TRATAMIENTO DIGITAL DE SEÑALES
Ingeniería de Telecomunicación (4º, 2º c)

Unidad 6ª: Estimación y regresión lineales

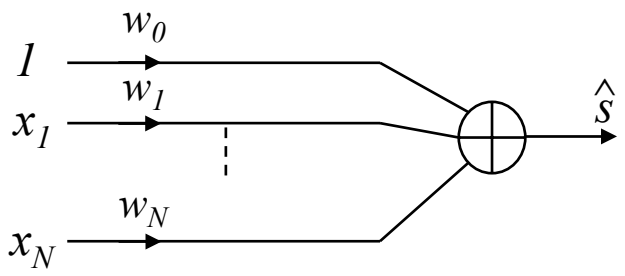
Aníbal R. Figueiras Vidal
Jesús Cid Sueiro
Ángel Navia Vázquez

Área de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

A: (Estimación Lineal Cuadrático Medio Mínima)

Como en el caso de los decisores, los estimadores lineales proporcionan ventajas por su sencillez y facilidad de interpretación, así que es frecuente recurrir a ellos. (Además, son estrictamente óptimos en algunas situaciones, como para el Problema General Gaussiano).

Supóngase que la va s se estima a partir de las N vas $\boldsymbol{x}$ *imponiendo la forma lineal* $\hat{s} = w_0 + \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}_e^T \boldsymbol{x}_e \quad \left(\boldsymbol{x}_e = \left[1 \; \boldsymbol{x}^T \right]^T \right)$



Determínese $\hat{\boldsymbol{w}}_e : \min_{\boldsymbol{w}_e} E\left\{ \left(s - \hat{s} \right)^2 \right\} = \min_{\boldsymbol{w}_e} E\left\{ e^2 \right\}$

$$e^2 = \left(s - \boldsymbol{w}_e^T \boldsymbol{x}_e\right)^2; \quad E\left\{e^2\right\} = E\left\{\left(s - \boldsymbol{w}_e^T \boldsymbol{x}_e\right)^2\right\}$$

$$\frac{\partial E\left\{e^2\right\}}{\partial w_n} = E\left\{\frac{\partial e^2}{\partial w_n}\right\} = E\left\{\frac{\partial e^2}{\partial e}\frac{\partial e}{\partial w_n}\right\} = 2E\left\{e\frac{\partial\left(s - \boldsymbol{w}_e^T \boldsymbol{x}_e\right)}{\partial w_n}\right\}$$

que igualaremos a 0

- *para* w_0*, resulta:* $E\{e\}\big|_{\boldsymbol{w}_e = \hat{\boldsymbol{w}}_e} = 0$

$$E\{s\} - \hat{w}_0 - \hat{\boldsymbol{w}}^T E\{\boldsymbol{x}\} = 0$$

$$\hat{w}_0 = E\{s\} - \hat{\boldsymbol{w}}^T E\{\boldsymbol{x}\}$$

- *para los demás pesos:* $E\{ex_n\}\big|_{\boldsymbol{w} = \hat{\boldsymbol{w}}_e} = 0$

en forma bloque: $E\left\{\hat{e}\,\boldsymbol{x}\right\} = \boldsymbol{0} = E\left\{\left(s - \hat{\boldsymbol{w}}_e^T \boldsymbol{x}_e\right)\boldsymbol{x}\right\}$

Principio de Ortogonalidad: $\hat{e} \perp x_i, \quad i = 1,\ldots,N$

(también en versión extendida)

(se comprueba fácilmente que es un mínimo, al ser la matriz de derivadas segundas un producto externo).

