Compression

Mladen Victor WICKERHAUSER

Washington University in St. Louis, Missouri
victor@wustl.edu
http://www.math.wustl.edu/~victor

Dimensionality Reduction and Manifold Estimation
PMF — University of Zagreb
Winter, 2022

Matrix Operator Norm

Setup: $m \times n$ matrix A, $\mathbf{C}$-valued coefficients a_{ij}.

Goal: for $A : \mathbf{E}^n \to \mathbf{E}^m$, estimate

$$\|A\|_{\mathrm{op}} \stackrel{\mathrm{def}}{=} \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \left(\sup_{x \neq 0} \frac{\|Ax\|^2}{\|x\|^2} \right)^{1/2} .$$

But $\|Ax\|^2 = \langle Ax, Ax \rangle = \langle x, A^*Ax \rangle$, for adjoint $A^* \stackrel{\mathrm{def}}{=} \bar{A}^T$.

A^*A is $n \times n$ and

- (hermitean) symmetric: $(A^*A)^* = A^*(A^*)^* = A^*A$.
- positive semidefinite: $\langle x, A^*Ax \rangle = \langle Ax, Ax \rangle = \|Ax\|^2 \geq 0$,

so by the spectral theorem for hermitean matrices, its eigenvalues are purely real and nonnegative.

Singular Values

Suppose $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of A^*A.
(It is traditional to list them in decreasing order, with multiplicity, possibly with multiple 0s at the lower end.)

Then $\sigma_k \stackrel{\text{def}}{=} \sqrt{\lambda_k}$, $k = 1, \ldots, n$ are called the *singular values* of A.

Lemma

- $\|A\|_{\text{op}} = \sigma_1 = \max\{\sigma_k : k = 1, \ldots, n\}$.
- *Rank of A is the number of nonzero singular values, counting multiplicity.*
- *Nonzero singular values of A are the same as for A^*, counting multiplicity.*

Proof.

Apply the spectral theorem, use the n linearly independent eigenvectors of A^*A. □

Singular Value Decomposition

Theorem

Matrix $A \in \mathbf{C}^{m \times n}$ can be factored as $A = USV^*$, where

- *U is $m \times m$ unitary (so $U^*U = \mathrm{I}$),*
- *V is $n \times n$ unitary (so $V^*V = \mathrm{I}$),*
- *S is $m \times n$ diagonal with $S_{kk} = \sigma_k$, $k = 1, \ldots, \min(m, n)$ being singular values σ_k of A, with multiplicity.*

Proof.

Diagonalize $A^*A = VS^*SV^*$ by spectral theorem to find unitary V. Likewise, diagonalize hermitean symmetric $AA^* = USS^*U^*$ with unitary U. □

SVD Properties

Let the rank of A be denoted by $r \leq \min(m, n)$. The matrices U, S, V may be chosen so that

- $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots$ gives the diagonal of S, with $S_{ij} = 0$ for all other $1 \leq i \leq m$ and $1 \leq j \leq n$.
- Columns $U_1, \ldots, U_r$ of U are an o.n.b. for column space $C(A)$.
- Columns $U_{r+1}, \ldots, U_m$ of U are an o.n.b. for cokernel $N(A^*)$.
- Rows $V_1^*, \ldots, V_r^*$ of V^* are an o.n.b. for the row space $R(A)$.
- Rows $V_{r+1}^*, \ldots, V_n^*$ of V^* are an o.n.b. for the kernel $N(A)$.

Put $\tilde{U} = [U_1, \ldots, U_r]$, $\tilde{V} = [V_1, \ldots, V_r]$, and $\tilde{S} = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_r)$ to get the *partial SVD*:

$$A = \tilde{U}\tilde{S}\tilde{V}^*,$$

which omits the singular values $\sigma_k = 0$ that contribute nothing.

SVD Applications 1

Let $A = USV^*$ be factored with the properties above, $\{U_k\}$ the columns of U, and $\{V_k\}$ the columns of V. Then,

- Put $P = [U_1, \ldots, U_r]$. Then $PP^* : \mathbf{C}^m \to C(A)$ is the orthogonal projection, while $P^*P = \mathrm{I} : \mathbf{C}^r \to \mathbf{C}^r$.
- Put $Q = [V_{r+1}, \ldots, V_n]$. Then $QQ^* : \mathbf{C}^n \to N(A)$ is the orthogonal projection, while $Q^*Q = \mathrm{I} : \mathbf{C}^{n-r} \to \mathbf{C}^{n-r}$.
- The *pseudoinverse* of $A = USV^*$ is $A^g \stackrel{\mathrm{def}}{=} VS^gU^*$, where

$$S^g_{ij} \stackrel{\mathrm{def}}{=} \begin{cases} 1/S_{ij}, & S_{ij} \neq 0, \\ 0, & S_{ij} = 0. \end{cases}$$

  It satisfies all the Penrose conditions: (1) $AA^gA = A$, (2) $A^gAA^g = A^g$, (3) $(AA^g)^* = AA^g$, and (4) $(A^gA)^* = A^gA$.

