Interactions Between Modern Harmonic Analysis and Statistical Estimation

Emmanuel Candès, California Institute of Technology

IDR Meeting, Singapore, August 2004

Outline of Lectures

- Lecture I. Classical estimation
- Lecture II. Nonlinear estimation
- Lecture III. Efficient representations and efficient estimation
- Lecture IV. Estimation in ill-posed linear inverse problems

Thanks...

To Carl de Boor for making corrections and suggesting changes

Lecture 1: Classical estimation

- Estimation of Gaussian processes
- Compression of Gaussian processes
- Smoothing
- Statistical theory

Principal components

- Stochastic process $X = (X_1, \ldots, X_T)$ with covariance matrix Σ

$$\Sigma(s,t) = E(X_s X_t) - E(X_s)E(X_t)$$

- Matrix of principal components

$$\Phi = \text{Col}(\varphi_1, \varphi_2, \ldots, \varphi_n)$$

- Φ diagonalizes the covariance matrix

$$D = \Phi \Sigma \Phi^T, \quad D = \text{diag}(\lambda_k^2).$$

- Principal component analysis

$$X' = \Phi^T X$$

1. The coordinates $X'_1, \ldots, X'_n$ are uncorrelated.
2. If X is multivariate normal, the coordinates $X'_1, \ldots, X'_n$ are independent.

Interpretation

Suppose $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_n^2$

- First principal component φ_1, $\|\varphi_1\| = 1$:

$$\text{Var}(X^T u) \leq \text{Var}(X^T \varphi_1), \forall u, \ \|u\| = 1,$$

  i.e. projection with maximal variance.

- Second principal component φ_2, $\|\varphi_2\| = 1$:

$$\text{Var}(X^T u) \leq \text{Var}(X^T \varphi_2), \forall u, \ \|u\| = 1, u \perp \varphi_1.$$

- Etc.

Principal components

X Gaussian process:

$$X \sim N(\mu, \Sigma), \quad f(x) \propto e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.$$

- Level sets of the density are ellipsoids

$$x \in \mathbf{R}^n, \quad (x-\mu)^T \Sigma^{-1} (x-\mu) = \text{Constant}.$$

- The principal components are the principal axes of this ellipsoid.

Karhunen Loeve Decomposition

Decomposition into principal components

- $(X_t), t = 0, \ldots, T-1$, is a Gaussian Process $X \sim N(0, \Sigma)$.
- Analysis

$$X' = \Phi^T X, \quad \text{Cov}(X') = D = \text{diag}(\lambda_k^2)$$

- Synthesis

$$X = \Phi X', \quad X'_k = \lambda_k Z_k,$$

  with Z white noise, Z_k i.i.d. $N(0,1)$.

- Karhunen-Loeve decomposition

$$X_t = \sum_k \lambda_k \varphi_k(t).$$

Principal Components

X_t, $0 \le t \le T-1$, a zero-mean Gaussian stochastic process with covariance Σ.

Assume $\Sigma < \infty$.

- Covariance Σ
  - Orthonormal eigenfunctions ϕ_k
  - Eigenvalues λ_k^2
- Karhunen-Loève (KL) Expansion

$$X_t = \sum_k Z_k \phi_k(t)$$

- KL Components:

$$\begin{aligned} Z_k &= \langle X, \phi_k \rangle \\ Z_k &\sim N(0, \lambda_k^2) \end{aligned}$$

Representation of Gaussian stochastic processes as superpositions of **independent components**

1. Analysis: Find the orthonormal eigenfunctions ϕ_k and the independent components Z_k.
2. Synthesis: Synthesize the process from the independent components using the orthonormal eigenfunctions.

Example

Stationary process on the circle

- Covariance Σ

$$\Sigma(s,t) = \gamma(s \ominus t), \quad 0 \le s,t < T$$

- Orthonormal eigenfunctions (T odd)

$$\begin{aligned} \phi_0(t) &= 1/\sqrt{T} \\ \phi_{2k}(t) &= \sqrt{2/T}\cos(2\pi kt/T), \quad k = 1, 2, \ldots, (T-1)/2 \\ \phi_{2k+1}(t) &= \sqrt{2/T}\sin(2\pi kt/T), \quad k = 1, 2, \ldots, (T-1)/2 \end{aligned}$$

