ITALIAN NATIONAL RESEARCH COUNCIL
"NELLO CARRARA" INSTITUTE FOR APPLIED PHYSICS
CNR FLORENCE RESEARCH AREA
Italy

TECHNICAL, SCIENTIFIC AND RESEARCH REPORTS

Vol. 1 - n. 64-1 (2009)

Leonardo Ciaccheri

Elaborazione di dati spettrali provenienti da sensori ottici tramite analisi multivariata

CNR-IFAC-TR-02/009

ISSN 2035-5831

Consiglio Nazionale delle Ricerche

Istituto di Fisica Applicata "Nello Carrara"

VIA MADONNA DEL PIANO, 10- BUILDING B – I50019 SESTO FIORENTINO (FI) - ITALY

Sintesi dell'attività svolta da Leonardo Ciaccheri nell'ambito del contratto di prestazione d'opera in regime di collaborazione coordinata e continuativa, Marzo 2008 – Febbraio 2009

Elaborazione di dati spettrali provenienti da sensori ottici tramite analisi multivariata

a cura di
Leonardo Ciaccheri

- Febbraio 2009 -

Sommario

1. Introduzione

La presente relazione riassume i risultati della sperimentazione svolta presso CNR-IFAC finalizzata ad idententificare la configurazione opto-geometrica e l'elaborazione dei dati ottimali per la rivelazione del tipo e della concentrazione di campione cellulare biologico.

- Sono stati considerati due tipi di campioni biologici, rispettivamente urina e pap-test in soluzione fisiologica. Per ogni tipo di campione sono state considerate 10 concentrazioni nell'intervallo 20-1000 cellule/mm^3, come riassunto in Tabella I.
- Per ogni tipo di campione sono state eseguite misure di assorbimento e di diffusione a 5 lunghezze d'onda di illuminazione e 4 angoli di rivelazione, tramite le quali sono state ottenute le curve di risposta in funzione della concentrazione cellulare.
- L'elaborazione di tali dati spettrali tramite tecniche di analisi multivariata ha permesso di identificare la tipologia del campione biologico.

Campioni biologici analizzati: urina e pap-test in soluzione fisiologica			
Campione #	Codice	Concentrazione (cells/mm^3)	Colore solo per i campioni di urina; mentre i campioni di pap-test sono sempre trasparenti
1	5000	20	Trasparente
2	2500	40	Giallo
3	1200	83	Giallo
4	900	111	Trasparente
5	700	143	Trasparente
6	500	200	Trasparente
7	300	333	Giallo chiaro
8	200	500	Giallo chiaro
9	150	667	Giallo chiaro
10	100	1000	Giallo chiaro

Tabella I. Caratteristiche dei campioni biologici sperimentati: urina e pap-test in soluzione fisiologica con concentrazioni cellulari nell'intervallo 20-1000 cells/mm^3.

2. Descrizione dell’apparato sperimentale di misura

La Figura 2.1 schematizza la configurazione del dispositivo optoelettronico utilizzato per effettuare le misure di assorbimento e di diffusione sui particolati biologici, mentre la Figura 2.2 illustra la sua realizzazione pratica.

- Il campione biologico è contenuto in una fiala di vetro (diametro esterno: 23 mm – altezza: 70 mm) alloggiata in un supporto cilindrico. Sulla superficie laterale del supporto sono inseriti una sorgente e quattro rivelatori.
- La sorgente è un LED o un diodo laser. Sono stati sperimentati 4 tipi differenti di LED, con emissione rispettivamente a 405, 525, 644 e 850 nm, ed un diodo laser con emissione a 635 nm. L’illuminazione del campione biologico contenuto nella fiala è stata effettuata tramite fibra ottica, connessa al ricettacolo contenente il LED o diodo laser ed affacciata alla superficie della fiala.
- I rivelatori sono PIN disposti sullo stesso piano della sorgente, orientati a 0°, 30°, 60° e 90° rispetto al fascio di illuminazione. Il rivelatore allineato con la sorgente effettua misure di assorbimento, mentre gli altri tre rivelatori effettuano misure di diffusione ai vari angoli. Le misure di assorbimento sono state ottenute tramite fibra ottica connesssa al rivelatore, mentre le misure di diffusione sono state effettuate senza utilizzare fibre ottiche, semplicemente posizionando i rivelatori ad i vari angoli.

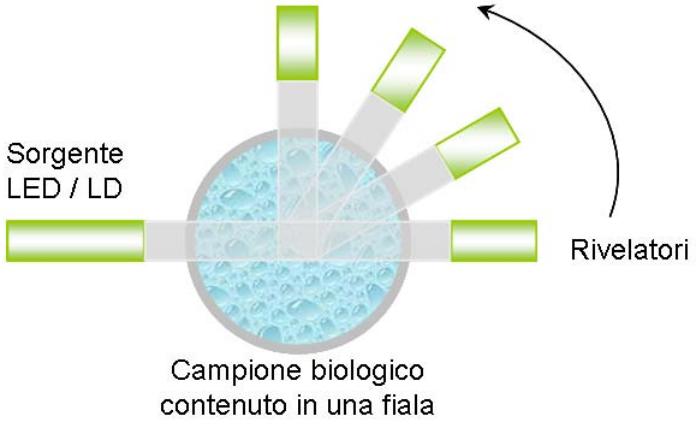


Figura 2.1. Schema del dispositivo optoelettronico utilizzato per effettuare le misure di assorbimento e di diffusione sui particolati biologici.

