Critic of Fast Approximate Factor Analysis

Daniel D. Beatty

May 27, 2003

Contents

Chapter 1

Karhunen-Loève / Singular Value Decomposition

Algorithm equivalence: singular value decomposition = factor analysis = principle component analysis = Karhunen-Loève transform

"SVD has algebraic interpretation (finding the eigenvalues of a matrix) and an analytic interpretation (finding the minimum of a cost function over the set of orthogonal matrices)."

"Two problems solved by principle orthogonal decomposition. First, distinguishing elements from a collection by making d measurements. Second, inverting a complicated map from a p-parameter configuration space to d-dimensional measurement space. "

"The problem is efficently distinguishing elements from a collection by making d measurements."

"Similarity of problems:

- Problem 1: Distinguising elements: The goal is find a discrete object given description in R^d.

- Problem 2: Inverting a complicated map from R^p to R^d. The goal is to find the parameters of R^p from the descripttion in R^p. "

"Combinations of measurements which root out the underlying parameters are called principle (orthogonal) components or factors which are computed by the SVD. The SVD in conventional form is an $O(d^3)$"

" The principle orthogonal or Karhunen-Loève coordinates for an ensemble $X = \{X_n \in R^d : n = 1, \ldots, N\}$ correspond to the choice of axes in R^d which minimizes the volume of the variance ellipsoid."

"the Karhunen-Loève basis eigenvectors are also called principle orthogonal components or principal factors, and computing them for a given ensemble X is also called factor analysis."

Question is what is the relevance of autocovariance in this analysis.

"The adjoint of K is the matrix that changes the standard coordinates into the Karhunen-Loève coordinates. This mapping is the Karhunen-Loève Transform."

"Finding these eigenvectors requires diagonalizing a matrix of order d, which has complexity $O(d^3)$."

"Suppose that the Karhunen-Loève eigenvectors has an optimization over the set of orthogonal transformations of the original ensemble X."

"It is possible to build a library of more than 2^d fast transforms U of R^d to use for "x" points.'

Question: What does this mean? This means that wavelets can be used to construct the fast transforms U which make up the "x"'. From these basis, a particular wavelet basis can be selected as the best choice. Or so this idea is implying.

A few defined terms :

- U is an orthogonal $d \times d$ matrix
- $Y = UX$ is short hand for $Y_n = UX_n \forall n \in Z \cup [1, N]$
- R^d indicates a set of $d \times d$ matrices

"This paper uses a notion of closeness that is derived from the function minimized by the Karhunen-Loeve transform."

Question: What does this mean? How does one identify an algorithm that is the closest to the Karhunen-Loève bases.

"Transform coding gain for an orthogonal matrix is defined by:

- $H(X) = \frac{1}{d} \sum_{i=1}^{d} \log \sigma(X)(\imath)$
- $G_{TC}(U) = \frac{Var(UX)}{e^{H(UX)}}$"

"H has various interpretations. It is the entropy of the direct sum of d independent Gaussian random variables with variances $\sigma(X)(\imath), i = 1, \ldots, d$"

1.1 Approximate Factor Analysis

"The approximate factor analysis algorithm is the search through a library of orthonormal basis for the one whose H is closest to that of the Karhunen-Loève basis, and cases of fast search methods the result is a fast approximate Karhunen-Loève algorithm. "

"The metric 'closeness' of a basis U to the Karhunen-Loève basis K can be measured by computing the transform coding gain of U and subtracting that of K:

$dist_X(U,V) = |H(U,X) - H(V,X)|$"

"This metric is considered degenerate metric on the orthogonal group since it gives a distance of zero between bases which have the same transform coding gain for X"

"A minimum for H(VX) is the Karhunen-Loève basis V=K, so minimizing H(UX) over fast transforms U is equivalent to minimizing $dist_X(U,K)$: it finds the closest fast transform for this ensemble in the transform coding sense. "

Note: This distance measure produces a problem in defining which orthonormal basis to chose. The premise is to find a basis whose transform coding gain metric is closest to the KL transform's. One way to consider this a data structure is that maintains the basis in each resolution. This would be a highly memory intensive problem.

One method is to make a binary tree. However, arrangement of this binary tree is at issue, and is certainly not clear. How are the elements to be arranged in this tree?

Conceivably, the tree could be arranged such that each element in the tree has an array that

is representative of component. Each left child would contain the average component. Each right child would contain a difference component. The root element would contain the original array. This is one possibility. However, it is memory intensive.

