主成分分析の適用例

例１．テスト成績の解析
共分散行列を用いた主成分分析
相関行列を用いた主成分分析

例２．ＲＧＢ信号の主成分分析

例３．分光反射率の解析

例４．画像圧縮

例５．パターン認識

例１－テストの成績の分析－
ある中学生の集合のテストの成績を主成分分析する

共分散行列：

	x1（国語	x2（社会	x3（数学	x4（理科	x5（音楽	x6（美術	x7（体育	x8（技家	x9（英語
x1（国語）	470								
x2（社会）	363	459							
x3（数学）	384	407	583						
x4（理科）	333	383	426	462					
x5（音楽）	356	369	421	391	531				
x6（美術）	259	229	245	224	263	300			
x7（体育）	248	113	106	66	186	201	708		
x8（技家）	321	326	320	344	301	209	95	500	
x9（英語）	493	501	576	485	479	304	232	404	874

各教科の平均と
分散

	平均	分散
x1（国語）	52	470
x2（社会）	39	459
x3（数学）	45	583
x4（理科）	50	462
x5（音楽）	42	531
x6（美術）	62	300
x7（体育）	57	708
x8（技家）	47	500
x9（英語）	39	874

得られた主成分を解釈する

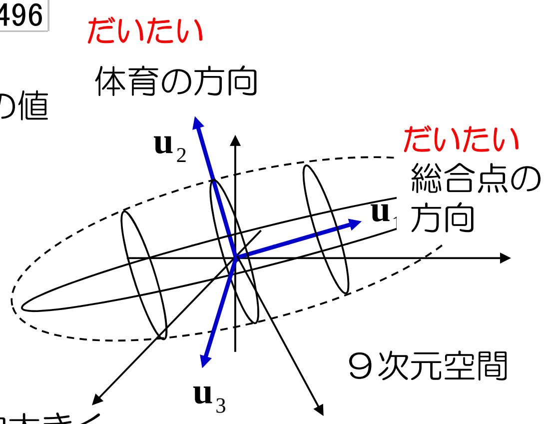
	第1主成分	第2主成分	第3主成分
x1(国語)	0.338	−0.121	0.095
x2(社会)	0.339	0.121	0.034
x3(数学)	0.377	0.174	−0.307
x4(理科)	0.337	0.207	0.099
x5(音楽)	0.349	0.004	−0.003
x6(美術)	0.227	−0.153	0.173
x7(体育)	0.172	−0.926	0.012
x8(技家)	0.298	0.126	0.781
x9(英語)	0.472	0.041	−0.496

	固有値
λ1	3232
λ2	700
λ3	260
λ4	208
λ5	124

第１主成分について：
 どの科目の係数も正で体育，英語を除いて同程度の値を取っている.
→ 総合点の意味合いをもつと解釈できる.

第２主成分について：
 体育は，主成分の係数の絶対値が特に大きい.
→ 体育の成績がばらつきを支配する.

第３主成分について：
技術家庭の係数が大きいが英語，数学なども比較的大きく，有意な解釈は与えられない.



データの標準化の必要性について

仮に，「美術」が１０点満点，他は１００点満点でで出されていたとする.

そのまま，単純に共分散行列を求めてしまうと，美術に関連した要素が著しく小さくなる.
この結果，主成分は以下のようになる.

	第1主成分	第2主成分	第3主成分
x1(国語)	0.345	−0.135	−0.112
x2(社会)	0.349	0.108	−0.0432
x3(数学)	0.389	0.157	0.3113
x4(理科)	0.347	0.196	−0.0948
x5(音楽)	0.357	−0.006	0.0444
x6(美術)	0.022	−0.013	−0.0092
x7(体育)	0.168	−0.945	−0.0551
x8(技家)	0.306	0.118	−0.8271
x9(英語)	0.487	0.011	0.4365

どの主成分でも，美術への荷重は小さく，美術がばらつきに寄与していない.

⇩

どの科目もデータのばらつきに等しく寄与させるためには，各科目の成績を（平均０分散１に）標準化してから，主成分分析を始めなければならない.