SVD Applications 2

If A is real, square, symmetric, and positive definite, then $U = V$ is real-valued orthogonal and $S = \Lambda$ is the diagonalised matrix of eigenvalues of A: $A = U\Lambda U^T$.

Example: A is the *covariance matrix* of a d-variate normal random variable $X \in \mathbf{R}^d$, so

$$A = \mathrm{cov}\,(X) = \mathrm{E}([X - \bar{X}][X - \bar{X}]^T) \in \mathbf{R}^{d \times d},$$

if X is a column vector.

The *screeplot* of the singular values $\{(k, \sigma_k) : k = 1, 2, \ldots, n\}$ (which are the eigenvalues of A) depicts the accumulation of variance by number of variates.

Karhunen-Loève

Setup: d-variate normal random vector $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ with mean $\mu = \mathrm{E}(\mathbf{x}) \in \mathbf{E}^d$ and covariance matrix $\Sigma \in \mathbf{R}^{d \times d}$:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}; \qquad \begin{aligned} \mu &= \mathrm{E}(\mathbf{x}); \\ \Sigma &= \mathrm{E}\left((\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\right). \end{aligned}$$

(Here $\mathbf{v}\mathbf{v}^T(i,j) = v_i v_j$, $1 \le i, j \le d$, with column vectors $\mathbf{v}$.)

Σ is symmetric positive (semi-)definite. By the spectral theorem,

$$\Sigma = UDU^T, \qquad U, D \in \mathbf{R}^{d \times d},$$

with orthogonal U and diagonal $D \ge 0$. Put $\sigma_i^2 \stackrel{\mathrm{def}}{=} D(i,i)$ to get:

Theorem

Coordinates of $U^T(\mathbf{x} - \mu)$ are $\mathcal{N}(0, \sigma_i^2)$, $i = 1, \ldots, d$, and independent.

Karhunen-Loève Basis

Columns $\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$ of unitary U are an orthonormal basis for $\mathbf{E}^d$, called the *Karhunen-Loève basis* for the random vector $\mathbf{x}$.

WOLOG, choose indices such that $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2 \geq 0$.

The $\mathbf{u}_1$ component has the greatest variance (σ_1^2) and contains the most information about position in S.

If $\sigma_k^2 = 0$ for $k > p$, then S is contained in a p-dimensional submanifold of $\mathbf{E}^d$.

Generalize to small variances: If $\sigma_k^2 < \sigma^2$ for $k > p$ and some threshold variance σ^2, say that S is *essentially p-dimensional*.

Hilbert Matrix

Generate examples of essentially p-dimensional data in $\mathbf{E}^d$ with $p \ll d$ using the $d \times d$ Hilbert matrix:

$$\mathrm{hilb}(d) \stackrel{\mathrm{def}}{=} \left(\frac{1}{i+j-1}\right)_{1 \le i,j \le d} = \begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{d} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{d} & \frac{1}{d+1} & \cdots & \frac{1}{2d-1} \end{pmatrix}$$

Symmetric and positive definite $\implies$ its singular values are its eigenvalues, with the estimate

$$\sigma_p = \lambda_p = O\left(2^{-5p}\right), \qquad p \to \infty.$$

K-L Transform Compression

Setup:

- samples $\{\mathbf{x}_n\} \in \mathbf{E}^d$ from r.v. $\mathbf{x}$ ("signals")
- K-L basis U for $\mathbf{x}$
- quality factor σ^2

Method:

- transform $\mathbf{y} = U^T \mathbf{x}$
- quantize $\mathbf{q} = \lfloor \mathbf{y}/\sigma \rfloor \subset \mathbf{Z}$
- remove redundancy in $\mathbf{q}$

JPEG

Joint Photographic Experts Group image compression algorithm:

- $\mathbf{x} \in \mathbf{E}^{8 \times 8}$ is an 8×8 subimage
- $\mathbf{x}(i, j) \in \mathbf{R}$ is the pixel at i, j
- model the covariance by

$$\mathrm{E}(\mathbf{x}(i,j)\mathbf{x}(i',j')) = f((i'-i)^2 + (j'-j)^2)$$

for some decreasing function f like $f(r) = e^{-r}$.