El Principio de Ortogonalidad para la formulación extendida conduce a:

$$\begin{array}{lllll}
\hat{w}_0 & +\hat{w}_1 E\{x_1\} & +\hat{w}_2 E\{x_2\}+ \cdots & +\hat{w}_N E\{x_N\} & = E\{s\} \\
\hat{w}_0 E\{x_1\} & +\hat{w}_1 E\{x_1 x_1\} & +\hat{w}_2 E\{x_2 x_1\}+ \cdots & +\hat{w}_N E\{x_N x_1\} & = E\{s x_1\} \\
\hat{w}_0 E\{x_2\} & +\hat{w}_1 E\{x_1 x_2\} & +\hat{w}_2 E\{x_2 x_2\}+ \cdots & +\hat{w}_N E\{x_N x_2\} & = E\{s x_2\} \\
\vdots & & & & \vdots \\
\hat{w}_0 E\{x_N\} & +\hat{w}_1 E\{x_1 x_N\} & +\hat{w}_2 E\{x_2 x_N\}+ \cdots & +\hat{w}_N E\{x_N x_N\} & = E\{s x_N\}
\end{array}$$

que puede reescribirse utilizando las covarianzas v : $E\{x_i x_j\} = r_{x_i x_j} =$

$= v_{x_i x_j} + E\{x_i\} E\{x_j\}$, *y análogamente para* $E\{s x_i\}$;

aparece entonces en la ecuación j+1

- *un sumando de la forma* $\left(\hat{w}_0 + \hat{w}_1 E\{x_1\} + \hat{w}_2 E\{x_2\} + \ldots + \hat{w}_N E\{x_N\}\right) E\{x_j\}$ *en el término de la izquierda;*
- *un sumando de la forma* $E\{s\} E\{x_j\}$ *en el de la derecha;*

que, al ser iguales (por la primera ecuación), pueden eliminarse; y debe procederse así, puesto que se eliminan términos de ***colinealidad*** *en las ecuaciones, y con ello se alivian los problemas numéricos en la solución.*

Así, se resolverá por un lado la primera ecuación, y, por otro, las llamadas ***Ecuaciones Normales*** *(N ecuaciones con N incógnitas)*

$$\begin{aligned}
\hat{w}_1 v_{x_1x_1} + \hat{w}_2 v_{x_2x_1} + \cdots + \hat{w}_N v_{x_Nx_1} &= v_{sx_1} \\
\hat{w}_1 v_{x_1x_2} + \hat{w}_2 v_{x_2x_2} + \cdots + \hat{w}_N v_{x_Nx_2} &= v_{sx_2} \\
\vdots & \quad \vdots \\
\hat{w}_1 v_{x_1x_N} + \hat{w}_2 v_{x_2x_N} + \cdots + \hat{w}_N v_{x_Nx_N} &= v_{sx_N}
\end{aligned}$$

(que son las únicas a resolver si $\boldsymbol{x}$ *y s tienen medias nulas).*

En forma matricial

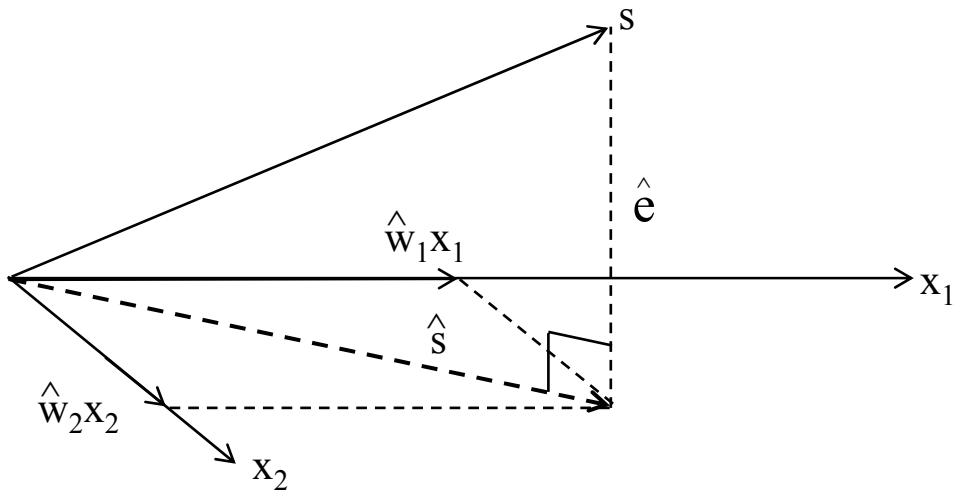
$$\boldsymbol{V}_{xx}\,\hat{\boldsymbol{w}} = \boldsymbol{v}_{sx} \qquad \textit{(nótese: } \left[\boldsymbol{V}_{xx}\right]_{i,j} = v_{x_jx_i} \textit{)}$$