相関行列に対する主成分分析

変数を標準化して相関行列を作り，これに対して主成分分析を行う.

相関行列：

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \\ r_{m1} & & r_{mm} \end{bmatrix}$$

各要素は標準化した変数間の共分散

$$r_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{x_{ik}-m_i}{\sigma_i}\cdot\frac{x_{jk}-m_j}{\sigma_j}$$

テスト成績の相関行列：

	x1(国語	x2(社会	x3(数学	x4(理科	x5(音楽	x6(美術	x7(体育	x8(技家	x9(英語
x1(国語)	1.000								
x2(社会)	0.782	1.000							
x3(数学)	0.734	0.787	1.000						
x4(理科)	0.715	0.832	0.821	1.000					
x5(音楽)	0.713	0.747	0.757	0.789	1.000				
x6(美術)	0.690	0.617	0.586	0.602	0.659	1.000			
x7(体育)	0.430	0.198	0.165	0.115	0.303	0.436	1.000		
x8(技家)	0.662	0.681	0.593	0.716	0.584	0.540	0.160	1.000	
x9(英語)	0.769	0.791	0.807	0.763	0.703	0.594	0.295	0.611	1.000

相関行列に対する主成分分析

	第1主成分	第2主成分	第3主成分
x1(国語)	0.362	−0.155	−0.057
x2(社会)	0.368	0.150	0.047
x3(数学)	0.358	0.184	0.372
x4(理科)	0.366	0.255	0.014
x5(音楽)	0.353	−0.003	0.236
x6(美術)	0.314	−0.299	−0.193
x7(体育)	0.145	−0.860	0.032
x8(技家)	0.314	0.154	−0.829
x9(英語)	0.359	0.037	0.274

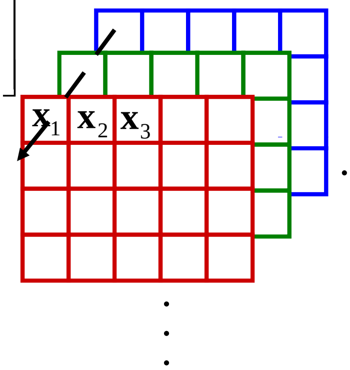
第１主成分について：
どの科目の係数も正で体育，英語を除いて同程度の値を取っている.
→ 総合点の意味合いをもつと解釈できる.

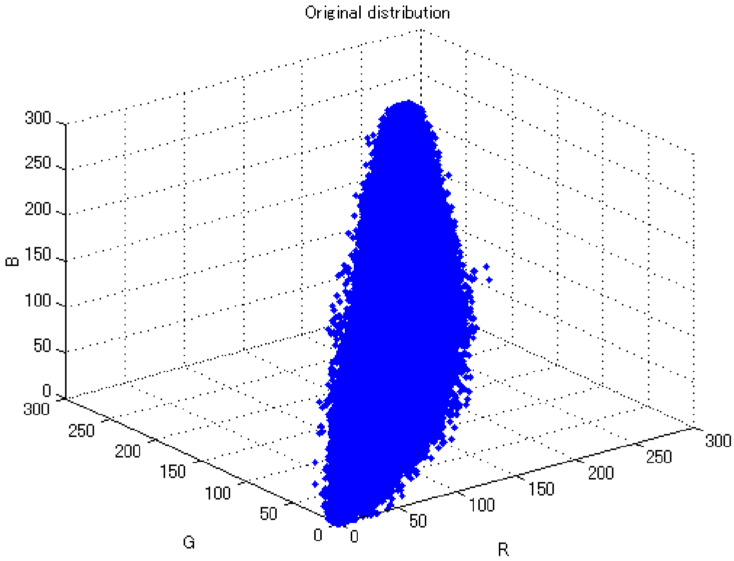
第２主成分について：
体育は，主成分の係数の絶対値が大きい.
美術や理科も比較的大きい.