- Eigenvalues

$$\lambda_k^2 = \phi_k^T \Sigma \phi_k, \quad k = 0, 1, 2, \ldots, T-1.$$

Estimation for Gaussian Processes

$$Y_t = X_t + \sigma Z_t, \quad 0 \le t < T$$

- Y observed
- X as before
- Z white noise, $Z \sim N(0, I)$
- Z independent of X

Problem: Recover X from data Y. Estimate $\hat{X} = T(Y)$

Find $\hat{X}$ such that MSE is minimum

$$MSE(X, \hat{X}) = E\|X - \hat{X}\|_2^2 = E\sum_{t=0}^{T-1}(X_t - \hat{X}_t)^2$$

The solution is given by

$$T(Y) = E\{X|Y\}$$

This is the classical regression problem!

Why? We want to choose $T(Y)$ such that $E\|X - T(Y)\|^2$ is minimum. In other words, we want to minimize

$$\int (T(y) - x)^2 p(x, y)\, dxdy$$

where p is the joint distribution of (x, y).

$$\int (T(y) - x)^2 p(x, y)\, dxdy = \int \left[\int (T(y) - x)^2 p(x|y)\, dx \right] dy$$

The solution is to minimize the integrand for each value of y.

$$T(y) = \text{argmin}_\mu \int (x - \mu)^2 p(x|y)\, dx$$

and the solution is

$$T(y) = \int x p(x|y)\, dx = E(X|Y = y).$$

Wiener's Filter

In our setup, MSE Estimation Problem:

How to get $E\{X|Y\}$

Solution:

$$E\{X|Y\} = \sum_{k=1}^{\infty} w_k \langle Y, \phi_k \rangle \phi_k$$

Wiener Filter weights

$$w_k = \frac{\text{Signal Power}}{\text{Signal+ Noise Power}} = \frac{\lambda_k^2}{\lambda_k^2 + \sigma^2}$$

Sequence Space View

$$Y_t = X_t + \sigma Z_t$$

Change of basis:

$$\begin{aligned} \langle Y, \phi_k \rangle &= \langle X, \phi_k \rangle + \sigma \langle Z, \phi_k \rangle \\ y_k &= \theta_k + \sigma z_k \end{aligned}$$

- z_k Gaussian white noise sequence, z_k i.i.d. $N(0, 1)$
- σ noise level
- $\theta_k = \langle X, \phi_k \rangle$ coordinates of X
  - The θ_k's are i.i.d. $N(0, \lambda_k^2)$
  - The θ_k's are independent of the z_k's

- Isometry

  $$\|X - Y\|^2 = \|\theta - y\|^2$$

  and, of course,

  $$E\|X - Y\|^2 = E\|\theta - y\|^2$$

- Solution

  $$\theta_k(y) = E(\theta_k|y) = E(\theta_k|y_k)$$

- From

  $$\begin{array}{ccccc} y_k & = & \theta_k & + & z_k \\ N(0, \lambda_k^2 + \sigma^2) & & N(0, \lambda_k^2) & & N(0, \sigma^2) \end{array}$$

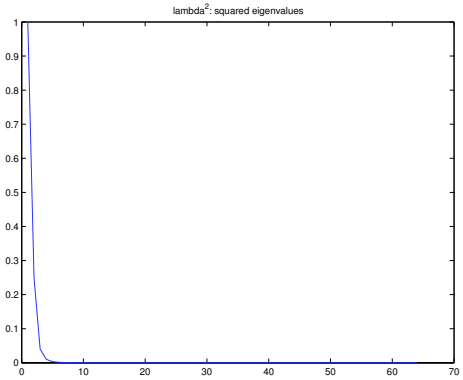
  obtain

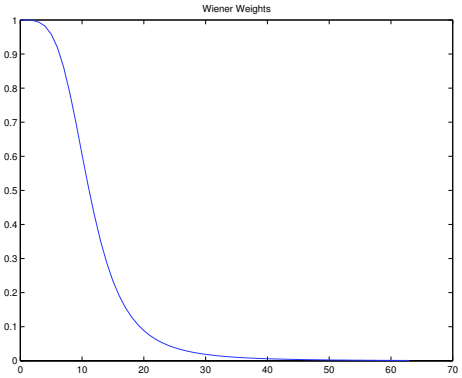
  $$E(\theta_k|y_k) = \frac{\lambda_k^2}{\lambda_k^2 + \sigma^2}\, y_k$$

  i.e.