For all such models, Σ commutes with translation (in i, j), hence is a convolution, hence is diagonalized by the discrete Fourier or cosine transform. So, use DCT as the K-L transform.

Details:
`https://www.iso.org/standard/18902.html`
`https://jpeg.org/jpeg/index.html`

Empirical K-L from Samples

Idea: estimate Σ from samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbf{E}^d$ without knowing the distribution of $\mathbf{x}$.

- Sample mean

$$\bar{\mathbf{x}} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- Sample covariance

$$\bar{\Sigma} \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- Approximate K-L basis: diagonalize $\bar{\Sigma}$ with orthogonal U

$$\bar{\Sigma} = UDU^T, \quad D = \mathrm{diag}(s_1^2, \ldots, s_d^2) \in \mathbf{R}^{d \times d}.$$

Face Coding by Empirical K-L

Figure: Face, minus average face, contains the important information.

Remark. Data is from Lawrence Sirovich, 143 images, 128×128 pixels each (so $d \approx 16,000$) in 8-bit grayscale, of male, Brown University students without facial hair, shifted and scaled to fixed eye points.

Tangent Space Estimation by Empirical K-L

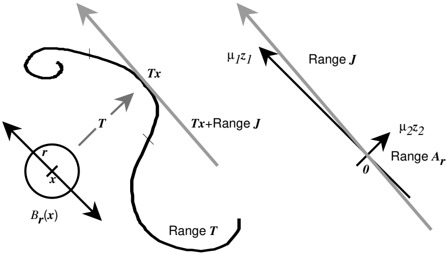


Figure: Principal components of a scatterplot are good candidates for an o.n.b. for the tangent space.

Complexity of Empirical K-L

For data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbf{E}^d$:

- Sample mean $\bar{\mathbf{x}}$ costs $O(nd)$
- Mean-subtracted samples $\{\mathbf{x}'_i\}$ cost $O(nd)$
- Sample covariance $\bar{\Sigma}$ costs $O(nd^2)$
- Diagonalizing $\bar{\Sigma}$ with orthogonal U costs $O(d^3)$.
- Transform one vector $\mathbf{x} \mapsto U^T\mathbf{x}$ costs $O(d^2)$.
- Transform n data vectors costs $O(nd^2)$.

Remark. U is sensitive to the data and may not be nice.

Empirical K-L variables are uncorrelated (over the samples), not independent.

Orthogonal Transform Coding

Given data $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbf{E}^d$ and an orthogonal transform $U : \mathbf{E}^d \to \mathbf{E}^d$, the empirical covariance of transformed data $U^T\mathbf{x}$ is:

$$\bar{\Sigma}_U \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (U^T[\mathbf{x}_i - \bar{\mathbf{x}}])(U^T[\mathbf{x}_i - \bar{\mathbf{x}}])^T = U^T \bar{\Sigma} U,$$

where $\bar{\Sigma}$ is the empirical covariance of untransformed $\mathbf{x}$.

Trace is invariant under this similarity transform:

$$\sum_{i=1}^{n} \bar{\Sigma}_U(i,i) = \operatorname{tr} \bar{\Sigma}_U = \operatorname{tr} U^T \bar{\Sigma} U = \operatorname{tr} \bar{\Sigma} = \sum_{i=1}^{n} \bar{\Sigma}(i,i).$$

Exercise: $\operatorname{tr} \bar{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$, the *mean squared deviation*.

Best Orthogonal Basis

$\bar{\Sigma}$ and $\bar{\Sigma}_U$ are symmetric positive definite matrices.

Idea: find good U by comparing diagonals of $\bar{\Sigma}_U$ and $\bar{\Sigma}$.

Start by normalizing the covariances:

$$p_i(U) \stackrel{\text{def}}{=} \frac{1}{\operatorname{tr} \bar{\Sigma}_U} \bar{\Sigma}_U(i,i), \qquad i = 1, \ldots, d.$$

Observe that $p(U) \stackrel{\text{def}}{=} \{p_i(U) : i = 1, \ldots, d\}$ is a pdf for every U:

$$(\forall i)\, p_i(U) \geq 0; \qquad \sum_{i=1}^{d} p_i(U) = 1.$$

Given: fixed library of orthogonal transforms $\mathbf{U} = \{U_1, \ldots, U_N\}$.