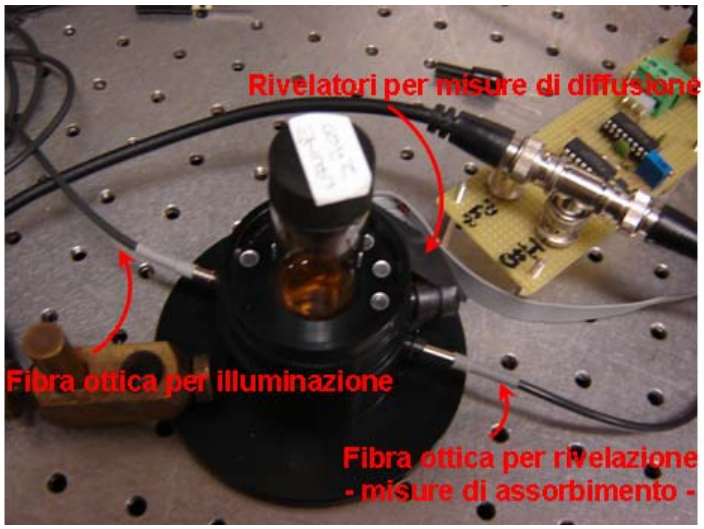


Figura 2.2. Realizzazione pratica del dispositivo optoelettronico utilizzato per effettuare le misure di assorbimento e di diffusione sui particolati biologici.

3. Risultati delle misure sperimentali

Tutti i campioni biologici riassunti nella Tabella I sono stati caratterizzati tramite misure di assorbimento e di diffusione a tutte le lunghezze d'onda e a tutti gli angoli disponibili. Tali misure sono state eseguite come segue:

- Misure di assorbimento: ottenute tramite il rivelatore disposto a 0° rispetto alla sorgente, rivelando l'intensità della luce trasmessa dal campione biologico normalizzata rispetto all'intensità della luce trasmessa da un campione di acqua distillata. Tali misure, pur essendo informative riguardo alla concentrazione del particolato biologico, sono dipendenti anche dal colore del campione.
- Misure di diffusione: ottenute tramite i rivelatori disposti a 30°, 60° e 90° rispetto alla sorgente, rivelando l'intensità della luce diffusa dal campione biologico normalizzata all'intensità della luce trasmessa a 0° dal campione stesso. Questo tipo di normalizzazione consente di ottenere l'informazione riguardante la concentrazione del particolato biologico in maniera indipendente dal colore del campione.

L'apparato sperimentale messo a punto consente di effettuare misure affette da un errore del 10%.

3.1 Misure di assorbimento

La Figura 3.1 illustra i risultati delle misure di assorbimento. Le curve di risposta per il campione di urina risultano differenti alle varie lunghezze, essendo il campione di colore giallo. Al contrario, per il campione di pap-test le curve di risposta alle varie lunghezze d'onda sono simili, non essendo influenzate dal colore trasparente del campione.

Queste curve di risposta sono riproducibili con un fitting dato dalla somma di due curve esponenziali [*]. Dai valori dei coefficienti di correlazione di tali curve, si evince che la curva di risposta che meglio rappresenta i dati sperimentali è ottenuta alla lunghezza d'onda di 850 nm, sia per l'urina che per il pap-test.

Con tale curva di risposta, l'errore percentuale di misura, dipendente dalla concentrazione cellulare, risulta rappresentato nella Fgura 3.2: per basse concentrazioni, inferiori a 200 cells/mm^3, risulta essere molto elevato, mentre assume valori compresi tra il 10% ed il 20% per concentrazioni superiori a 200 cells/mm^3.