例２　RGBカラー信号の主成分分析

オリジナル画像

$$\mathbf{x}_1 = \begin{bmatrix} R_1 \\ G_1 \\ B_1 \end{bmatrix}$$



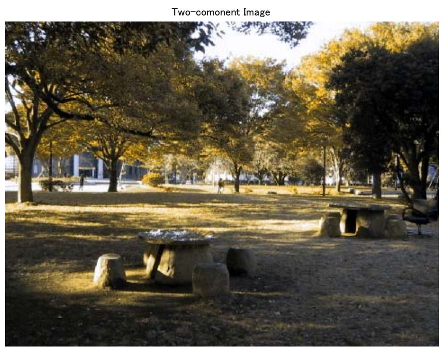


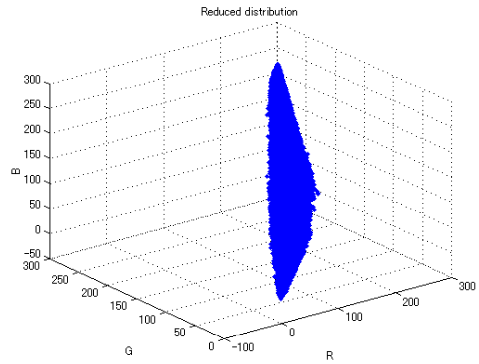
RGB空間での画素値の分布

Program name:PCAdemoRGB.m

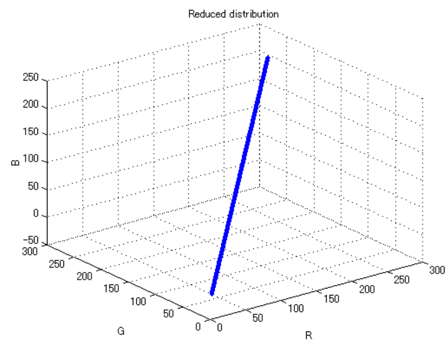
例２　RGBカラー信号の主成分分析

第１および第２主成分のみ





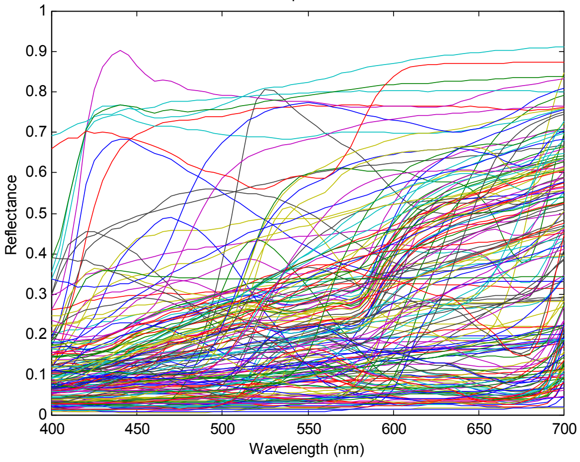
第１主成分のみ