Define the *best (orthogonal) basis* from $\mathbf{U}$ to be the optimum of some concentration function on the pdf $p(U)$.

Transform Coding Gain

Transform coding gain from U measures concentration on the diagonal:

$$G(U) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \log \frac{1}{p_i(U)} = \log \prod_{i=1}^{d} \frac{1}{p_i(U)},$$

with $p(U)$ defined as before. This requires $p_i(U) > 0$, all i.

Note: If $\bar{\Sigma}_U$ is diagonal, then

$$G(U) = \log \frac{(\operatorname{tr} \bar{\Sigma})^d}{\det \bar{\Sigma}}$$

Exercise: Transform coding gain $G(U)$ is maximized by the empirical K-L basis U for $\mathbf{x}$. (Hint: use Cholesky factorization.)

Entropy

An alternative measure of (inverse) concentration is the Shannon-Weaver *entropy* of the normalized diagonal elements:

$$H(U) \stackrel{\text{def}}{=} \sum_{i=1}^{d} p_i(u) \log \frac{1}{p_i(U)} = \log \prod_{i=1}^{d} \left(\frac{1}{p_i(U)} \right)^{p_i(U)},$$

with $p(U)$ as before and the convention that $0 \log(1/0) = 0$.

The *theoretical dimension*, an estimate of the number of coordinates that contain most of the variance of $\mathbf{x}$ in the U basis, is $\exp H(U)$.

Lemma

Entropy $H(U)$ is minimized by the empirical K-L basis U for $\mathbf{x}$.

Improvements from Constraints

Requiring that the underlying orthogonal functions be smooth has advantagse.

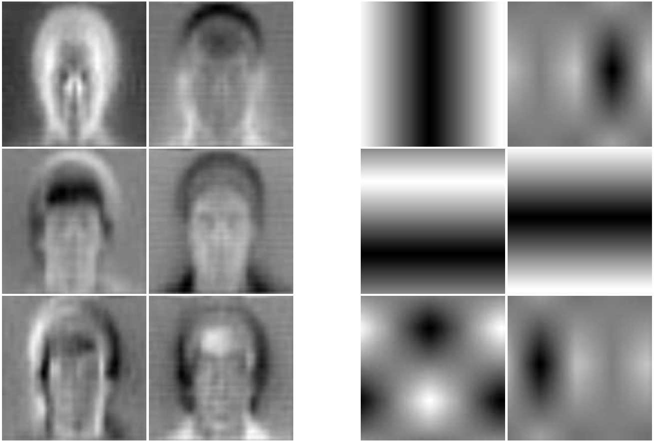


Figure: Eigenfaces versus principal components with imposed smoothness

Haar Basis

Problem: find a basis for $L^2(\mathbf{R})$ that is countable, orthonormal, and simple (piecewise constant and compactly supported).

Theorem (Haar, 1910)

Let

$$w(t) = \begin{cases} 1, & 0 \le t < \frac{1}{2} \\ -1, & \frac{1}{2} \le t < 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Then $W \stackrel{\text{def}}{=} \{2^{-j/2} w(2^{-j}t - k) : j, k \in \mathbf{Z}\}$ *is an o.n.b. for* $L^2(\mathbf{R})$.

Proof.

Decompose $L^2(\mathbf{R}) = \oplus \sum_{j \in \mathbf{Z}} W_j$ "by scales," where

$$W_j \stackrel{\text{def}}{=} \overline{\text{span}} \{2^{-j/2} w(2^{-j}t - k) : k \in \mathbf{Z}\}.$$

Show orthogonality directly, and density by construction. □

Fast Haar Transform 1

Idea: find W_j coefficients by successive approximation. Let

$$\mathbf{1}(t) \stackrel{\text{def}}{=} \begin{cases} 1, & 0 \le t < 1 \\ 0, & \text{otherwise.} \end{cases}, \quad B_0 \stackrel{\text{def}}{=} \{\mathbf{1}(t-k) : k \in \mathbf{Z}\},$$

denote the closed span by $V_0 \stackrel{\text{def}}{=} \overline{\text{span}}\, B_0 \subset L^2(\mathbf{R})$, and put:

$$P_0 : L^2(\mathbf{R}) \to V_0, \qquad P_0 f(t) \stackrel{\text{def}}{=} \sum_{b \in B_0} \langle f, b \rangle b(t).$$

(B_0 is an orthonormal basis for V_0, the piecewise constant square integrable functions with jumps at integers.)