1.2 The Entropy Metric

The entropy of U is computed by

- $\mathcal{H}(U) = -\sum_{i=1}^{d}\sum_{j=1}^{d} |U(\imath,\jmath)|^2 \log(|U(\imath,\jmath)|^2)$
- defined for the inequality $0 \leq \mathcal{H}(U) \leq d \log d$

"$\mathcal{H}$ is a functional on O(d) , the compact Lie group of orthogonal linear transformations of R^d."

Question: What is a compact Lie group of orthogonal transformations?

"The entropy metric on the orthogonal group is the function

$dist(U,V) = dist_O(U,V) \stackrel{def}{=} \sqrt{H\ (U \cdot V)}$"

Chapter 2

The approximate KL Transform

Otherwise called the library of rapidly computable orthonormal wavelet packet bases which are geared to address the KL Transform. These are constructed to take advantage of the rapid growth of the number of subtrees of a binary tree.

Question: What is a QF?

"H and G are applied recursively a total of L times to get a complete wavelet packet analysis down to level L. We arrange the resulting sequences into a binary tree, which now contains very many basis subsets."

Given:

- Signal $x = \{x_0, \ldots, x_{d-1}\}$ such that $d = M2^L$
- A complete wavelet packet down to level L
- An arrangement to place the resulting sequences in a binary tree.

Question: This arrangement is very vague. How are the wavelet packets to be arranged?

Supposition: "Suppose that the sequence of coefficients in block f of level s for signal X_n isdesignated as $\lambda_{s,f}^{(n)}(p)$. We then sum the coefficients of the N signal trees into two accumulator trees in location p of block f at level s which are as follows:

- a tree of means which contains $\sum_{n=0}^{N-1} \lambda_{s,f}^{(n)}(p)$
- a tree of squares which contains $\sum_{n=0}^{N-1} (\lambda_{s,f}^{(n)}(p))^2$

Question: From what arrangement, and block of sequences do these means and squares trees come from? It would seem that arrangement would be a crucial issue, and this is treated like a magic hand wave.

"Cost of computing all of the blocks in an L-level tree starting from d samples is $O(dL) = O(d \log d)$ operations for a random vector. If there N vectors, then the total search is $O(Nd \log d)$."

"A binary tree of variances at index p of block f at level s can be produced."

"The tree of variances may now be searched for the graph basis which minimizes the transform coding gain."

Finding the approximate Karhunen-Loève basis

- Expand N vectors $\{X_n \in R^d : n = 1, 2, \ldots, N\}$ into wavelet packets coefficients: $O(Nd \log d)$
- Summing squares into the variance tree: $O(d \log d)$
- Searching the variance tree for a best basis: $O(d + d \log d)$

- Sorting the best basis vectors into decreasing order $O(d \log d)$
- Diagonalizing the auto-covariance matrix of the top d' best basis vectors $O(d'^3)$

Total complexity of the Approximate Karhunen-Loève basis: $O(Nd \log d + d'^3)$. Since $d' \ll d$, it safe to say the approximate solution is faster than the conventional one.

The approximate Karhunen-Loève transform of one vector

- Computing the wavelet packet coefficients of one vector $O(d \log d)$
- Applying the $d' \times d'$ matrix K'^* : $O(d'^2)$

Updating the approximate Karhunen-Loève basis

- Expanding one vector into wavelet packet coefficients
- Adding the coefficients into the means tree
- Adding the squared coefficients into the squares tree
- Forming the variance tree and computing the new information costs
- Searching the variance tree for the joint best basis

The classification in large data sets apply to both rogues' gallery problem, fingerprint classification problem, and rank reduction for complex classifiers.

Chapter 3

Jacobians of complicated maps

Problems with numerical computation of Jacobian are: difference quotiient is ill-conditioned, the Jacobian might itself be an ill-conditioned matrix, and few methods of estimating the condition number of J. If it is replaced by an approximation based on the KL transform for the positive matrix JJ*, what happens? The hypothesis is that the error soley exists in the approximation. The first Justification for this hypothesis is that the KL transform is orthogonal and perfectly condition in its very nature. The SVD of J*J provides an estimation of the condition number of J. $cond(J) = \sqrt{cond(j * J)} = \sqrt{\mu_1/\mu_p}$ such that μ_1 is the first singular value and μ_p is the nth singular value of the estimate J*J.

Note May 21, 2003: Discussion with Dr. Eric Sinzinger: The issue with KL transform is to establish a set of common features, and extractions from a set of basis functions. Of course, the KL is using this against a system matrix for which is already a basis for.