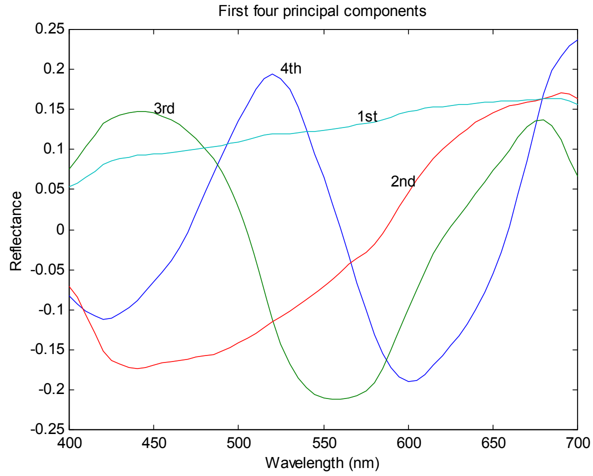


例３　～分光反射率の近似～

波長：400,404,...,700nm -> 61次元

自然界の分光反射率サンプル　→ 主成分分析 →　主成分（第１～第４）

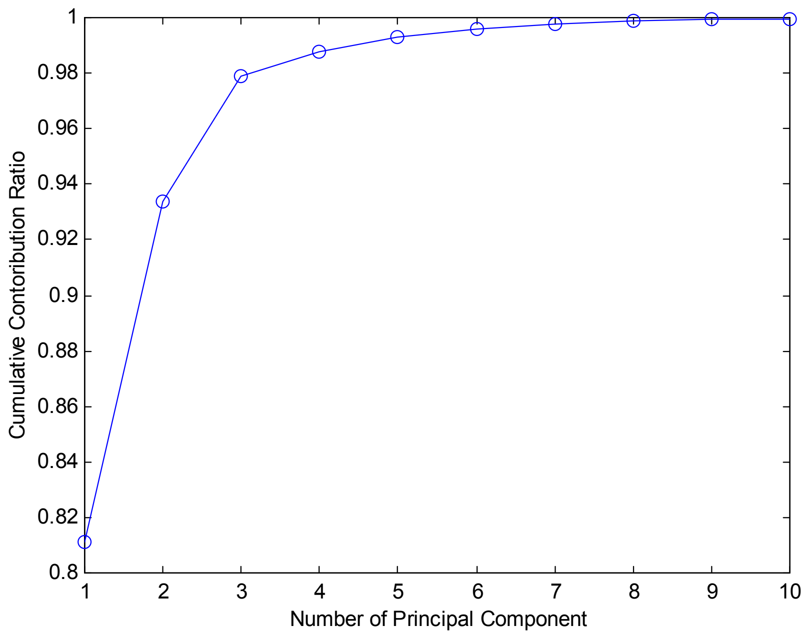




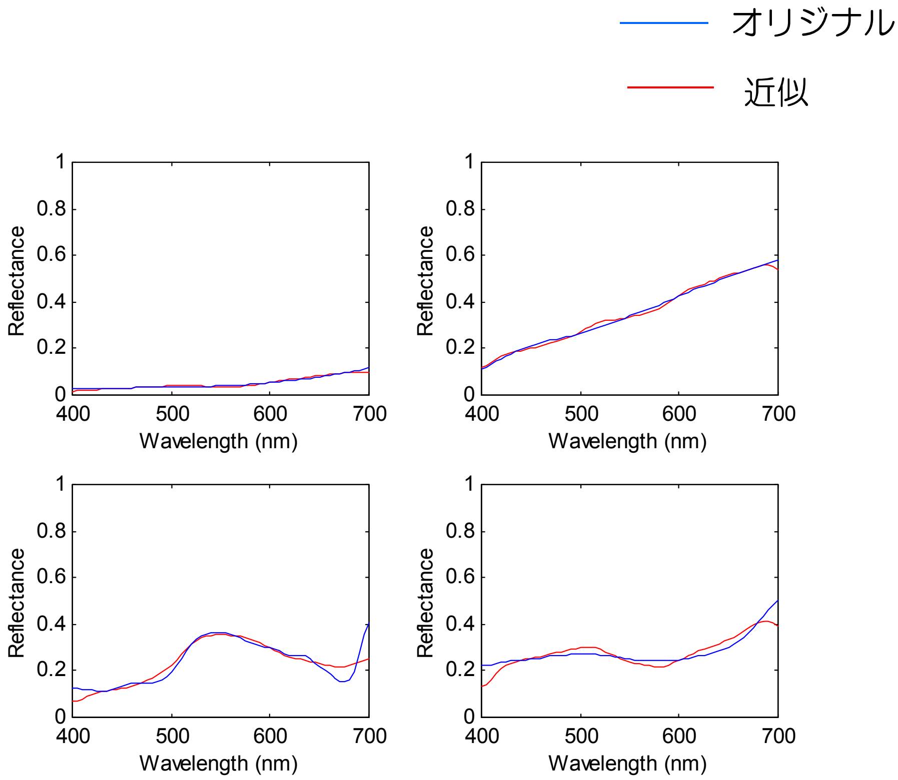
$$r(\lambda) = m(\lambda) + \sum_{i=1}^{61} k_i u_i(\lambda) \longrightarrow r(\lambda) \approx m(\lambda) + \sum_{i=1}^{4} k_i u_i(\lambda)$$

低い次元数で元の信号を表現できる.