  $$E(X|Y) = \sum_k w_k y_k\, \phi_k$$

Wiener's Weights





Example

$$
\begin{aligned}
\phi_0(t) &= 1/\sqrt{T} \\
\phi_{2k-1}(t) &= \sqrt{2/T}\cos(2\pi kt/T), \quad k = 1, 2, \ldots, (T-1)/2 \\
\phi_{2k}(t) &= \sqrt{2/T}\sin(2\pi kt/T), \quad k = 1, 2, \ldots, (T-1)/2 \\
\phi_{T-1}(t) &= (-1)^t/\sqrt{T}
\end{aligned}
$$

Synthesize a Gaussian process:

$$X = \sum_k \lambda_k Z_k \phi_k(t)$$

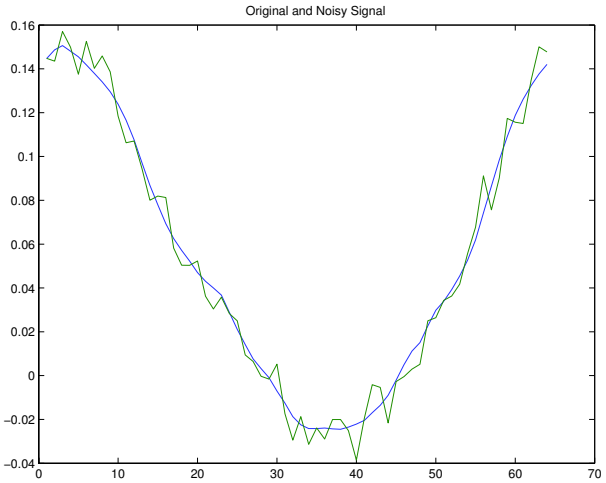
```
n = 64;
k = 0:n-1;
lambda = 1./(1 + k.^2);
figure; plot(lambda.^2)

Z = randn(1,n);
Xp = lambda.*Z;
X = inv_my_dct(Xp);
```


Synthesize noisy data:

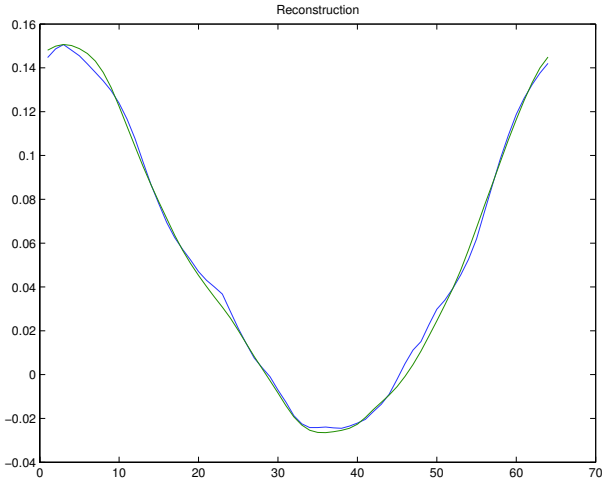
```
sigma = .003*sqrt(n);
Y = X + sigma*randn(1,n);
```

Apply Wiener's filter:

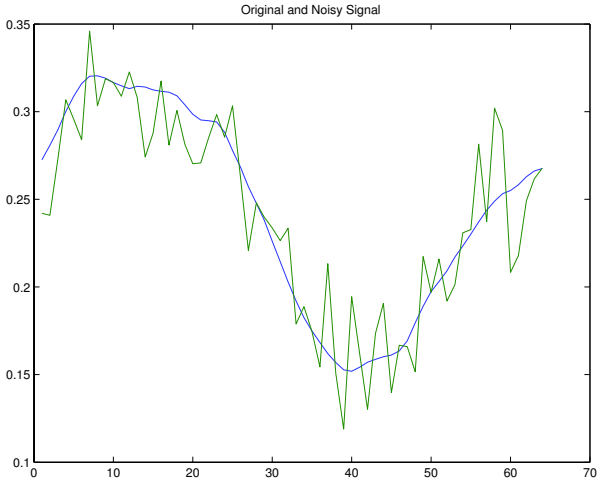
```
FY = my_dct(Y);
Weights = lambda.^2 ./(lambda.^2 + sigma.^2);
FR = FY.*Weights;
R = inv_my_dct(FR);
```