Exercise: P_0 is an orthogonal projection.

Fast Haar Transform 2

Note that $\langle w, \mathbf{1} \rangle = 0$ and $w(t) = \mathbf{1}(2t) - \mathbf{1}(2t-1)$. This suggests:

$$V_j \stackrel{\text{def}}{=} \overline{\text{span}}\, B_j, \quad \text{with o.n.b. } B_j \stackrel{\text{def}}{=} \{2^{-j/2}\mathbf{1}(2^{-j}t - k) : k \in \mathbf{Z}\},$$

just like W_j has o.n.b. $\{2^{-j/2}w(2^{-j}t - k) : k \in \mathbf{Z}\}$. Then

$$w \in V_0^{\perp} \cap V_{-1} \implies W_0 \subset V_0^{\perp} \cap V_{-1} \implies W_0 \subset V_0^{\perp} \cap V_{-1}.$$

Get the Haar expansion from successive approximation:

Lemma

- $\{0\} \subset \cdots \subset V_{+1} \subset V_0 \subset V_{-1} \subset \cdots \subset L^2(\mathbf{R})$,
- *for all* $j \in \mathbf{Z}$, $V_j = V_{j+1} \oplus W_{j+1}$.

Fast Haar Transform 3

Denote by P_j and Q_j the orthogonal projections onto V_j and W_j, respectively

Use the o.n.b.s for V_j and W_j to factor these orthogonal projections into transforms H, G acting on coefficient sequences:

$$P_j \stackrel{\text{def}}{=} HH^*, \qquad Q_j \stackrel{\text{def}}{=} GG^*,$$

where H and G are linear transformations on $\ell^2(\mathbf{Z})$ composed with *decimation by 2*, keeping only half the output.

Remark. For simplicity of analysis, H and G are assumed to be the same for all j. This suffices for the fast Haar decomposition. It may be generalized with a family of operators $\{(H_j, G_j) : j \in \mathbf{Z}\}$ to control size properties, something done by Morten Nielsen.

Fast Haar Transform 4

Generate coefficient sequences with H and G:

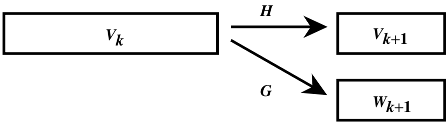


Figure: Low-pass (H) and high-pass (G) filtering

Perfect reconstruction using adjoints H^* and G^*:

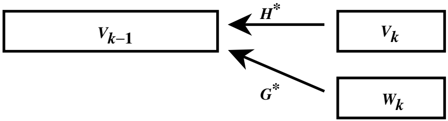


Figure: Reconstruct from low (V) and high (W) frequency components

Fast Haar Transform 5

Apply filtering to coefficient sequences recursively:

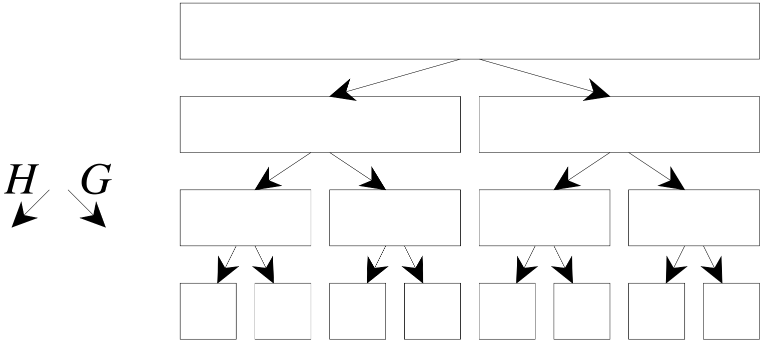


Figure: Recursive splitting algorithm

Low-Pass H and High-Pass G

General Conditions on H, G:

- $HH^* = I$ and $GG^* = I$, so H^*H and G^*G are orthogonal projections;
- $HG^* = GH^* = 0$, so H and G project onto orthogonal subspaces;
- $H^*H + G^*G = I$, so H and G together allow perfect reconstruction.

Additional conditions:

- $\phi(x) = H\phi(x) \stackrel{\text{def}}{=} \sum_k h_k \phi(2x - k)$ has a fixed point in $L^2(\mathbf{R}) \cap L^1(\mathbf{R})$ with $\|\phi\| = 1$.
- ϕ is nice, for example smooth and compactly supported.