分光反射率データの累積寄与率



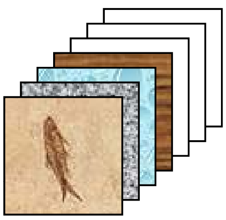
４次までの主成分で近似した例



例４　画像圧縮

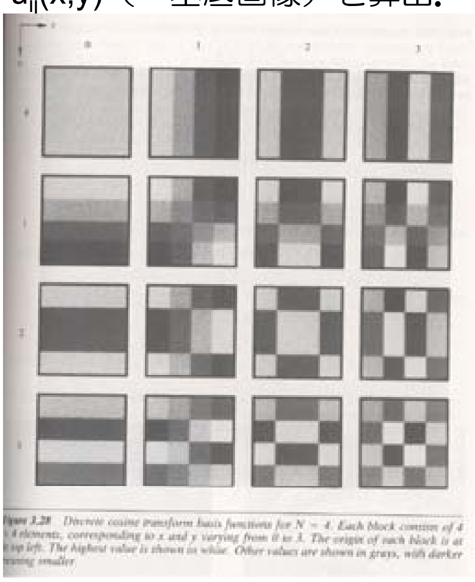
PCA ＝ Hotelling Transform or Karhunen-Loeve transform

オリジナル画像群



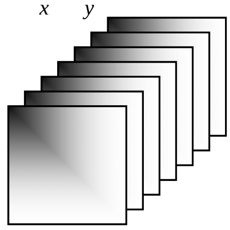
$f(x,y)$

K-L変換（主成分分析）により,
分散の大きい順番に基底ベクトル
$u_{ij}(x,y)$（＝基底画像）を算出.

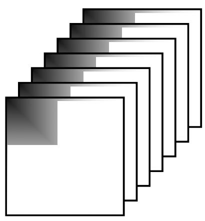


係数算出：（内積演算）

$$k(i,j)=\sum_x \sum_y f(x,y)u_{ij}(x,y)$$

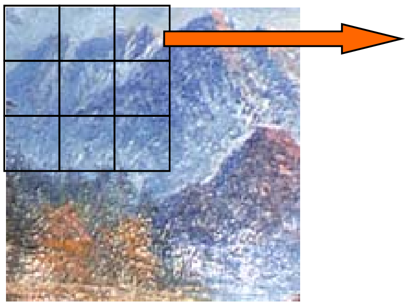


分散の大きい係数のみ保存

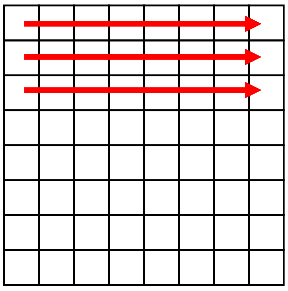


戻すときは逆の手順で

画像の主成分分析例



I番目のブロック（8x8=64画素）



$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_{64}^{(i)} \end{bmatrix}$$ 画素値