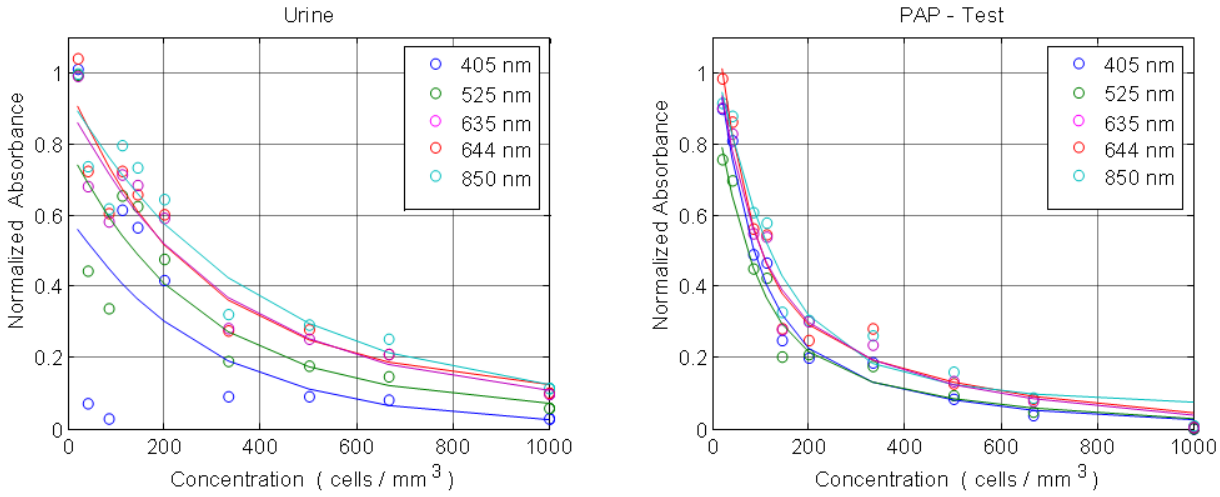


Figura 3.1. Assorbanza normalizzata in funzione della concentrazione cellulare misurata alle varie lunghezze d'onda – Sinistra: campioni di urina; Destra: campioni di pap-test.

[*] vedi Appendice II

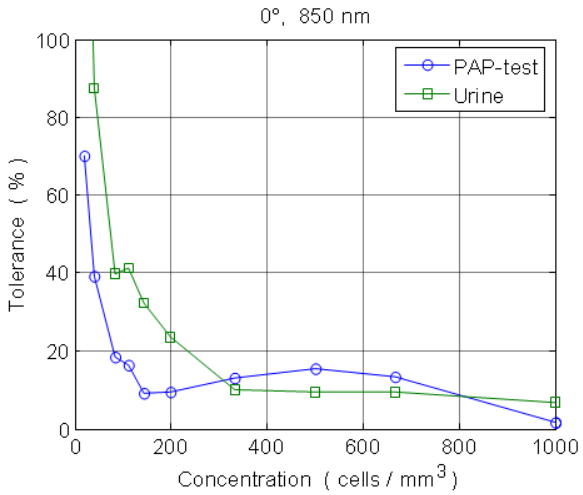


Figura 3.2. Errore percentuale di misura considerando la curva di calibrazione a 850 nm come curva rappresentativa dei dati misurati.

3.2 Misure di diffusione

Le Figure 3.3, 3.4 e 3.5 illustrano i risultati delle misure di diffusione, rispettivamente agli angoli 30°, 60° e 90°, eseguite a tutte le lunghezze d'onda. Sia per i campioni di urina, che per quelli di pap-test, tali misure risultano indipendenti dal colore, essendo l'intensità della luce diffusa normalizzata all'intensità della luce trasmessa a 0° dal campione stesso.

Queste curve di risposta sono riproducibili con un fitting dato da una sigmoide [*]. Dai valori dei coefficienti di correlazione di tali curve, si evince che la curva di risposta che meglio rappresenta i dati sperimentali è ottenuta alla lunghezza d'onda di 850 nm, sia per l'urina che per il pap-test.

Con tale curva di risposta, l'errore percentuale di misura, dipendente dalla concentrazione cellulare, risulta rappresentato nella Fgura 3.6: per basse concentrazioni, inferiori a 200 cells/m^3 risulta essere variabile tra il 25% ed il 10%, mentre assume valori inferiori al 7% per concentrazioni superiori a 200 cells/mm^3.