The idea of KL via wavelets is to use a series wavelet transforms against the basis functions

to represent the original with fewer elements. In this basis set, the basis vectors are a collective set. Therefore, the selection of the wavelet pattern has be a close approximation to the original to be valid. Therefore a comparison is made based on a collective variance (recall that variance is a statistical property and related to standard deviation. $\sum_i (x_i - \bar{x})^2$) The wavelet packet sequence whose pattern which achieves the best sparse-and-closeness combination is then used for the calculation of the KL transform.

The above method is also valid against the SVD. Thus it may be very valuable.

Dr. Sinzinger iterated the point, that focus on my first two objectives (Matrix Multiplication and Matrix Inversion) and possibly PDE's should be my primary focus. For academic purposes I agree with him. However, there must be balance between academics and the job that is paying me.

Dr. Sill points out to me that funding is being provided for me via SDSS and Microsoft.

In the mean time, the objective is short description on wavelet based matrix multiplication, and general look at the problem.

Chapter 4

Wavelet Based Matrix Multiplication

The idea is to make a dense collection of wavelet packets $\mathcal{W}(R)$ such that it is dense in $L^2(R)$. Naturarlly this occurs with the Haar Wavelet. Could also be ported to other wavelet schemes such as Daub4 or Coeffliets?

There is an ordering of wavelet packets that is proposed by FAFA. A lexicography is defined from left to right for packet frequency, scale, and position. There is also an adjoint order that exchanges order. "The adjoint order $< *$ just exchanges X and Y indicies $\psi_X \otimes \psi_Y < *\psi'_X \otimes \psi'_Y$ if and only if $\psi_Y \otimes \psi_X < *\psi'_Y \otimes \psi'_X$. It is a total order."

What about projections?

Claim: "There is also a natural injection $J^1 : L^2(R) \to W^1$ given by $J^1 x = \{\lambda_{sf}(p)\}$ for $x \in L^2(R)$ with $\lambda_{sf} =< x, \psi^<_{sfp} >$ being the sequence of backwards inner products with functions in $W(R)$. If B is a basis subset, then the composition J^1_B of J^1 with projection onto the subsequences indexed by B is also injective. J^1_B is an isomorphism of $L^2(R)$ onto $l^2(B)$, which is defined to be

the square-summable sequences of W^1 whose indices belong to B."

Claim: There is an inverse matrix mapping $R^1 : W^1 \to L^2(R)$ defined by: $R^1\lambda(t) = \sum_{(s,f,p)\in Z^3} \bar{\lambda}_{sf}(p)\psi'^{<}_{sfp}(t)$.

What about applying operators to vectors? Notation used was letting X and Y be two named copies of R. Here $x \in L^2(R)$ is a vector whose coordinates with respect to wavelets and wavelet packets form the sequence: $J^1x = \{\langle x, \psi^{<}_X\rangle : \psi_x \in \mathcal{W}(X)\}$. Also included in the notation is a Hilbert-Schmidt operator ($M : L^2(X) \to L^2(Y)$.)

Question: What does this Hilbert-Schmidt operator have to do tensor products? The claim is that with respect to the complete set of tensor products wavelet packets from the sequence $J^2M = \{\langle M, \psi^{<}_X \otimes \psi^{<}_Y\rangle : \psi_X \in \mathcal{W}(X), \psi_Y \in \mathcal{W}(Y)\}$. Supposedly this tensor product leads to the identity:

$$\langle Mv, \psi^{<}_Y\rangle = \sum_{\psi_x\in\mathcal{W}(X)} \langle M, \psi^{<}_X \otimes \psi^{<}_Y\rangle\langle x, \psi^{<}_X\rangle.$$

Question: What is meant by the statement, "This identity generalizes to a linear action of $\mathcal{W}^2$ on $\mathcal{W}^1$ defined by $c(x)_{sfp} = \sum_{(s'f'p')} \lambda_{sfps'fp'}v_{s'f'p'}$." with all of its confusing subscripts. I am presuming that a linear action includes a matrix multiply, since a matrix is of R^{N^2}

Question: What is meant by "We can lift the action of M on x to these larger spaces via the commutative diagram?" His diagram included :

$$\begin{array}{ccc} \mathcal{W}^1 & \overset{J^2_B M}{\to} & \mathcal{W}^1 \\ J^1 \uparrow & \bigcirc & \downarrow R^1 \\ L^2(R) & \overset{M}{\to} & L^2(R) \end{array}$$

FAFA Example of Matrix Multiplication The illustration considers nonstandard multiplication

by square matrices of order 16. In particular, wavelets were used to obtain a “best isotropic basis’.’

What does this mean?

Chapter 5

Matrix Multiplication Second Look

Chosen Wavelet for example is the Haar Wavelet. Wavelet packets are chosen for the multi-resolution method. The key to this method is the best basis (by variance).