Example: Haar-Walsh Splitting

For $x \in \ell^2(\mathbf{Z})$, define

$$\begin{aligned} Hx(n) &= [x(2n) + x(2n+1)]/2; \\ Gx(n) &= x(2n+1) - x(2n). \end{aligned}$$

$$H^*x(n) = \begin{cases} x(\frac{n}{2}), & \text{if } n \text{ is even;} \\ x(\frac{n-1}{2}), & \text{if } n \text{ is odd;} \end{cases}$$

$$G^*x(n) = \begin{cases} -\frac{1}{2}x(\frac{n}{2}), & \text{if } n \text{ is even;} \\ \frac{1}{2}x(\frac{n-1}{2}), & \text{if } n \text{ is odd.} \end{cases}$$

Exercise: $HH^*x = x$, $GG^*x = x$, and $x = H^*Hx + G^*Gx$.

Fast Wavelet and Wavelet Packet Transforms

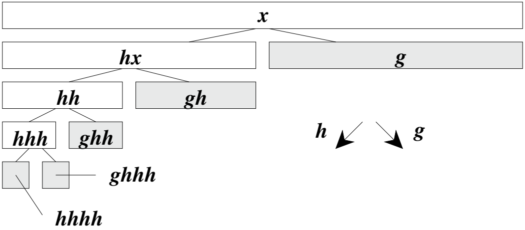


Figure: Fast Haar wavelet tranform (Mallat)

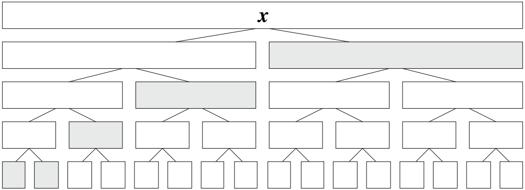


Figure: Complete wavelet packet decomposition

Fast Transforms Into Other Bases

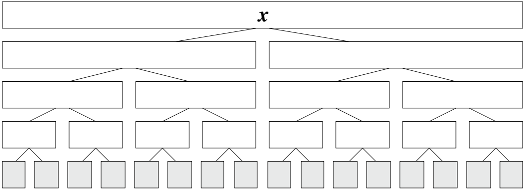


Figure: Fast Fourier-like subband transform

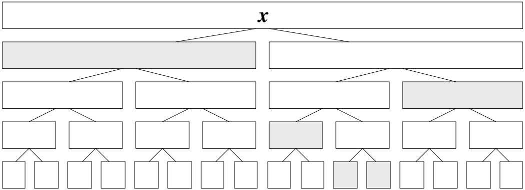


Figure: Some other adapted wavelet packet transform

How Many Such Bases?

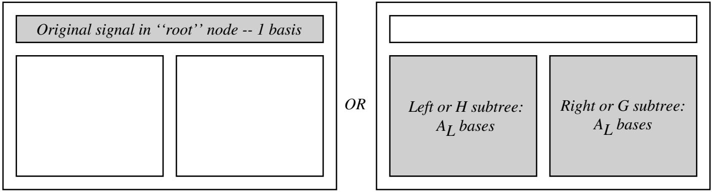


Figure: Counting bases obtainable by Mallat's algorithm

Let A_L be the number of distinct orthonormal bases with L levels of decomposition. Then $A_0 = 1$ and $A_1 = 2$, while

$$A_{L+1} = 1 + A_L^2 \qquad \Longrightarrow A_{L+1} > 2^{2^L}, \quad L > 0.$$

Approximate K-L Basis

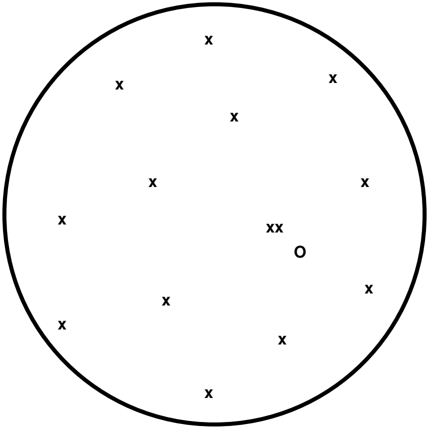


Figure: Best basis (xx) among the fast bases (x) among all the orthonormal bases including empirical K-L basis (o).

Concentration of Variance

Compute the transform coding gain by comparing cumulative variance:

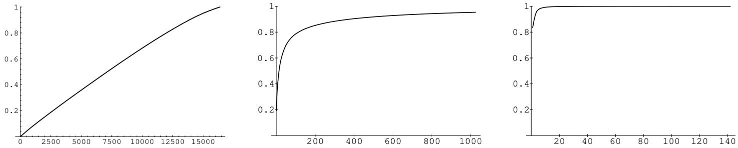


Figure: Example cumulative variance in: original basis (L), best fast basis (M), and empirical K-L basis (R).