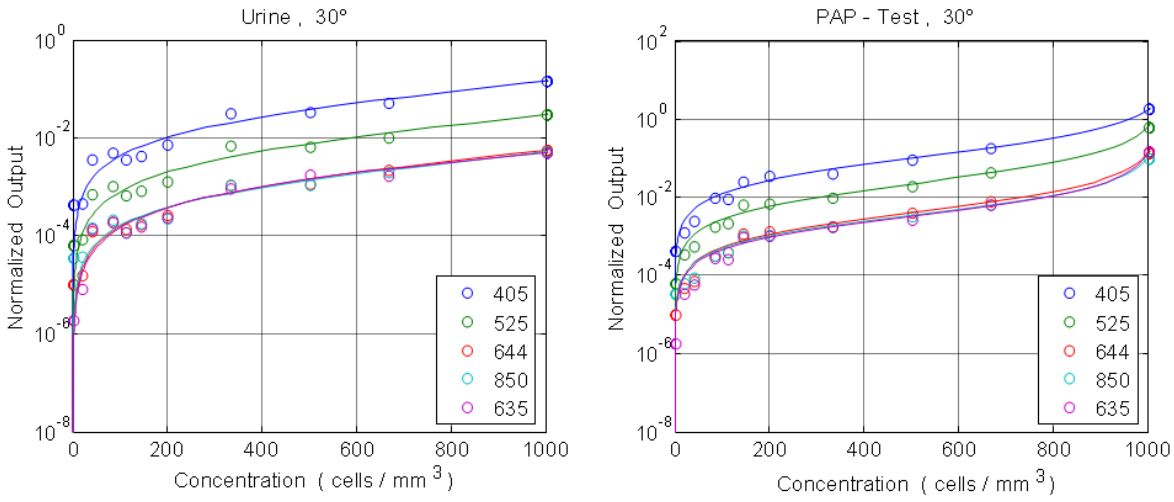
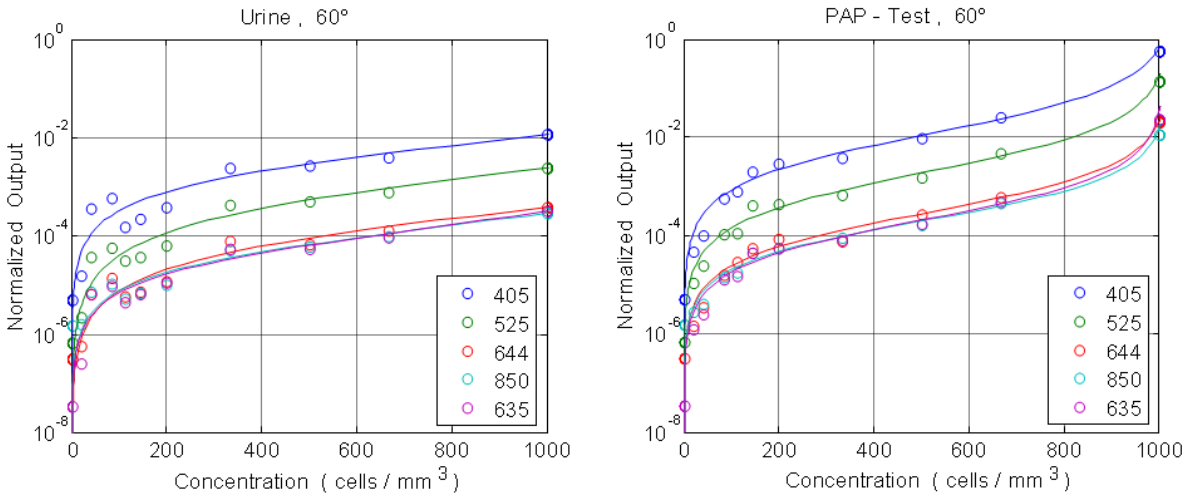


Figura 3.3. Intensità diffusa a 30°, normalizzata all'intensità trasmessa a 0°, in funzione della concentrazione cellulare, misurata alle varie lunghezze d'onda – Sinistra: campioni di urina; Destra: campioni di pap-test.

[*] vedi Appendice II



F igura 3.4. Intensità diffusa a 60°, normalizzata all'intensità trasmessa a 0°, in funzione della concentrazione cellulare, misurata alle varie lunghezze d'onda – Sinistra: campioni di urina; Destra: campioni di pap-test.

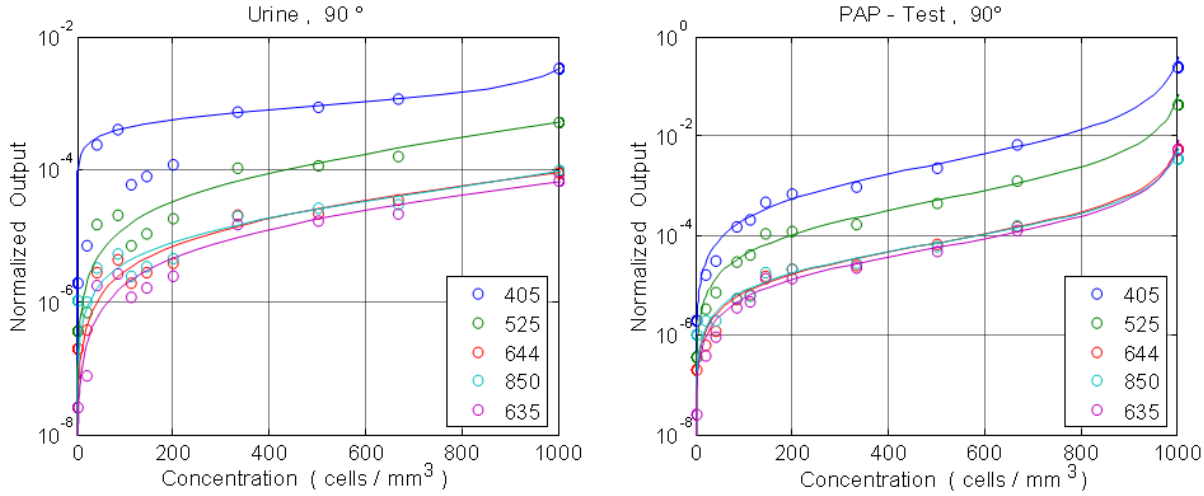


Figura 3.5. Intensità diffusa a 90°, normalizzata all'intensità trasmessa a 0°, in funzione della concentrazione cellulare, misurata alle varie lunghezze d'onda – Sinistra: campioni di urina; Destra: campioni di pap-test.

