Optimal Filtering
II: Linear Regression and Correlation Matrices
ECE416 Adaptive Algorithms

Prof. Fred L. Fontaine

Department of Electrical Engineering
The Cooper Union

Spring 2022

Linear Regression

Let $\mathbf{X} = [X_1 \; X_2 \; \cdots \; X_M]^T$ denote a random vector, and Y a random variable. We seek a vector $\mathbf{w}$ such that:

$$Y \approx \mathbf{w}^H \mathbf{X} = \sum_{m=1}^{M} w_m^* X_m$$

To be more precise:

$$\mathbf{w} = \arg\min E\left(\left|Y - \mathbf{w}^H X\right|^2\right)$$

Linear Regression

Via the orthogonality principle, if $e = Y - \mathbf{w}^H \mathbf{X}$ is the estimation error, we need $e \perp X_m$ for each $1 \leq m \leq M$, which we can write as:

$$E\left(\mathbf{X} e^*\right) = \mathbf{0}$$

There are M linear equations in the w_m's, called the orthogonality conditions. Since e is a scalar, $e^* = e^H$:

$$E\left(\mathbf{X}\left(Y - \mathbf{w}^H \mathbf{X}\right)^H\right) = E\left(\mathbf{X} Y^*\right) - E\left(\mathbf{X}\mathbf{X}^H\right)\mathbf{w} = \mathbf{0}$$

Linear Regression

Let $R = E\left(\mathbf{X}\mathbf{X}^H\right)$ be the autocorrelation matrix of $\mathbf{X}$, and $\mathbf{p} = E\left(\mathbf{X}Y^*\right)$ be the cross-correlation vector between $\mathbf{X}$ and Y (each component is $p_i = E\left(X_i Y^*\right)$).

Then the optimal vector $\mathbf{w}_0$ satisfies $\mathbf{p} - R\mathbf{w}_0 = \mathbf{0}$ or:

$$\mathbf{R}\mathbf{w}_0 = \mathbf{p}$$

Thus:

$$\mathbf{w}_0 = \mathbf{R}^{-1}\mathbf{p}$$

FUNDAMENTAL EQUATION

The entire course can be encapsulated in that one equation:

$$\mathbf{w}_0 = R^{-1}\mathbf{p}$$

.

Linear Regression

Define e_0 as the optimal error when we use $\mathbf{w}_0$:

$$Y = \mathbf{w}_0^H \mathbf{X} + e_0 = \sum w_{0m}^* X_m + e_0$$

where $e_0 \perp X_m$ for all m. This is actually the definition of a linear regression model: the error term is orthogonal (uncorrelated) with the "model" representing Y, i.e., $\mathbf{w}_0^H \mathbf{X}$.

Here we are assuming all quantities are 0-mean *so orthogonal = uncorrelated. In the Gaussian case, we get* **independence** *as well!*

Linear Regression

In many cases, a linear regression model also contains a constant term, but that can be obtained by augmenting $\mathbf{X}$ with a constant. That is, let:

$$\mathbf{X}' = [1 \; X_1 \; \cdots X_M]^T$$

so the regression model takes the form:

$$Y = a_0 + w_1^* X_1 + \cdots + w_M^* X_M + e$$

where a_0 is a deterministic constant, and e is 0-mean random error.

If say the X_m's are all zero mean with correlation matrix R, then the correlation matrix of $\mathbf{X}'$ is:

$$R' = \begin{bmatrix} 1 & \mathbf{0}_M^T \\ \mathbf{0}_M & \mathbf{R} \end{bmatrix}$$

where $\mathbf{0}_M$ is the $M \times 1$ zero vector, so R' is invertible if R is.