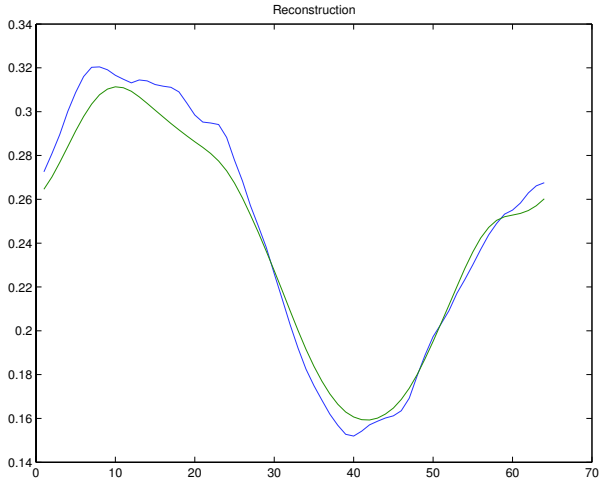
21



22

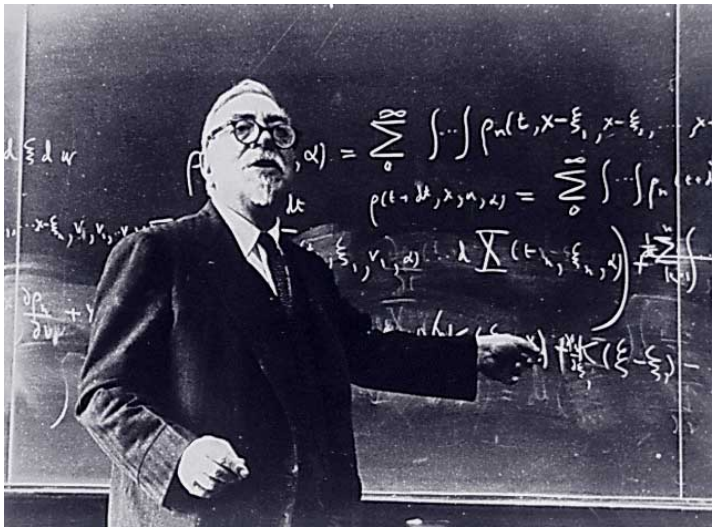


23

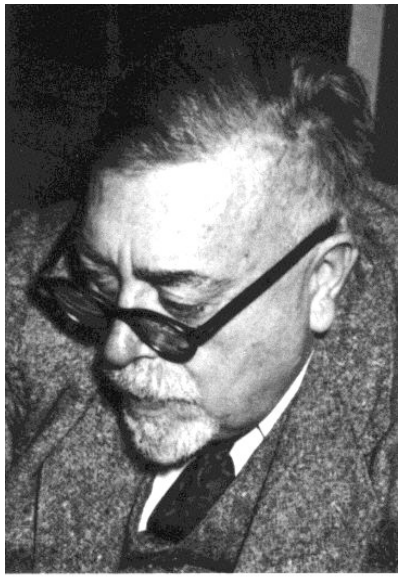


24

Norbert Wiener



Shannon and Wiener



NORBERT WIENER
1894–1964

Compression of Gaussian Processes

Rate-Distortion Theory, Shannon 1948.

Lossy data compression of continuous-valued stochastic processes. " How many bits are required to approximately represent sample paths of a stochastic process?" That is to represent $X_1, \ldots, X_n$.

X_t, $t = 1, \ldots, n$, a zero-mean Gaussian stochastic with covariance Σ.

Encoder/Decoder pair

- Encoder

$$X \mapsto \{0,1\}^L$$

  i.e. maps a point in R^n into a bitstring of length L.

- Decoder

$$\{0,1\}^L \mapsto \hat{X}$$

  i.e. maps a bitstring of length L into a point in R^n.

Terminology

- The 2^L elements $\hat{X}(b)$ (b is a bitstring) are the codewords
- The collection of all codewords is the codebook.

Central Question

Let $N(D, X)$ denote the minimal number of codewords needed in a codebook $\mathcal{C} = \{\hat{X}\}$ so that

$$E \min_{X' \in \mathcal{C}} \|X - \hat{X}\|^2_{L_2(T)} \leq D.$$

Encoding using closest-point mapping:

$$X^* = \text{argmin}_{\hat{X} \in \mathcal{C}} \|X - \hat{X}\|_{L_2(T)}$$

Number of bits required $\sim \log(N(D, X))$.