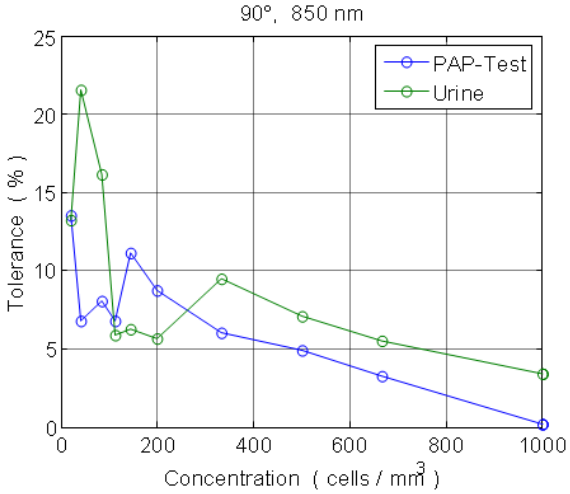


Figura 3.6. Errore percentuale di misura considerando la curva di calibrazione a 90° ed 850 nm come curva rappresentativa dei dati misurati.

4. Elaborazione dei dati spettrali – discriminazione del campione biologico tramite analisi multivariata [*]

I risultati mostrati nelle precedenti sezioni evidenziano le differenti caratteristiche diffusive dei campioni di pap-test rispetto a quelli di urina. È quindi possibile, in linea di principio sfruttare queste differenze per realizzare un algoritmo di classificazione del tipo di particolato, indipendentemente dalla concentrazione.

Allo scopo di rendere l'analisi indipendente dall'intensità delle sorgenti è stato innanzitutto normalizzata l'intensità misurata all'angolo di scattering θ ed alla lunghezza d'onda λ, dividendola per quella misurata a 0°, alla stessa lunghezza d'onda. Si sono poi calcolati i rapporti tra le intensità normalizzate relative allo stesso angolo di scattering e a due diverse lunghezze d'onda, λ_1 e λ_2. Ciò consente, in approssimazione di scattering multiplo al 1° ordine, di fornire una quantità dipendente solo dal rapporto tra le sezioni differenziali di scattering relative alle due configurazioni considerate e quindi indipendente dalla concentrazione.

Sia $S(\theta, \lambda_1, \lambda_2)$ la quantità di cui sopra, dove θ può assumere i valori 30°,60° o 90°; fissate le due lunghezze d'onda, ogni campione sarà caratterizzato da un vettore 1×3, ottenendo pertanto una matrice di dati 20×3. Tale matrice è stata calcolata per ogni possibile coppia di lunghezze d'onda e quindi, tramite la *Principal Component Analysis* (PCA), si sono proiettati i punti rappresentativi dei vari campioni in una mappa bidimensionale, per valutare l'eventuale separazione tra le due classi di preparati.

Le mappe con miglior capacità discriminante si sono rivelate essere quelle relative alle coppie di lunghezze d'onda (405 nm, 525 nm) e (525 nm, 644 nm). Tuttavia, come si può apprezzare in Figura 4.1, nessuna delle due, singolarmente, è in grado di fornire una discriminazione netta dei due tipi di particolato.