Error Performance Surface

Here we tried to find $\mathbf{w}$ to minimize the cost function:

$$J(\mathbf{w}) = E\left(\left|y - \mathbf{w}^H\mathbf{x}\right|^2\right)$$

Considering J as a function of $\mathbf{w}$ is the *error performance surface.*

It is customary to assume $\mathbf{x}$, y are 0-mean, so $E\left(|y|^2\right) = \sigma_y^2$. Expanding out gives:

$$\begin{aligned} J(\mathbf{w}) &= E\left(|y|^2\right) - \mathbf{w}^H E\left(\mathbf{x}y^*\right) - E\left(y\mathbf{x}^H\right)\mathbf{w} + \mathbf{w}^H E\left(\mathbf{x}\mathbf{x}^H\right)\mathbf{w} \\ &= \sigma_y^2 - \mathbf{w}^H\mathbf{p} - \mathbf{p}^H\mathbf{w} + \mathbf{w}^H R_x \mathbf{w} \end{aligned}$$

Error Performance Surface

Rather than directly using the orthogonality principle, we can compute:

$$\nabla J = -2\mathbf{p} + 2R\mathbf{w} = \mathbf{0}$$

which again gives us the optimal solution:

$$\mathbf{w}_0 = R^{-1}\mathbf{p}$$

Note:

$$\mathbf{p} = R\mathbf{w}_0$$

Error Performance Surface

Substituting this in gives $J_{\min} = J(\mathbf{w}_0)$ given by:

$$\begin{aligned} J_{\min} &= \sigma_y^2 - \mathbf{p}^H R^{-1}\mathbf{p} - \mathbf{p}^H R^{-1}\mathbf{p} + \mathbf{p}^H R^{-1} R R^{-1}\mathbf{p} \\ &= \sigma_y^2 - \mathbf{p}^H R^{-1}\mathbf{p} \\ &= \sigma_y^2 - \mathbf{w}_0^H R \mathbf{w}_0 \end{aligned}$$

Error Performance Surface

Note that $\mathbf{w}_0$ reduces the variance of y, σ_y^2, by the amount $\mathbf{w}_0^H R \mathbf{w}_0$.

We now want to go back to $J(\mathbf{w})$ and obtain a different expression. We could use the orthogonality principle but let us work directly in the vector form:

$$\begin{aligned}\mathbf{w}^H R \mathbf{w} &= (\mathbf{w}-\mathbf{w}_0)^H R (\mathbf{w}-\mathbf{w}_0) + \mathbf{w}_0^H R (\mathbf{w}-\mathbf{w}_0) \\ &\quad + (\mathbf{w}-\mathbf{w}_0)^H R \mathbf{w}_0 + \mathbf{w}_0^H R \mathbf{w}_0\end{aligned}$$

$$\begin{aligned}\mathbf{w}^H R \mathbf{w} &= (\mathbf{w}-\mathbf{w}_0)^H R (\mathbf{w}-\mathbf{w}_0) + \mathbf{p}^H (\mathbf{w}-\mathbf{w}_0) + \\ &\quad (\mathbf{w}-\mathbf{w}_0)^H \mathbf{p} + \mathbf{w}_0^H R \mathbf{w}_0\end{aligned}$$

Error Performance Surface

$$\begin{aligned}\mathbf{w}^H\mathbf{p}+\mathbf{p}^H\mathbf{w} &= \mathbf{w}_0^H\mathbf{p}+\mathbf{p}^H\mathbf{w}_0+\left(\mathbf{w}-\mathbf{w}_0\right)^H\mathbf{p}+\mathbf{p}^H\left(\mathbf{w}-\mathbf{w}_0\right)\\ &= 2\mathbf{w}_0^H R\mathbf{w}_0+\left(\mathbf{w}-\mathbf{w}_0\right)^H\mathbf{p}+\mathbf{p}^H\left(\mathbf{w}-\mathbf{w}_0\right)\end{aligned}$$

Subtracting this from $\mathbf{w}^H R\mathbf{w}$, adding σ_y^2 yields:

$$J\left(\mathbf{w}\right)=\sigma_y^2-\mathbf{w}_0^H R\mathbf{w}_0+\left(\mathbf{w}-\mathbf{w}_0\right)^H R\left(\mathbf{w}-\mathbf{w}_0\right)$$

Error Performance Surface

Our final expression is:

