17. Best Linear Unbiased Estimators (BLUE)

ECE 830, Spring 2014

Linear Estimation

Linear estimators are an important class of estimators because of their

- simplicity
- dependence on the first and second order moments only
- ease of implementation

This last point is especially important for filtering problems where we must process data in real-time.

Best Linear Unbiased Estimation

The MVUE is often not computable

- CRLB or Rao-Blackwell not applicable
- intractable mathematical model
- randomness is only known up to first and second order moments

In such cases, we must be content with a suboptimal estimator.
Our approach is to compute the BLUE :

Definition: Best Linear Unbiased Estimator (BLUE)

The BLUE is the linear estimator of the form

$$\widehat{\theta}(x) = Ax, \quad A \in \mathbb{R}^{p \times N}$$

with smallest variance among all linear, unbiased estimators.

Note that for $\widehat{\theta}$ to be unbiased we must have

$$\theta = \mathbb{E}[\widehat{\theta}] =$$

Therefore the mean of the data must obey a linear relationship with the true parameter. This relationship will not always be true (e.g. $\mathbb{E}[x_n] = \cos\theta$), so even the BLUE isn't always feasible. However, there is an important class of problems where it does hold.

Linear Models

Suppose x and θ are related through

$$x = H\theta + W$$

where

- θ is fixed but unknown
- H is $N \times p$ and known, full rank
- W is random with $\mathbb{E}[W] = 0$ and $C := \mathbb{E}[WW^T]$ is known and positive definite but not necessarily Gaussian

Can you think of an A such that $A\mathbb{E}[x] = \theta$?

Taking $A =$, we find

$$A\mathbb{E}[x] =$$

So the pseudoinverse is an unbiased estimator. But is its variance minimal?

The BLUE

The BLUE for Linear Models:

If the data are of the general linear model form

$$x = H\theta + w$$

where H is a known $N \times p$ matrix, rank p, θ is $p \times 1$ parameter, and w is an $N \times 1$ zero-mean noise vector with covariance C (the pdf of w is otherwise unspecified or unknown), then the BLUE of θ is

$$\widehat{\theta} = (H^T C^{-1} H) H^T C^{-1} x$$

the (minimum) variance of $\widehat{\theta}_i$ is $\mathrm{Var}(\widehat{\theta}_i) = [(H^T C^{-1} H)^{-1}]_{ii}$, and the covariance of $\widehat{\theta}$ is

$$C_{\widehat{\theta}} = (H^T C^{-1} H)^{-1}$$

Proof

The unbiased constraint means

$$\theta = \mathbb{E}[Ax] = AH\theta + A\mathbb{E}[W] = AH\theta \quad \text{or} \quad AH = I_{p\times p}$$

The vector variance is

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}) &= \mathbb{E}[(\widehat{\theta}-\theta)^T(\widehat{\theta}-\theta)] = \sum_{i=1}^{p}\mathbb{E}(\widehat{\theta}_i-\theta_i)^2 \\
&= \sum_{i=1}^{p}\mathbb{E}[((Ax)_i-\theta_i)^2] = \sum_{i=1}^{p}\mathbb{E}[((AH\theta+AW)_i-\theta_i)^2] \\
&= \sum_{i=1}^{p}\mathbb{E}[((\theta+AW)_i-\theta_i)^2] = \sum_{i=1}^{p}\mathbb{E}[((AW)_i)^2] \\
&= \sum_{i=1}^{p}\mathbb{E}[a_i^T WW^T a_i] = \sum_{i=1}^{p} a_i^T C a_i
\end{aligned}$$

where $a = [a_1 \cdots a_p]^T$.

Optimization:

$$\begin{aligned} \text{minimize}_A \quad & \sum_{i=1}^{p} a_i^T C a_i \\ \text{subject to} \quad & AH = I \end{aligned}$$

Now the Lagrangian is

$$J = \sum_{i=1}^{p} a_i^T C a_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_j^{(i)} (a_i^T h_j - \delta_{ij})$$

Taking the gradient we get

$$\frac{dJ}{da_i} = 2Ca_i + \sum_{i=1}^{p} \lambda_j^{(i)} h_j = 2Ca_i + H\lambda^{(i)}$$

where $\lambda^{(i)} = [\lambda_1^{(i)} \cdots \lambda_p^{(i)}]^T$. Thus

$$a_i^* = -\frac{1}{2} C^{-1} H \lambda^{(i)}.$$

From the constraint we have

$$
\begin{aligned}
H^T a_i^* &= \boldsymbol{e}_i = [0 \cdots 1 \cdots 0] \\
&= \\
\lambda^{(i)} &= \\
a_i^* &=
\end{aligned}
$$

and so

$$A^* = \left(C^{-1} H (H^T C^{-1} H)^{-1} I_{p \times p}\right)^T = (H^T C^{-1} H)^{-1} H^T C.$$