There is a ordering given to wavelet packets, in lexical order. The lexicographical order is read from left to right, based on scale, frequency, and position. This basis displayed in a tree.

- Each level is 1 scale
- Each leaf is a low pass and high pass filter from its respective parent.
- The prime element for each leaf is an array which is a condensed version of the original.

This can be applied in the 2-D world with tensor products which are designated mathematically as:

$$\psi_X = \otimes\psi_Y$$

The lexicographical rules apply this inequality:

$$\psi_X = \otimes\psi_Y < \psi'_X = \otimes\psi'_Y$$

- $x < x'$
- $x = x'$ and $y < y'$

If either of these are true then the inequality is true, too.

As far as practical wavelet components are concerned, this is represented in a quad tree. The leaves lexicography is such that the leaves have two possible orders:

1. A-H-V-D applies if the average is in top left corner
2. H-A-V-D applies if the average is in the lower left corner of the transform matrix
3. D-V-H-A applies if the average is in the lower right corner of the transform matrix

The claimed isomorphism

- Natural injection in vector/ 1-D space
- $J : L^2(R) \rightarrow W^1$
- $J^1 x = \{\lambda_{sf}(p)\}$ such that λ_{sf} is a wavelet packet
- Indexing of position can be based on the basis set.
- There is also an inverse map, too (inverse transform.)

- Conclusion: These are mathematical representations

Natural Injection in matrix/ 2-D array for objects in $L^2(R^2)$

- Hilbert Schmidt operators $M \rightarrow < M_j, \psi_X \otimes \psi_j >$

- Increase operators exist $R^2 : W^2 \rightarrow L^2(R^2)$

Identity generalization of $W^2 \rightarrow W^1 < Mv, \psi_X^< >= \sum_\mu < M, \psi_X^< \otimes \psi_Y^< >< x, \psi_X^2 >$

Some how there is a lifting act such that the choice of R (filter ψ index scheme) "can reduce the complexity of map $J_B^2 M$ and therefore the complexity of the operator applications."

linear Action $W^2 \rightarrow W^1$ defined:

$$c(x)_{sfp} = \sum_{(s'f'p')} \lambda_{sfps'f'p'} v_{s'f'p'}$$

1. Inject domain space into a ?16 log 16? dimensional space by expansion into the complete domain wavelet packet analysis tree (bottom)

2. Multiply the injected elements elements with the best basis coefficients in the center matrix square.

3. Sum the products into range wavelet packet synthesis tree.

4. Project the range wavelet pocket synthesis into 16 dimensional range space by the adjoint of wavelet packet expansion.

5.1 Conventional Matrix Multiplication

Conventional multiplication is spelled out as

$$c_{i,j} = \sum_k a_{i,k} b_{k,j}$$

For a 2×2 matrix, there is the following:

$$\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{pmatrix}$$

For a wavelet transform on both matrix A and B, the results are:

$$W(A) = \frac{1}{2} \begin{pmatrix} a_{1,1} + a_{2,1} + a_{1,2} + a_{2,2} & a_{1,1} + a_{2,1} - a_{1,2} - a_{2,2} \\ a_{1,1} - a_{2,1} + a_{1,2} - a_{2,2} & a_{1,1} - a_{2,1} - a_{1,2} + a_{2,2} \end{pmatrix}$$

$$W(B) = \frac{1}{2} \begin{pmatrix} b_{1,1} + b_{2,1} + b_{1,2} + b_{2,2} & b_{1,1} + b_{2,1} - b_{1,2} - b_{2,2} \\ b_{1,1} - b_{2,1} + b_{1,2} - b_{2,2} & b_{1,1} - b_{2,1} - b_{1,2} + b_{2,2} \end{pmatrix}$$

To state $W(A)$ and $W(B)$ simply:

$$W(A) = \begin{pmatrix} w^a_{1,1} & w^a_{1,2} \\ w^a_{2,1} & w^a_{2,2} \end{pmatrix}$$

$$W(B) = \begin{pmatrix} w^b_{1,1} & w^b_{1,2} \\ w^b_{2,1} & w^b_{2,2} \end{pmatrix}$$

The product of the averages:

$w^a_{1,1}w^b_{1,1} = \frac{1}{4}(b_{1,1}a_{1,1}+b_{1,1}a_{2,1}+b_{1,1}a_{1,2}+b_{1,1}a_{2,2}+b_{2,1}a_{1,1}+b_{2,1}a_{2,1}+b_{2,1}a_{1,2}+b_{2,1}a_{2,2}+b_{1,2}a_{1,1}+b_{1,2}a_{2,1}+b_{1,2}a_{1,2}+b_{1,2}a_{2,2}+b_{2,2}a_{1,1}+b_{2,2}a_{2,1}+b_{2,2}a_{1,2}+b_{2,2}a_{2,2})$