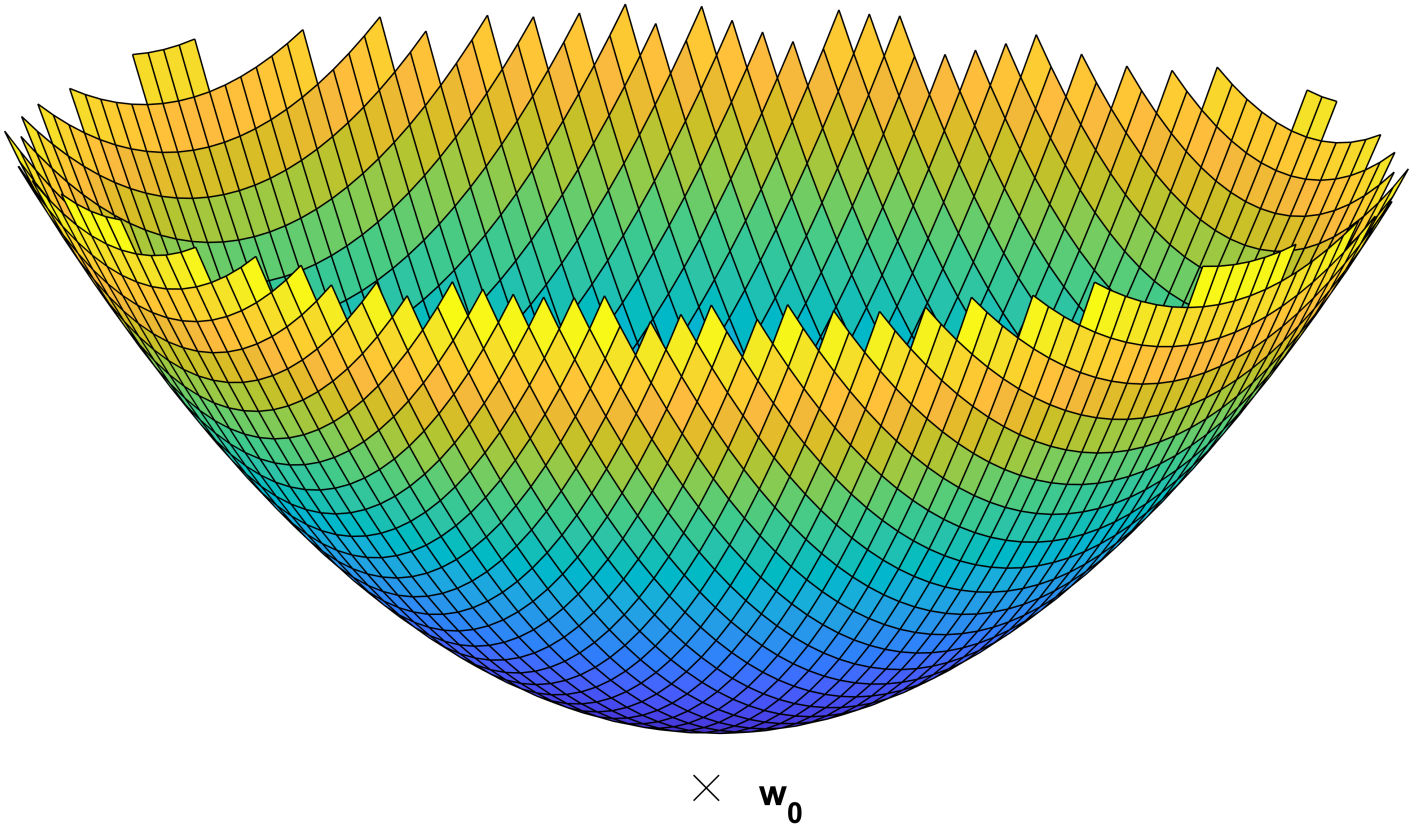
$$J(\mathbf{w}) = J_{\min} + (\mathbf{w} - \mathbf{w}_0)^H R (\mathbf{w} - \mathbf{w}_0)$$

Since R is pd, this has a *unique* global minimum at $\mathbf{w} = \mathbf{w}_0$. The surface has the shape of a *paraboloid*.