画像を小ブロックに分割

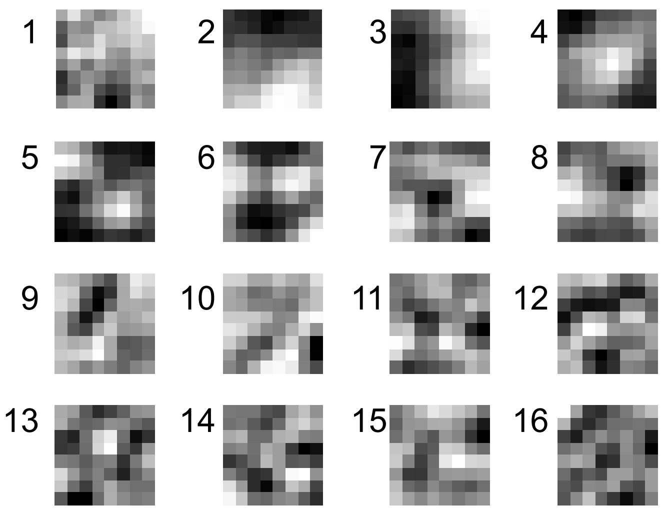
各小ブロック内の画素をラスタスキャンの順に並べて列ベクトルにする

データセット

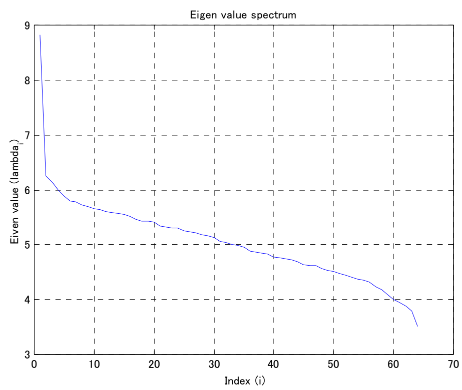
$$\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$$

に対して主成分分析を行い，主成分ベクトルを求め，画像圧縮に用いる.

画像の主成分分析例



固有値の大きい順に並べた
最初の１６枚の主成分画像



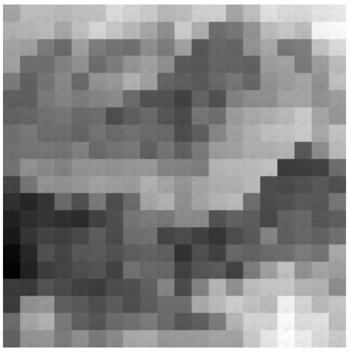
固有値スペクトル

オリジナルおよび復元画像

オリジナル

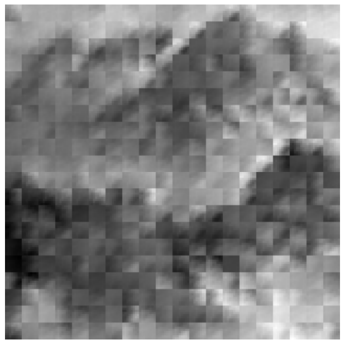
成分数：１

# of synthesized components: 1



成分数：３

# of synthesized components: 3



成分数：１０

# of synthesized components: 10

成分数：２０

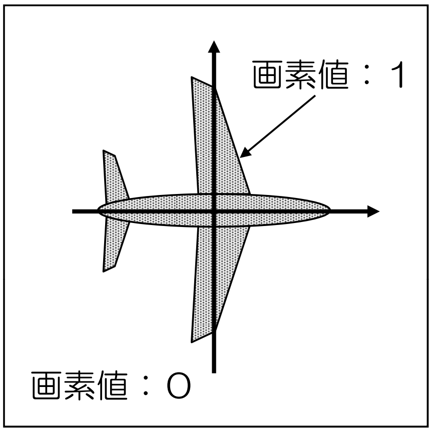
# of synthesized components: 20

成分数：３０

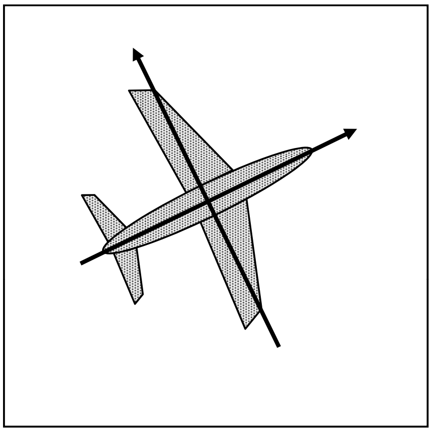
# of synthesized components: 30

例５　パターン認識

２値画像



２値画像



画素値１のプロットの共分散行列
から主成分方向とばらつきを算出

画素値１のプロットの共分散行列
から主成分方向とばらつきを算出

－固有値（λ_1,λ_2）からパターンの一致度がわかる
－固有ベクトルの方向から物体の回転が検出できる