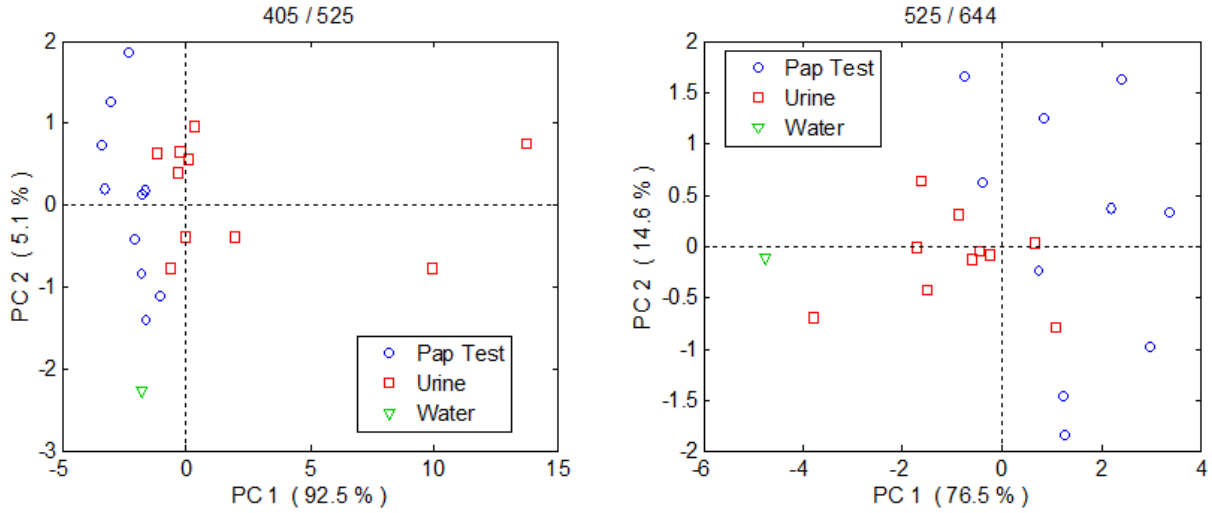


Figura 4.1. Risultati dell'elaborazione dei dati spettrali tramite PCA.

Per meglio discriminare le due differenti tipologie di campione biologico è stato eseguito il prodotto cartesiano dei sottospazi rappresentati nelle due mappe di Fig. 4.1, creando quindi uno spazio quadridimensionale all'interno del quale è stata applicata la *Linear Discriminant Analysis* (LDA) trovando quindi la direzione di miglior separazione tra i due *cluster* di punti.

Il risultato è mostrato in Figura 4.2-sinistra, dove sono mostrate le proiezioni dei punti rappresentativi lungo l'asse di miglior separazione (DF1 = *Discriminating Function* 1); la stessa quantità è stata rappresentato sia sulle ascisse che sulle ordinate per maggior chiarezza. La Figura 4.2-destra mostra invece l'istogramma della distribuzione dei punti lungo tale asse.

[*] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. De Jong, D.J. Lewi, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics*, Elsevier Science BV, Amsterdam, 1998.
M.J. Adams, *Chemometric in analytical spectroscopy*, Royal Society of Chemestry, Cambridge, 1995.

Per valutare la capacità discriminante del metodo è stata fatta una *cross-validation*, usando 4/5 dei campioni, a rotazione, come *training set* per classificare il restante quinto. La prova ha dato come risultato un 85% di campioni correttamente classificati.

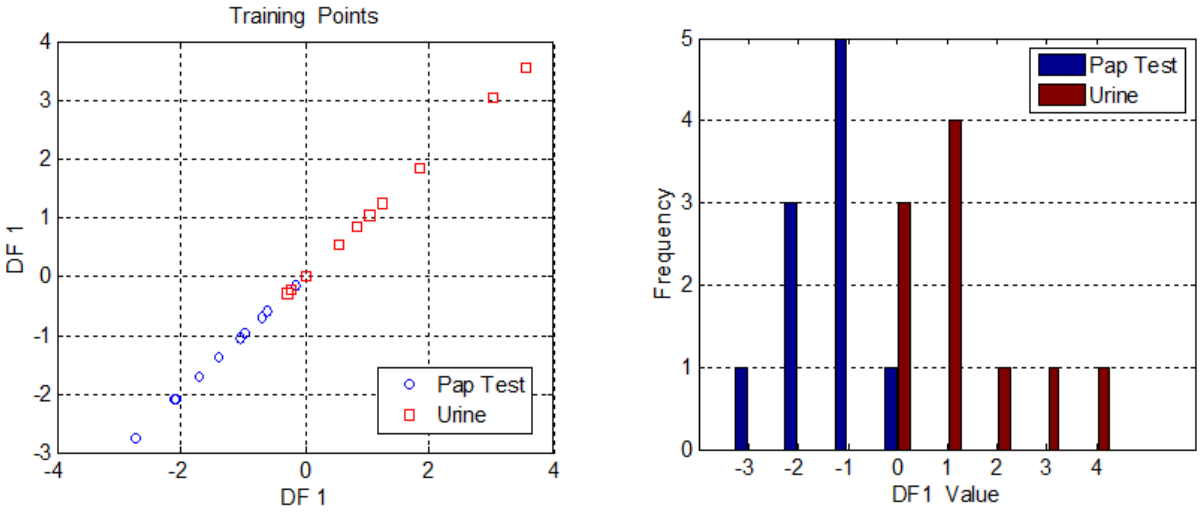


Figura 4.2. Discriminazione del campione diologico tramite elaborazione LDA.

5. Conclusioni e prospettive

Le misure di assorbimento e diffusione effettuate alle 5 lunghezze d'onda ed ai 4 angoli individuati, opportunamente eleborate con tecniche di analisi multivariata PCA ed LDA, consentono di discriminare la tipologia del campione biologico, urina o pap-test.