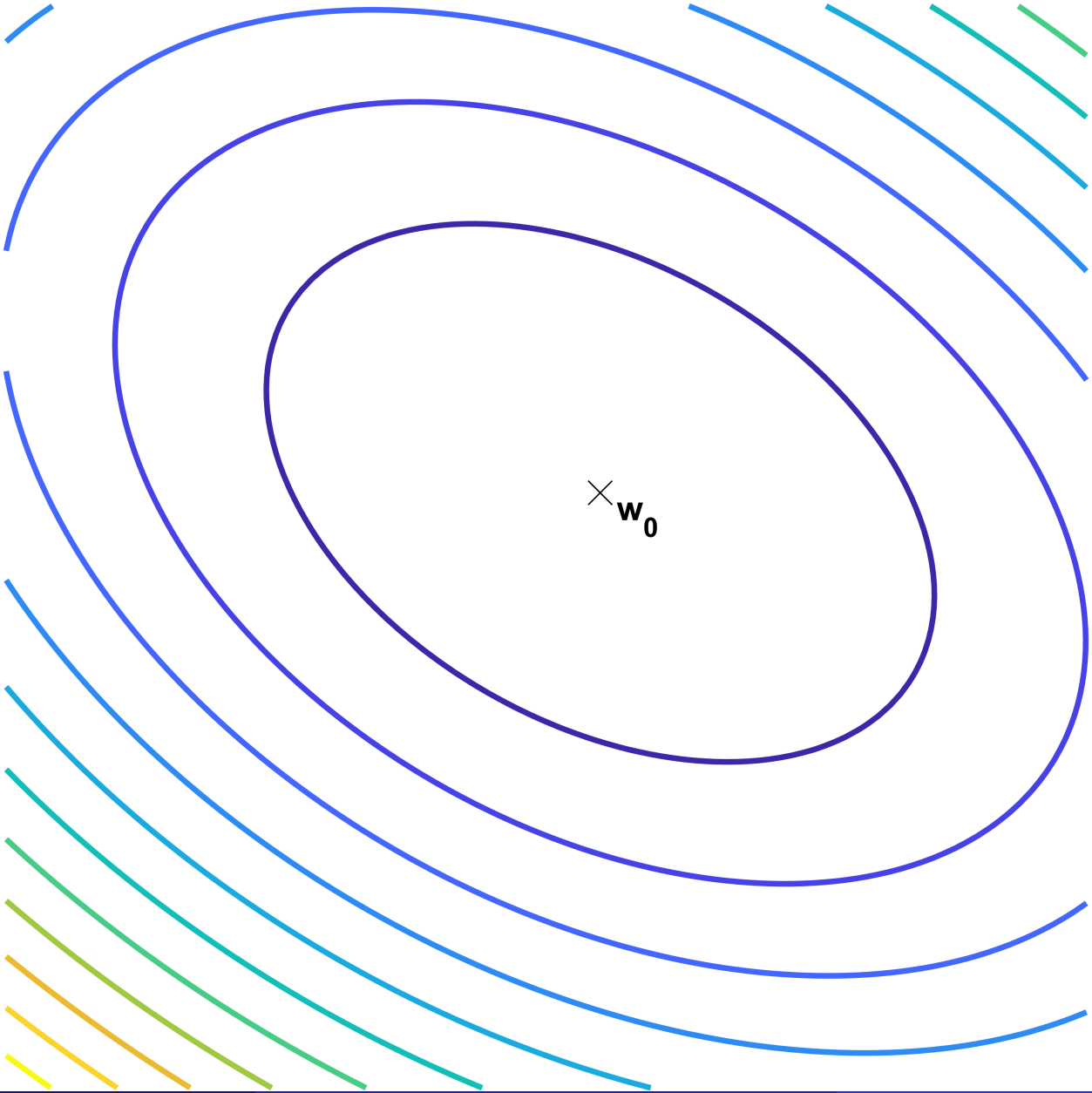
When $\mathbf{w} \neq \mathbf{w}_0$, we have an *excess error:*

$$J_{ex}(\mathbf{w}) = J(\mathbf{w}) - J_{\min} = (\mathbf{w} - \mathbf{w}_0)^H R (\mathbf{w} - \mathbf{w}_0)$$

Error Performance Surface



Error Performance Surface: Contour Plot



Ellipsoidal Contours

If $\{\mathbf{q}_i\}$ are the orthonormal eigenvectors of R, we can decompose:

$$\boldsymbol{\varepsilon} = \mathbf{w} - \mathbf{w}_0$$

into eigencomponents as:

$$\boldsymbol{\varepsilon} = \sum \varepsilon_i \mathbf{q}_i$$

with $\varepsilon_i = \mathbf{q}_i^H \boldsymbol{\varepsilon}$.

