of papaya seeds. *Seed Science and Technology*, **37**, 776-780.
- Simak, M. and Gustafsson, A. (1953). X-ray photography and sensitivity in forest tree species. *Hereditas*, **39**, 458-468.
- Simak, M. (1991). Testing of forest tree and shrub seeds by X-radiography. In *Tree and Shrub Seed Handbook* (eds. A.G. Gordon, P. Gosling and B.S.P. Wang), pp. 1-28, International Seed Testing Association, Bassersdorf, Switzerland.
- Tang, S., Wang, Y. and Chen, Y. (2012). Application of ICA to X-ray coronary digital subtraction angiography. *Neurocomputing*, **79**, 168-172.
- Tiilikainen, J., Bosund V., Tilli, J-M., Sormunen, J., Mattila, M., Hakkarainen, T. and Lipsanen, H. (2007). Genetic algorithm using independent component analysis in x-ray reflectivity curve fitting of periodic layer structures. *Journal of Physics D: Applied Physics*, **40**, 6000-6004.
- Van der Burg, W.J., Aartse, J.W., Van Zwol, R.A., Jalink, H. and Bino, R.J. (1994). Predicting tomato seedling morphology by X-ray analysis of seeds. *Journal of American Horticulture Science*, **119**, 258-263.

## CONSIDERAÇÕES FINAIS

Com este trabalho, propôs-se uma nova metodologia a ser usada na análise de imagens radiográficas de sementes. Pode-se considerar que a análise de componentes independentes foi eficaz na extração de características das imagens de raios X. O estudo de simulação descrito no primeiro artigo demonstrou isso pelo desempenho do classificador que obteve ótimos resultados ao usar as características extraídas por ICA como parâmetros de entrada. Quando aplicado a novas sementes reais, os resultados refletiram que o uso da análise discriminante ainda permanece vulnerável à subjetividade da classificação prévia realizada visualmente pelo analista da imagem, sendo esta uma das possíveis causas nos erros identificados na discriminação das sementes.

Nos diferentes testes que se realizam em avaliação de lotes de sementes é importante observar as peculiaridades de cada espécie. Desta forma deve-se também considerar que a abrangência deste estudo está limitada a sementes de girassol, apesar de na maioria das vezes ter sido usado o termo “sementes” de forma genérica. As associações observadas entre classificação física e potencial germinativo neste estudo ratificam resultados de trabalhos anteriores com sementes de girassol.

Como perspectiva de trabalhos futuros pode-se estender a aplicação desta metodologia a sementes de outras espécies. Outra proposta é utilizar diferentes classificadores, como rede neurais e/ou máquina de vetor de suporte, que já mostraram ser eficientes em outros tipos de aplicações.

Uma vez que o uso da metodologia proposta mostrou-se viável, deve-se incrementar a aplicação de técnicas de processamento de imagens que auxiliem na automatização da aquisição e no melhoramento da qualidade das imagens individuais das sementes. Desta forma todo o processo de avaliação pode-se tornar

mais rápido e seguro, contribuindo significativamente na análise de sementes de girassol e de outras espécies.