Individuato il tipo di campione biologico, la determinazione della concentrazione può essere ottenuta tramite una selezione delle misure, come segue:

- La miglior curva di calibrazione derivante dalle misure di assorbimento, normalizzate all'intensità trasmessa da un campione di acqua distillata, è stata ottenuta alla lunghezza d'onda di 850 nm:
  - è costituita dalla somma di due curve esponenziali e dipende dal colore del campione biologico analizzato;
  - presenta un errore percentuale di misura dipendente dalla concentrazione cellulare, molto elevato per concentrazioni inferiori a 200 cells/mm^3, mentre assume valori compresi tra il 10% ed il 20% per concentrazioni superiori a 200 cells/mm^3.
- La miglior curva di calibrazione derivante dalle misure di diffusione ai vari angoli, normalizzate all'intensità trasmessa a 0°, è stata ottenuta a 90° alla lunghezza d'onda di 850 nm:
  - è costituita da una sigmoide e non dipende dal colore del campione biologico analizzato;
  - presenta un errore percentuale di misura dipendente dalla concentrazione cellulare, compreso tra il 25% ed il 10% per concentrazioni inferiori a 200 cells/mm^3, mentre assume valori inferiori al 7% per concentrazioni superiori a 200 cells/mm^3.

Appendice I – Specifiche tecniche di sorgenti e rivelatori

Componente	**Tipo**	**Indirizzo web**
Diodo laser λ = 635 nm	RLT6303MG	http://www.roithner-laser.com/All_Datasheets/Laserdiodes/RLT6303MG.pdf
LED λ = 405 nm	RLU405-9-30	http://www.roithner-laser.com/All_Datasheets/LEDs/LED375_385_395_405_series.pdf
LED λ = 525 nm	B5-433-B525	http://www.roithner-laser.com/All_Datasheets/LEDs/B5-433-B525.pdf
LED λ = 644 nm	Toshiba TLRH180P Codice RS 2609441	http://www.rs-components.it/
LED λ = 850 nm	HIR33	http://www.roithner-laser.com/All_Datasheets/LEDs/hir333.pdf
Rivelatore per misure di assorbimento	V500	http://www.optiphase.com/data_sheets/v-600_datasheetrevb1.pdf
Rivelatori per misure di diffusione	TSL12S	http://www.taosinc.com/

Appendice II – Dettagli sulle caratteristiche delle curve di calibrazione

Misure di assorbimento

Le curve di risposta sono riproducibili con un fitting dato dalla somma di due curve esponenziali:

$$\mathbf{P_0(\,C\,) = A\ exp\,(\,k_1\,C\,) + B\ exp(\,k_2\,C\,)}$$

Dove $P_0(\,C\,)$ rappresenta l'intensità trasmessa dal campione biologico, normalizzata a quella trasmessa da un campione di acqua distillata. La tabella sottostante riassume i valori delle quattro costanti A, B, k_1 e k_2, calcolate con una procedura iterativa.

PAP-Test Fit Coefficients

	k1	A	k2	B
405	-0.0133	0.8800	-0.0025	0.2703
525	-0.0130	0.7042	-0.0023	0.2587
644	-0.0151	0.8689	-0.0022	0.3868
850	-0.0080	0.9422	-0.0007	0.1474
635	-0.0139	0.7181	-0.0024	0.4101

Urine Fit Coefficients

	k1	A	k2	B
405	-0.0035	0.5913	-0.0001	0.0096
525	-0.0038	0.7075	-0.0004	0.0863
644	-0.0042	0.7551	-0.0006	0.2161
850	-0.0027	0.8573	-0.0002	0.0821
635	-0.0035	0.7239	-0.0008	0.1902

Misure di diffusione

Le curve di risposta sono riproducibili con un fitting dato dalla sigmoide:

$$\mathbf{C\,(\,x\,) = A\ x^{\,k}\ /\ (\,B + x^{\,k}\,)}$$

dove $x = P\,(\,\theta\,) / P\,(\,0\,)$ è l'intensità della luce diffusa all'angolo θ, normalizzata all'intensità della luce trasmessa a 0°. La tabella sottostante riassume i valori delle tre costanti A, B e k, calcolate con una procedura iterativa.

**** PAP-TEST ****				**** URINE ****			
config.	A	B	k	config.	A	B	k
30-405	1053	9.9e-002	1.05	30-405	1828	1.6e-001	0.86
30-525	1032	2.0e-002	1.04	30-525	1894	4.9e-002	0.83
30-644	1027	3.2e-003	1.05	30-644	1842	1.1e-002	0.84
30-850	1024	1.9e-003	1.11	30-850	1770	7.0e-003	0.89
30-635	1016	2.0e-003	1.10	30-635	2704	2.7e-002	0.78
60-405	1033	2.3e-002	0.85	60-405	1832	1.9e-002	0.85
60-525	1023	5.0e-003	0.85	60-525	2254	1.9e-002	0.70
60-644	1023	7.1e-004	0.89	60-644	2217	3.8e-003	0.73
60-850	1016	2.8e-004	0.96	60-850	1597	4.5e-004	0.88
60-635	1011	5.0e-004	0.90	60-635	1484	4.3e-004	0.87
90-405	1018	7.2e-003	0.84	90-405	1034	4.6e-009	2.76
90-525	1018	1.5e-003	0.86	90-525	1786	1.4e-003	0.84
90-644	1015	1.3e-004	0.94	90-644	2027	3.6e-004	0.85
90-850	1017	9.1e-005	0.98	90-850	1484	3.0e-005	1.05
90-635	1012	1.1e-004	0.93	90-635	2203	5.5e-004	0.80