Ellipsoidal Contours

Note that, with $R = Q^H \Lambda Q$:

$$\boldsymbol{\varepsilon}^H R \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^H Q \Lambda Q^H \boldsymbol{\varepsilon} = \sum \lambda_i \left| \varepsilon_i \right|^2$$

This means the contours of constant $J(\mathbf{w})$ are *ellipsoids* with axes in the eigenvector directions (lengths equal to $1/\sqrt{\lambda_i}$).

Remark: This has a stochastic interpretation: if $\mathbf{w} \sim \mathbf{N}\left(\mathbf{w}_0, R^{-1}\right)$, i.e., the mean vector is $\mathbf{w}_0$ and the covariance matrix is R^{-1}, then these are *isoprobability* contours (the pdf $f(\vec{w})$ is constant on them).

Numerical Issues

$$\mathbf{w}_0 = R^{-1}\mathbf{p}$$

There are two problems in practice, closely related.

1. We do not know R and can only estimate it.
2. Whether we know R exactly or only an estimate, computing R^{-1} can be problematic.

Note that in many cases, the data is first "centered" by subtracting off the (sample) mean, in which case we focus on the covariance matrix C instead of R. Nevertheless, the same issues hold.

Numerical Issues

- R may be **huge,** with thousands, tens of thousands or even more entries.
- R may be close to singular. Usually with the presence of noise, R will be invertible, but it may have small eigenvalues close to 0. Recall the eigenvalues of R^{-1} are $1/\lambda(R)$ (here $\lambda(R)$ means eigenvalues of R).
- When we only have an estimate of R, estimation errors may be magnified greatly via the inversion process. Indeed, the smallest eigenvalues of R determine the largest eigenvalues of R^{-1}.

Condition Number

The *condition number* of a square matrix $\chi(A)$ is often used as a measure of how difficult it is to work with A from a numerical perspective.

Use $\|A\|$ to denote the spectral norm of a matrix, i.e., $\|A\| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$:

$$\chi(A) = \|A\| \cdot \left\|A^{-1}\right\|$$

Condition Number

The condition number roughly represents the spread of magnitudes of the entries of A and A^{-1}. For example:

$$A = \begin{bmatrix} 10 & 0.8 \\ 0.3 & 0.025 \end{bmatrix} \longrightarrow A^{-1} = \begin{bmatrix} 2.5 & -80 \\ -30 & 1000 \end{bmatrix}$$

In this case $\|A\| \approx 10.04$, $\left\|A^{-1}\right\| \approx 1004$ so $\chi(A) \approx 10,007$.

Condition Number

There is a property of spectral norm (does not just apply to square matrices):

$$\|AB\| \leq \|A\| \, \|B\|$$

From this we get:

$$1 \leq \chi(A) \leq \infty$$

where again ∞ is if A^{-1} does not exist.

A large condition number means handling the matrix (not even necessarily trying to invert it) is sensitive to numerical errors.

When $\chi(A)$ is large, we say A is *ill-conditioned*.

Error Analysis (Golub)

Let d be a vector, A a matrix, and:

$$Ad = w$$

Represent d, A to some precision yielding errors δd, δA. Let:

$$\epsilon = \max\left(\frac{|\delta d|}{|d|}, \frac{\|\delta A\|}{\|A\|}\right)$$

where $|\cdot|$ for a vector is Euclidean length, $\|\cdot\|$ for a matrix is spectral norm.

Error Analysis (Golub)

$$(A + \delta A)(d + \delta d) = w + \delta w \approx w$$

With small perturbations we get:

$$\frac{|\delta w|}{|w|} \leq \epsilon \chi(A)$$

Eigenvalue Spread

For a *Hermitian* matrix A:

$$\chi(A) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}}$$

where $|\lambda|_{\max}$, $|\lambda|_{\min}$ denote the maximum and minimum absolute values of the eigenvalues (remember they are all real). This is called the *eigenvalue spread*.

The reason is that $\|A\| = |\lambda|_{\max}$, and $\lambda\left(A^{-1}\right) = 1/\lambda(A)$ (the eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A), so $\left\|A^{-1}\right\| = 1/|\lambda|_{\min}$.