Connection to linear Gaussian models

Recall, this is precisely the form of the estimator for the linear Gaussian model

$$w \sim \mathcal{N}(0, C)$$

Here our model is more general

$$x = H\theta + \underbrace{w}_{x - \mathbb{E}[x]}$$

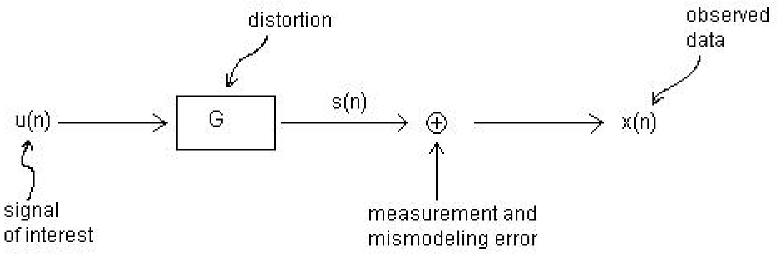
where w is a noise vector with covariance C but is not necessarily Gaussian.

$\Rightarrow$ If w is Gaussian, then BLUE is also MVUE estimator

$\Rightarrow$ If w is non-Gaussian, then BLUE is not necessarily MVUE

Deconvolution Example

Suppose that we want to recover a signal from a noisy and distorted observation.



Assume that

$$
\begin{aligned}
s(n) &= u(n) * g(n) \quad \leftarrow \text{impulse response of } G \\
&= \sum_{k=0}^{M-1} g(k)u(n-k)
\end{aligned}
$$

and assume $u(n) = 0$ for $n < 0$.

Define

$$x = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} \quad \theta = \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(N-1) \end{bmatrix}$$

Distorted signal:

$$x = \underbrace{\begin{bmatrix} g(0) & 0 & \cdots & \cdots & 0 \\ g(1) & g(0) & 0 & \cdots & \\ g(2) & g(1) & g(0) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & g(M-1) & \cdots & 0 \\ 0 & \cdots & 0 & g(M-1) & g(0) \end{bmatrix}}_{H_{N\times N}} \underbrace{\begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(N-1) \end{bmatrix}}_{\theta_{N\times 1}} + w$$

LS Estimator:

$$\widehat{\theta} = (H^\top H)^{-1} H^\top x \qquad \text{or, more generally} \qquad \widehat{\theta} = (H^\mathsf{H} H)^{-1} H^\mathsf{H} x$$

Example:

Impulse response length $M = 2$
Signal/observation dimension $N = 5$

$$H = \begin{bmatrix} g(0) & 0 & 0 & 0 & 0 \\ g(1) & g(0) & 0 & 0 & 0 \\ 0 & g(1) & g(0) & 0 & 0 \\ 0 & 0 & g(1) & g(0) & 0 \\ 0 & 0 & 0 & g(1) & g(0) \end{bmatrix}$$

Notice that

$$H \approx H_c = \begin{bmatrix} g(0) & 0 & 0 & 0 & g(1) \\ g(1) & g(0) & 0 & 0 & 0 \\ 0 & g(1) & g(0) & 0 & 0 \\ 0 & 0 & g(1) & g(0) & 0 \\ 0 & 0 & 0 & g(1) & g(0) \end{bmatrix}$$

H_c is a circulant matrix.

Example: (cont.)

What do we know about circulant matrices? They are diagonalized by the Fourier transform! That is,

$$H_c = U^{\mathsf{H}} D U,$$

where U denotes the Fourier transform matrix, U^{H} is the inverse Fourier transform, and D is a diagonal matrix.

$$\begin{aligned}
\Rightarrow \widehat{\theta} &= (H^{\mathsf{H}} H)^{-1} H^{\mathsf{H}} x \\
&\approx (H_c^{\mathsf{H}} H_c)^{-1} H_c^{\mathsf{H}} x \\
&= \\
&=
\end{aligned}$$

$$\begin{aligned}
\widehat{\theta} &\approx \\
&=
\end{aligned}$$

Example: (cont.)

1. Compute DFT of observation x $(DFT \equiv U)$
2. Weight each DFT coefficient by $\frac{1}{d_{ii}}$, where

$$D = \begin{bmatrix} d_{0,0} & & 0 \\ & \ddots & \\ 0 & & d_{N-1,N-1} \end{bmatrix}$$

3. Compute inverse DFT of result

This is simply a frequency domain filtering operation. The DFT and inverse DFT can be computed in operations using the FFT algorithm. Frequency domain weighting requires operations. Compare to complexity required for matrix inversion of $(H^{\mathsf{H}}H)^{-1}$.