Appendice III – Programmi realizzati in Matlab per l'elaborazione dei dati spettrali con PCA ed LDA

```
% HOSPI_LDA_SCRIPT  Linear Discriminant Analysis for Pap Test vs. Urine
%                   samples discrimination.
clear

% Initialization
w1 = 405;
w2 = 525;
w3 = 644;

% PCA routines
[pc12,I,J,vlab,subdiv] = hospi_ratio2_pca(w1,w2);
[pc23,dummy1,dummy2,dummy3,dummy4] = hospi_ratio2_pca(w2,w3,0);
clear dummy*

% Data Fusion
X = cat(2,pc12.scr,pc23.scr);
X(end,:) = [];
subdiv(end,:) = '';

% LDA
ok = crval_lda(X,subdiv,5);
ldm = crval_lda(X,subdiv,1,1);
disp(' ')
disp(['SECV = ',num2str(100-ok),'%'])

% Histogram
x = linspace(-3,4,8)';
p = subdiv(:,1) == 'P';
u = subdiv(:,1) == 'U';
Np = hist(ldm.scr(p),x);
Nu = hist(ldm.scr(u),x);
figure
bar(x,[Np;Nu]')
set(gca,'fontsize',14,'ygrid','on')
legend(subdiv(1,:),subdiv(end,:))
xlabel('DF1  Value')
ylabel('Frequency')
```

```
% HOSPI_RATIO2_PCA  Principal Component Analysis of two-wavelebths
%                   scattering data.
%
%   The program evaluates the ratio of scattering cross-sections at
%   two diffrent wavelength, 'w1' and 'w2', for three different scattering
%   angles: 30°, 60° and 90°. Then PCA is carried on the data matrix.
%
%   Available wavelengths (nm): 405, 525, 644, 850, 635 (laser).
function [pcm,I,J,vlab,subdiv] = hospi_ratio2_pca(w1,w2,clw)

% Input checks
if nargin < 3, clw = 1; end
if clw, close all, end

% Data loading
[Sp,sub_p,cp,wl,th] = build_mat('auto_pap','Pap Test');
[Su,sub_u,cu,wl,th] = build_mat('auto_uri','Urine');

% Concatenating
S = cat(1,Sp,Su);
subdiv = strvcat(sub_p,sub_u);
conc = cat(1,cp,cu);

% Initialization
N = size(S,1);
M = size(S,2);
olab = int2str(conc);
vlab = int2str(th(1:4));

% Subdivinding samples by type
[I{1},I{2}] = logind(subdiv);
I{3} = ['ob';'sr';'vg'];
[J{1},J{2}] = logind(int2str(wl));
J{3} = ['ob';'sr';'vg';'dm';'pc'];

% Ratio of outputs
u1 = find(wl([1 4 7 10 13]) == w1);
st1 = 3 * (u1 - 1) + 1;
ed1 = 3 * u1;
u2 = find(wl([1 4 7 10 13]) == w2);
st2 = 3 * (u2 - 1) + 1;
ed2 = 3 * u2;
X = S(:,st1:ed1) ./ S(:,st2:ed2);

% Pre processing
[X0,stat] = prepro(X,1);

% PCA
pcm = mypca05(X0,2);
pcm.stat = stat;
pcm.data = X;

% Plots
pcaplot(pcm.scr,1,2,pcm.eig,olab,I)
title([int2str(w1),' / ',int2str(w2)])
pcaplot(pcm.ldg,1,2,pcm.eig,vlab)
title([int2str(w1),' / ',int2str(w2)])

%                 *** End of main module ***
```

```
% ------------------------------------------------------------

% BUILD_MAT  Data matrix assembling routine
function [X,subdiv,conc,wl,th] = build_mat(filename,mark)

load(filename)
N = size(P,1);
K = size(P,2) - 1;
M = size(P,3);

% Unfolding patterns
for n = 1:N
    temp = P(n,2:end,1) / P(n,1,1);
    if n == 1
        th = theta(2:end);
        wl = wlen(1) * ones(K,1);
    end
    for m = 2:M
        temp = cat(2,temp,P(n,2:end,m) / P(n,1,m));
        if n == 1
            th = cat(1,th,theta(2:end));
            wl = cat(1,wl,wlen(m)*ones(K,1));
        end
    end
    X(n,:) = temp;
    subdiv(n,:) = mark;
end

% Water pattern
if mark(1) == 'U'
    temp = H(2:end,1)' / H(1,1);
    for m = 2:M
        temp = cat(2,temp,H(2:end,m)' / H(1,m));
    end
    X = cat(1,X,temp);
    subdiv = strvcat(subdiv,'Water');
    conc = cat(1,conc,0);
end
```