Eigenvalue Spread

For a correlation matrix, we can drop the absolute values:

$$\chi(R) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Going back to:

$$w_0 = R^{-1}p$$

we see that the eigenvalue spread of R plays a significant role in the numerical issues associated with estimation problems.

Correlation Matrix for WSS Signals

Throughout much of this course we will work with $u[n]$ WSS with correlation $r[m]$, PSD $S(\omega)$. Define:

$$u_M[n] = \begin{bmatrix} u[n] \\ \vdots \\ u[n-M+1] \end{bmatrix}$$

With the index $1 \leq i \leq M$ as usual, we see that the i^{th} element is $u[n-i+1]$. With $R = E\left(u_M[n]\, u_M^H[n]\right)$, the ij^{th} element is:

$$E\left(u[n-i+1]\, u^*[n-j+1]\right) = r[(n-i+1)-(n-j+1)] = r[j-i]$$

Toeplitz

Let R_M denote the $M \times M$ correlation matrix of $u_M[n]$. Using $r_m = r[m]$, for example for $M = 5$ we get:

$$R_5 = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 & r_4 \\ r_{-1} & r_0 & r_1 & r_2 & r_3 \\ r_{-2} & r_{-1} & r_0 & r_1 & r_2 \\ r_{-3} & r_{-2} & r_{-1} & r_0 & r_1 \\ r_{-4} & r_{-3} & r_{-2} & r_{-1} & r_0 \end{bmatrix}$$

The matrix has *Toeplitz symmetry:* it is constant on the diagonals. Since $r_m = r_{-m}^*$, it is Hermitian as well. It is also pd (unless there is a degeneracy).

Toeplitz Correlation Matrices

The Toeplitz symmetry leads to interesting "embeddings." For example, R_4 can be seen inside R_5.

Let us define the $\mathbf{r}_M$ vector as:

$$\mathbf{r}_M = \begin{bmatrix} r_{-1} \\ r_{-2} \\ \vdots \\ r_{-M} \end{bmatrix} = \begin{bmatrix} r_1^* \\ r_2^* \\ r_3^* \\ r_4^* \end{bmatrix}$$

Also, for a column vector $\mathbf{x}$, define $\mathbf{x}^B$ by flipping its components upside-down:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \longrightarrow \mathbf{x}^B = \begin{bmatrix} x_N \\ \vdots \\ x_2 \\ x_1 \end{bmatrix}$$

Finally $\mathbf{x}^*$ means just conjugating the entries (no transpose).

Toeplitz Correlation Matrices

$$R_{M+1} = \begin{bmatrix} R_M & \mathbf{r}_M^{B*} \\ \mathbf{r}_M^{BT} & r_0 \end{bmatrix} = \begin{bmatrix} r_0 & \mathbf{r}_M^{H} \\ \mathbf{r}_M & R_M \end{bmatrix}$$

Correlation and PSD for WSS Processes

Given a vector $\mathbf{w} = [w_0\ w_1\ \cdots\ w_{M-1}]^T$, define:

$$W(\omega) = \sum_{m=0}^{M-1} w_m^* e^{-j\omega m}$$

Then:

$$\mathbf{w}^H R_M \mathbf{w} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |W(\omega)|^2 S(\omega)\, d\omega$$

Correlation and PSD for WSS Processes

To verify this:

$$
\begin{aligned}
|W(\omega)|^2 &= \sum_{m=0}^{M-1}\sum_{n=0}^{M-1} w_m^* w_n e^{j\omega(n-m)} \\
\frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 S(\omega)\, d\omega &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_m^* w_n \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\omega(n-m)} S(\omega)\, d\omega\right] \\
&= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_m^* w_n r[n-m] \\
&= \mathbf{w}^H R_M \mathbf{w}
\end{aligned}
$$

Note that the property $S(\omega) \geq 0$ is equivalent to the positive semi-definiteness of R_M. In fact, if R_M fails to be positive definite, then $S(\omega)$ must be 0 at all but a finite number of points.

Correlation and PSD for WSS Processes

If $\|\mathbf{w}\|^2 = \sum |w_m|^2 = 1$, then:

$$\lambda_{\min}(R_M) \leq \mathbf{w}^H R_M \mathbf{w} \leq \boldsymbol{\lambda}_{\max}(R_M)$$

with the equalities on either side attained when $\mathbf{w}$ is a corresponding eigenvector.