Baby Example

We want to use one bit to encode $X \sim N(0, \sigma^2)$

- The bit should distinguish whether $X > 0$ or not
- To minimize the mean squared error, the symbol should be at the conditional mean of its region.

$$\hat{X}(1) = E(X|X > 0) = 2\int_0^\infty x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\, dx = \sqrt{\frac{2}{\pi}}\sigma.$$

- Solution: send bit = 1, if $X > 0$ and decode with $.7979\sigma$; otherwise, send bit = 0 and decode with $-.7979\sigma$;

Rate-Distortion Theory

Shannon proposed that in an asymptotic sense

$$\log(N(D, X))/R(D, X) \to 1, \quad \text{as } n \to \infty$$

where $R(D, X)$ is the rate-distortion function for X

$$R(D) = \inf I(X, Y) \text{ s.t. } E\|X - Y\|^2_{L_2(T)} \leq D,$$

with $I(X, Y)$ the usual mutual information

$$I(X, Y) = \int\int p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, dxdy$$

Here the minimum is to be understood over all the conditional probability distributions $p_{Y|X}$.

- Formal solutions are difficult to evaluate.
- Actual calculations of rates have been carried out in only a few very simple cases; e.g. X Gaussian process.

Partial Formal Solution

Lagrange's method

$$\int\int \left[p(x, y) \log \frac{p(x, y)}{p(x)p(y)} + \mu p(x, y)(x - y)^2 + \nu(x)p(x, y)\right] dxdy$$

- Variational equation (take first variation in $p(x, y)$)

$$p(x|y) = B(x)e^{-\lambda(x-y)^2},$$

where $B(x)$ obeys

$$\int B(x)e^{-\frac{1}{2\lambda^2}(x-y)^2}\, dx = 1.$$

This says that $B(x)$ is a constant equal to B such that $B\int e^{-\lambda u^2}\, du = 1$.

$$p(x|y) = Be^{-\lambda(x-y)^2}$$

Hence, the conditional distribution of X given $Y = y$ is Gaussian with mean y and variance $\lambda^2 = D$.

- Marginals

$$p(x) = \int p(x|y)p(y)\,dy.$$

Assume $X \sim N(0, \sigma^2)$, this gives

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} = Be^{-\lambda x^2} \int e^{2\lambda xy}\, e^{\lambda y^2} p(y)\,dy.$$

This says that the Laplace transform of $e^{\lambda y^2} p(y)$ must be Gaussian, hence $e^{\lambda y^2} p(y)$ must be Gaussian and, hence, $p(y)$ must be Gaussian.

- Solution

$$\begin{array}{ccccc} X & = & Y & + & Z \\ N(0,\sigma^2) & & N(0,\sigma^2 - D) & & N(0,D) \end{array}$$

Example I

$X \sim N(0, \sigma^2)$.

- The variable X^* that minimizes the mutual information has the following structure

$$\begin{array}{ccccc} X & = & X^* & + & Z \\ N(0,\sigma^2) & & N(0,\sigma^2 - D) & & N(0,D) \end{array}$$

- The rate-distortion function is given by

$$R(D) = E\left(\log \frac{p(X, X^*)}{p(X)p(X^*)}\right) = E\left(\log \frac{p(X|X^*)}{p(X)}\right),$$

with

$$\log \frac{p(X|X^*)}{p(X)} = -\frac{\log D}{2} - \frac{(X - X^*)^2}{2D} + \frac{\log \sigma^2}{2} + \frac{X^2}{2\sigma^2}.$$

Hence,

$$R(D) = \begin{cases} \frac{1}{2}\log(\sigma^2/D) & \sigma^2 > D \\ 0 & \sigma^2 \le D \end{cases},$$

Example II

$X = (X_1, \ldots, X_n)$ with $X_i \sim N(0, \sigma_i^2)$ and the X_i's are independent

- The process X^* which minimizes the mutual information has the following structure

$$\begin{array}{ccccc} X_i & = & X_i^* & + & Z \\ N(0,\sigma_i^2) & & N(0,\sigma_i^2 - D_i) & & N(0,D_i) \end{array}$$

and the X_i^*'s are independent.

- How to allocate the bits?
- Overall distortion

$$E\|X - X^*\|^2 = \sum_i E(X_i - X_i^*)^2 = \sum_i D_i$$

- Overall Rate

$$R(D) = \sum_i \frac{1}{2} \max(0, \log(\sigma_i^2/D_i))$$

- Minimize $R(D)$ subject to $\sum_i D_i = D$.