Remark. Data is from Lawrence Sirovich, mentioned above.

Haar Wavelet Packets

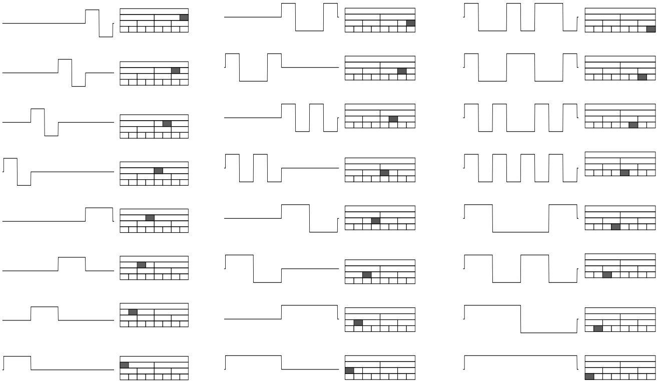


Figure: Haar wavelet packets from three levels

Wavelet and Best-Basis Image Compression

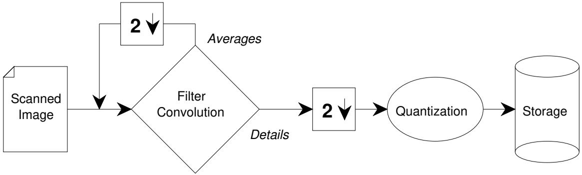


Figure: Wavelet compression using fast Haar tranform

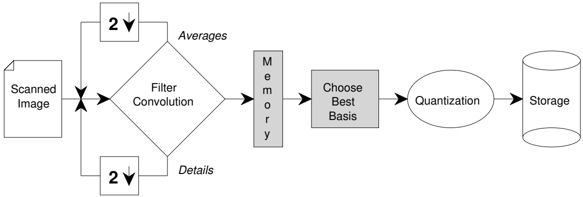


Figure: Best-basis compression using wavelet packets

Phase Plane

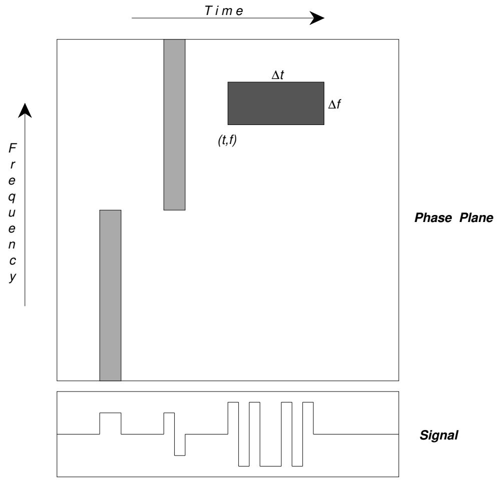


Figure: Idealized depiction of the phase plane with orthogonal atoms

Dirac and Fourier Atoms in the Phase Plane

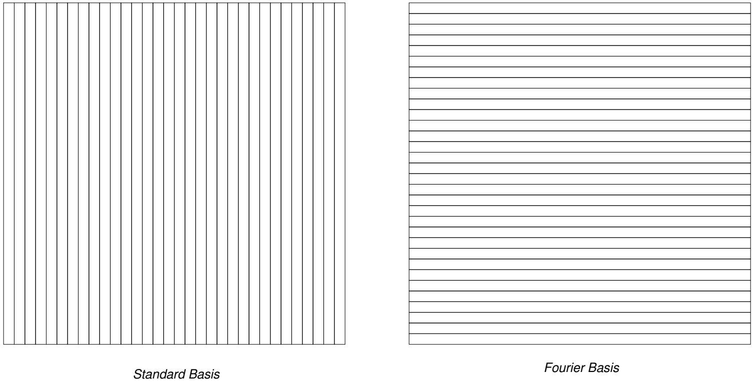


Figure: Samples (Dirac atoms) versus pure frequencies (Fourier atoms)