But by Parseval's theorem, $\frac{1}{2\pi} \int_{-\pi}^{\pi} |W(\omega)|^2 d\omega = 1$ as well.

Therefore, the integral is trapped between $\min S(\omega)$ and $\max S(\omega)$.

Correlation and PSD for WSS Processes

Therefore, for all M:

$$\min S(\omega) \leq \lambda_{\min}(R_M) \leq \lambda_{\max}(R_M) \leq \max S(\omega)$$

This bounds the condition number for all R_M:

$$\chi(R_M) \leq \frac{\max(S(\omega))}{\min(S(\omega))}$$

As $M \to \infty$, we can concentrate $|W(\omega)|^2$ more tightly at the points where $S(\omega)$ achieves its maximum, or minimum, so we expect $\chi(R_M)$ to converge to this bound.

Szegö's Theorem

The relationship between R_M and $S(\omega)$, and in particular since each $\lambda(R_M)$ has the form $\mathbf{w}^H R_M \mathbf{w}$ when $\mathbf{w}$ is a corresponding unit eigenvector, suggests intuitively the following result called *Szegö's Theorem:*

If $g(\lambda)$ is a "nice function" for $\{\lambda \geq 0\}$, and with $\left\{\lambda_m^{(M)}\right\}_{1 \leq m \leq M}$ the eigenvalues of R_M:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} g\left(\lambda_m^{(M)}\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(S(\omega))\, d\omega$$

Theorem:

$$\lim_{M\to\infty} \log\left(\det\left(R_M\right)\right)^{1/M} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S\left(\omega\right) d\omega$$

Proof:

$$\left(\det R_M\right)^{1/M} = \left(\prod_{m=1}^{M} \lambda_m^{(M)}\right)^{1/M}$$

is the geometric mean of the eigenvalues.

Entropy

Taking log:

$$\log\left(\det\left(R_M\right)\right)^{1/M} = \frac{1}{M}\sum_{m=1}^{M}\ln\lambda_m^{(M)}$$

and we then apply Szegö's theorem.

We shall use this later in the context of linear prediction.

Random Mix of Deterministic Signals

Let us consider another model for a type of random signal. Let $\{\mathbf{s}_\ell\}_{1\leq\ell\leq L} \subset \mathbb{C}^M$ be L *deterministic* M-dimensional vectors, and the following is thus $M \times L$:

$$S = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_L \end{bmatrix}$$

Let $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_L]^T$ be a *random vector* and:

$$\mathbf{x} = S\boldsymbol{\alpha} = \sum_{\ell=1}^{L} \alpha_\ell^* \mathbf{s}_\ell$$

In other words, $\mathbf{x}$ is a random linear combination of deterministic signals. With R_α the correlation matrix of $\boldsymbol{\alpha}$ we get:

$$R_x = SR_\alpha S^H$$

Random Mix of Deterministic Signals

As a special case, if $\boldsymbol{\alpha}$ are 0-mean uncorrelated with $\sigma_\ell^2 = var\left(\alpha_\ell\right)$:

$$R_x = \sum_{\ell=1}^{L} \sigma_\ell^2 \mathbf{s}_\ell \mathbf{s}_\ell^H$$

If the $\mathbf{s}_\ell$'s are linearly independent (this requires, in particular, $L \leq M$), then R_x has rank L. However, unless the $\mathbf{s}_\ell$'s are orthogonal, they are *not* the eigenvectors of R_x, and σ_ℓ^2 are not the eigenvalues.

Random Mix of Sinewaves

A special case is a sinewave:

$$\mathbf{s}(\omega) = \begin{bmatrix} 1 \\ e^{j\omega} \\ \vdots \\ e^{j(M-1)\omega} \end{bmatrix}$$

Theorem: Given any $\{\omega_\ell\}_{1\leq\ell\leq L}$ with $L \leq M$ and $|\omega_i - \omega_j| \neq$ integer multiple of 2π for any $i \neq j$, the collection $\{\mathbf{s}(\omega_\ell)\}_{1\leq\ell\leq L}$ are linearly independent.

OPTIMAL FILTERING
II: LINEAR REGRESSION AND CORRELATION MATRICES

ECE416 ADAPTIVE ALGORITHMS