- Lagrange's method

$$J(D) = \sum_i \frac{1}{2} \log(\sigma_i^2/D_i) + \lambda \sum_i D_i$$

First variation

$$\partial J/\partial D_i = 0 = -\frac{1}{2D_i} + \lambda$$

+ Kuhn Tucker conditions.

- Solution:

$$D_i = \min(\theta, \sigma_i^2)$$

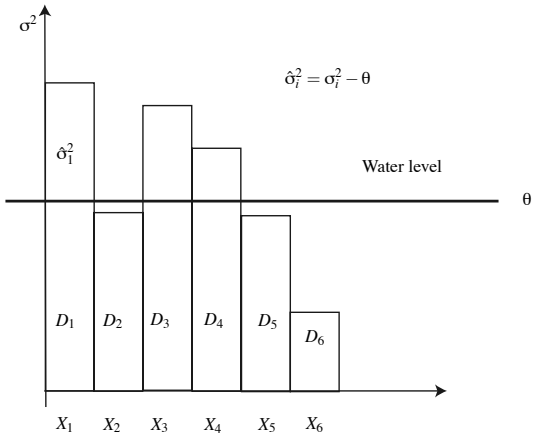
- The rate-distortion function is given by

$$R(D_\theta) = \sum_{\sigma_i^2 > \theta} \frac{1}{2} \log(\sigma_i^2/\theta)$$

where

$$D_\theta = \sum_i \min(\theta, \sigma_i^2).$$

Reverse water-filling



Coding to reach $R(D)$

$X = (X_1, \ldots, X_n)$ with $X_i \sim N(0, \sigma_i^2)$ and the X_i's are independent

- X^* with minimum mutual information

$$X_i^* \sim N(0, \sigma_i^2 - \theta)$$

- Construction of a codebook. Sample from the process X^* described above and obtain codewords $X^{*(1)}, X^{*(2)}, \ldots, X^{*(N)}$.
- Random codebook compression, pick k such that

$$\|X - X^{*(k)}\|$$

is minimum.

Compression of Gaussian processes

- $X \sim N(0, \Sigma)$
- $Z = \Phi^T X$, Z_i $i.i.d.$ $N(0, \lambda_i^2)$
  $\phi_i(t)$ are the eigenfunctions of the covariance matrix Σ.
- Isometry

$$E\|X - X'\|^2 = E\|\Phi^T X - \Phi^T X'\|^2 = E\|Z - Z'\|^2$$

Just studied the coding of Z_i $i.i.d.$ $N(0, \lambda_i^2)$

- Z^* minimizes the mutual information
- Random codebook $Z^{*(1)}, Z^{*(2)}, \ldots, Z^{*(N)}$.
- X^* minimizes the mutual information
- Random codebook $X^{*(1)} = \Phi Z^{*(1)}, \ldots$

Formally,

$$X^*(t) = \sum_i Z_i^* \phi_i(t), \quad Z_i^* \sim N(0, \mu_i^2), \quad \mu_i^2 = \lambda_i^2 - \theta.$$

Recall

$$X(t) = \sum_i Z_i \phi_i(t), \quad Z_i \sim N(0, \lambda_i^2)$$

Process X^*

- Covariance with the same eigenfunctions as that of X,
- but the eigenvalues are reduced in size

$$\mu_i^2 = (\lambda_i^2 - \theta)_+$$

- Only finitely many nonzero coefficients. Only coefficients above a certain water level are described in the coding process.

Example: Stationary Process

- Covariance $\Sigma_{ij} = \sigma(i-j)$
- Orthonormal basis of eigenfunctions $\phi_i(t)$ is Fourier

Assume λ_i is non-increasing as frequency index increases.

For a fixed distortion D, we get the water-level θ.