The product of vertical

$w^a_{1,1}w^b_{1,1} = \frac{1}{4}(b_{1,1}a_{1,1}+b_{1,1}a_{2,1}-b_{1,1}a_{1,2}-b_{1,1}a_{2,2}+b_{2,1}a_{1,1}+b_{2,1}a_{2,1}-b_{2,1}a_{1,2}-b_{2,1}a_{2,2}-b_{1,2}a_{1,1}-b_{1,2}a_{2,1}+b_{1,2}a_{1,2}+b_{1,2}a_{2,2}-b_{2,2}a_{1,1}-b_{2,2}a_{2,1}+b_{2,2}a_{1,2}+b_{2,2}a_{2,2})$

The product of the horizontal

$w^a_{1,1}w^b_{1,1} = \frac{1}{4}(b_{1,1}a_{1,1}-b_{1,1}a_{2,1}+b_{1,1}a_{1,2}-b_{1,1}a_{2,2}-b_{2,1}a_{1,1}+b_{2,1}a_{2,1}-b_{2,1}a_{1,2}+b_{2,1}a_{2,2}+b_{1,2}a_{1,1}-b_{1,2}a_{2,1}+b_{1,2}a_{1,2}-b_{1,2}a_{2,2}-b_{2,2}a_{1,1}+b_{2,2}a_{2,1}-b_{2,2}a_{1,2}+b_{2,2}a_{2,2})$

The product of the diagonal

$w^a_{1,1}w^b_{1,1} = \frac{1}{4}(b_{1,1}a_{1,1}-b_{1,1}a_{2,1}-b_{1,1}a_{1,2}+b_{1,1}a_{2,2}-b_{2,1}a_{1,1}+b_{2,1}a_{2,1}+b_{2,1}a_{1,2}-b_{2,1}a_{2,2}-b_{1,2}a_{1,1}+b_{1,2}a_{2,1}+b_{1,2}a_{1,2}-b_{1,2}a_{2,2}+b_{2,2}a_{1,1}-b_{2,2}a_{2,1}-b_{2,2}a_{1,2}+b_{2,2}a_{2,2})$

The section multiplication is defined:

$$W(A) \overset{+}{*} W(B) = \begin{pmatrix} w^a_{1,1}w^b_{1,1} & w^a_{1,2}w^b_{1,2} \\ w^a_{2,1}w^b_{2,1} & w^a_{2,2}w^b_{2,2} \end{pmatrix}$$

The inverse wavelet transform of the section multiply is:

$$W^{-1}(W(A) \overset{+}{*} W(B)) = \frac{1}{2}\begin{pmatrix} (w^a_{1,1}w^b_{1,1} + w^a_{1,2}w^b_{1,2} + w^a_{2,1}w^b_{2,1} + w^a_{2,2}w^b_{2,2}) & (w^a_{1,1}w^b_{1,1} - w^a_{1,2}w^b_{1,2} + w^a_{2,1}w^b_{2,1} - w^a_{2,2}w^b_{2,2}) \\ (w^a_{1,1}w^b_{1,1} + w^a_{1,2}w^b_{1,2} - w^a_{2,1}w^b_{2,1} - w^a_{2,2}w^b_{2,2}) & (w^a_{1,1}w^b_{1,1} - w^a_{1,2}w^b_{1,2} - w^a_{2,1}w^b_{2,1} + w^a_{2,2}w^b_{2,2}) \end{pmatrix}$$

$$\frac{1}{2}\begin{pmatrix} (w^a_{1,1}w^b_{1,1} + w^a_{1,2}w^b_{1,2} + w^a_{2,1}w^b_{2,1} + w^a_{2,2}w^b_{2,2}) & (w^a_{1,1}w^b_{1,1} - w^a_{1,2}w^b_{1,2} + w^a_{2,1}w^b_{2,1} - w^a_{2,2}w^b_{2,2}) \\ (w^a_{1,1}w^b_{1,1} + w^a_{1,2}w^b_{1,2} - w^a_{2,1}w^b_{2,1} - w^a_{2,2}w^b_{2,2}) & (w^a_{1,1}w^b_{1,1} - w^a_{1,2}w^b_{1,2} - w^a_{2,1}w^b_{2,1} + w^a_{2,2}w^b_{2,2}) \end{pmatrix}$$