Gabor Atoms in the Phase Plane

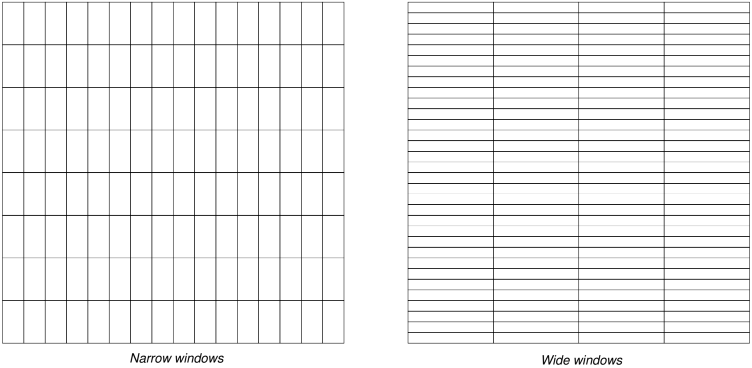
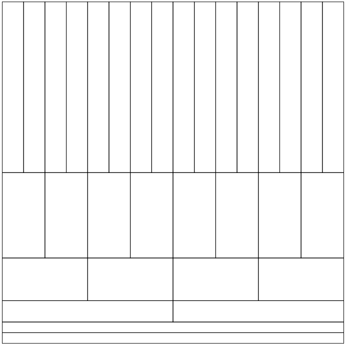


Figure: Windowed or Short-Time Fourier bases (Gabor atoms)

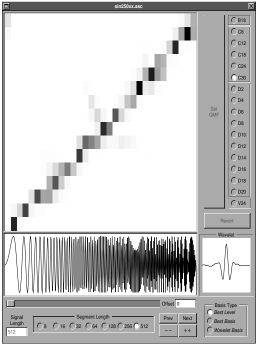
Wavelets Decomposition of the Phase Plane



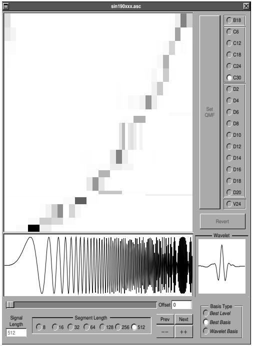
Wavelet basis

Figure: Wavelet orthonormal basis (Littlewood-Paley atoms)

Best Basis Analysis of Chirps in the Phase Plane



Linear chirp, best level



Quadratic chirp, best basis

Figure: Atomic decomposition of linear and quadratic chirps

CUR Approximation

Given: Samples $X_1, \ldots, X_m \in \mathbf{R}^n$ from r mixed n-variate normals.

Goal: Identify the r principal axes of the component distributions from the mixture matrix $A \in \mathbf{R}^{m \times n}$,

$$A = \begin{pmatrix} X_1(1) & \cdots & X_1(n) \\ \vdots & \ddots & \vdots \\ X_m(1) & \cdots & X_m(n) \end{pmatrix},$$

using columns and rows selected from A.

Example: `cur-talk.pdf`, (Mark Embree, p.7)

Idea: Choose r columns $C \in \mathbf{R}^{m \times r}$ and r rows $R \in \mathbf{R}^{r \times n}$ and an $r \times r$ unitary U so as to minimize

$$\|A - CUR\|_{\mathrm{op}}.$$

CUR Algorithm

Method: factor $A = VSW^T$ by SVD. Either fix r or fix a threshold $\sigma > 0$ from which the effective rank r will be computed by $\sigma_k < \sigma$ if $k > r$.

- Choose top r rows of V by norm; use those rows of A in R.
- Choose top r columns of W^T by norm; use those columns of A in C.
- Compute $U = C^g A R^g$, where X^g is the Moore-Penrose pseudoinverse of X.

Claim: $\|A - CUR\|_{\mathrm{op}}$ is minimal over all such choices.

Proof: `CS6220-Lecture14-CUR.pdf` (Anil Damle, p.4)

Remark. Complexity is moderate, requiring three SVDs.

References

- Gilbert Strang. “Chapter 7: The Singular Value Decomposition.” *Introduction to Linear Algebra*, 4th Ed., Wellesley-Cambridge, 2009.
- Pavel Šťovíček. “Spectral Analysis of the Hilbert Matrix.” Lecture, Department of Mathematics, Faculty of Nuclear Science, Czech Technical University in Prague.
- Michael W. Mahoney and Petros Drineas. “CUR matrix decompositions for improved data analysis.” Proceedings of the National Academy of Sciences, 106:3 (2009), pp.697–702.
- Morten Nielsen. *Size Properties of Wavelet Packets*. Ph.D. thesis, Washington University in St. Louis, 2000.
- Mladen Victor Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. AK Peters/CRC Press, Boston, 1994.