Interpretation

- Go into the frequency domain
- Extract low-frequency coefficients
- Compare with codebook entries

Kernel Smoothing, I

Statistical model

$$y_i = f(t_i) + \epsilon_i, \quad 1 \le i \le n$$

- y: data (available)
- ϵ: iid stochastic errors
- f: object of interest we wish to recover (unknown)
- Nonparametrics: f completely unknown
- Goal: estimate f from data y

Kernel Smoothing, II

- Graph of averages
- $K(t)$ is a kernel (usually, symmetric density)
- $K \ge 0$, $\int K(t)\, dt = 1$.
- Examples:
  - Box: $K(t) = 1_{\{-1/2 \le t < 1/2\}}$
  - Gaussian: $K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Kernel smoother is of the form ($\hat{f}$ is a weighted average)

$$\hat{f}(t) = \frac{\sum_i w_i y_i}{\sum_i w_i}, \quad w_i = K((t - t_i)/h)$$

h is the bandwidth

- Closer points $\longrightarrow$ larger weights
- Small bandwidth $\longrightarrow$ averaging over very few points
- Large bandwidth $\longrightarrow$ averaging over many points

Kernel Smoothing, III

Assume

- Equispaced design: $t_i = (i-1)/n$
- The estimand is periodic

Useful assumptions for getting simple results

Frequency-side picture:

$$g = k * y$$

with $k_j = K(j/nh)$ and $\sum_j k_j = 1$. In the frequency-domain

$$\hat{g}_j = \hat{k}_j \hat{y}_j$$

Same structure: damp high frequencies

Spline Smoothing

$$g_\lambda = \text{argmin} \sum_i (y_i - g(t_i))^2 + \lambda \int_0^1 |g''(u)|^2 \, du$$

Trade-off

- Goodness of fit
- Complexity of the fit

Complexity-penalized fitting

Where does the name come from? Solution is a cubic spline with nodes (t_i)

Spline Smoothing: Frequency Viewpoint

- Model
$$y_i = f_i + \epsilon_i$$
- Fit (discrete approximation for simplicity)

  $\text{argmin} \sum_i (y_i - g_i)^2 + \lambda \sum_i |(D^2 g)_i|^2, \quad (D^2 g)_i = g_{i+1} - 2g_i + g_{i-1}.$
- Fourier isometry
$$\text{argmin} \sum_k (\hat{y}_k - \hat{g}_k)^2 + \lambda \sum |d_k|^2 |\hat{g}_k|^2$$
  with $d_k := 2(1 - \cos(2\pi k/n))$
- Solution
$$\hat{g}_k = \frac{\hat{y}_k}{1 + \lambda |d_k|^2}$$
- Same structure: damp high frequencies.

Statistical Theory

$\mathcal{F} = \{f : \|f\|_{W_2^m} \leq C\}$

Data

$$Y(dt) = f(t)dt + \epsilon W(dt) \quad t \in [0,1)$$

- $Y(dt)$ observed
- $W(dt)$ white noise
- Sobolev norm: $\|f\|^2_{W_2^m} = \int |f(t)|^2 + |f^{(m)}(t)|^2 \, dt$

Seek estimator $\hat{f}$ so that

$$\sup_{\mathcal{F}} E\|\hat{f} - f\|_2^2 = \min$$

- Makes no reference to basis – totally intrinsic
- Puts no restriction on estimator – all measurable procedures allowed.

Pinsker's Solution

- Answer (asymptotic)

$$\hat{f} = \sum_{k \geq 0} w_{k,\epsilon} \langle Y, \phi_k \rangle \phi_k$$

- Fourier Basis:

$$\phi_{2k}(t) = \cos(2\pi k t), k \geq 0; \quad \phi_{2k-1}(t) = \sin(2\pi k t), k \geq 1.$$

- Pinsker weights:

$$w_{k,\epsilon} = (1 - \lambda(\epsilon) k^{2m})_+$$

Interpretation: Go into Fourier Basis, Throw out High Frequencies.

Minimax Estimation Over Ellipsoid

Sequence Model

$$\begin{aligned} \langle Y, \phi_k \rangle &= \langle f, \phi_k \rangle + \epsilon \langle W, \phi_k \rangle \\ y_k &= \theta_k + \epsilon z_k \end{aligned}$$

- $\theta_k = \langle X, \phi_k \rangle$, Fourier coefficients
- $f \in \mathcal{F}^m(C)$

$$\sum_k k^{2m} (|\theta_{2k-1}^2 + |\theta_{2k}|^2) \leq C$$

- Problem is equivalent to recovering $\theta \in \Theta$ from

$$y \sim N(\theta, \epsilon I)$$

Conclusion

- Intrinsic problems
  - statistical estimation
  - data compression (rate distortion theory)
- Optimal representations emerge naturally
- Solid statistical theory