cuya solución es

$$\hat{\boldsymbol{w}} = \boldsymbol{V}_{xx}^{-1}\boldsymbol{v}_{sx}$$

Interpretación

El principio de Ortogonalidad tiene una interpretación obvia considerando que las va se pueden representar como vectores con módulo igual a la raíz cuadrada de su varianza (medias nulas):



el tamaño del error es mínimo si s se proyecta ortogonalmente sobre el subespacio generado por las **x**; y el estimador lineal buscado no es otra cosa que la **proyección resultante**, ŝ. Por ello, se habla también de **Teorema de Proyección**.

Debe notarse que $E\{\hat{e}\hat{s}\} = E\{\hat{e}\hat{\mathbf{w}}^T\mathbf{x}\} = 0$

Ejercicios de discusión

D: ¿Cuál es la contribución a la reducción del error de cada variable x_i? (Prescíndase de las medias).

En la misma figura que se ha utilizado para ilustrar el Principio de Ortogonalidad se observa que la intervención de x_1 en el valor de $\hat{e}$ (y de $\hat{s}$)... depende de cómo sea x_2.

Analíticamente:

$$E\left\{\hat{e}^2\right\} = E\left\{\left(s - \hat{\boldsymbol{w}}^T \boldsymbol{x}\right)\hat{e}\right\} = E\left\{s\,\hat{e}\right\} = E\left\{s\left(s - \hat{\boldsymbol{w}}^T \boldsymbol{x}\right)\right\} = E\left\{s^2\right\} - \hat{\boldsymbol{w}}^T \boldsymbol{v}_{sx}$$

que es

$$E\left\{\hat{e}^2\right\} = E\left\{s^2\right\} - \boldsymbol{v}_{sx}^T \boldsymbol{V}_{xx}^{-1} \boldsymbol{v}_{sx}$$

donde queda de manifiesto lo dicho: y se percibe que la covarianza entre s y x_i no indica directamente la importancia de x_i para reducir el error (salvo que se trate de elegir una variable de entre dos).

Por tanto, si se tratase de seleccionar variables, no se puede hacer de acuerdo con sus covarianzas individuales con respecto a s. Tampoco de acuerdo con sus coeficientes $\hat{\boldsymbol{w}}$*: ya que sus valores dependen de todas las demás variables.*

Naturalmente, en el caso de que las $\boldsymbol{x}$ *sean* ***incorrelacionadas*** *(varianzas cruzadas nulas),* $\boldsymbol{V}_{xx}$ *se diagonaliza, y resulta manifiesto que cada variable influye de acuerdo con* $v_{x_i x_i}^{-1} v_{sx_i}^2 \left(w_i v_{sx_i} \right)$.

Y también es obvio que, si se trata de añadir una nueva variable a un estimador lineal ***fijado****, ha de elegirse la más correlacionada con el* ***error****.*

El problema de selección de variables es importante, como se sabe: debe observarse que la utilización de variables linealmente dependientes convierte $\boldsymbol{V}_{xx}$ *en singular.*

T: Métodos de selección de variables para estimación lms

D: Las variables dato se pueden "ortogonalizar" (medias nulas) de acuerdo con el procedimiento de Gram-Schmidt.

a) Formule algorítmicamente el procedimiento.

b) ¿Es el procedimiento útil para selección de variables?

a) Supuesto que se ha fijado un orden $x_1, \ldots, x_N$, se trata de ir creando variables que sean ortogonales a las anteriores: para ello se restan las partes no ortogonales

$$x'_1 = x_1$$

$$x'_2 = x_2 - \frac{E\{x'_1\, x_2\}}{E\{x'_1\, x'_1\}} x'_1$$

$$x'_3 = x_3 - \frac{E\{x'_1\, x_3\}}{E\{x'_1\, x'_1\}} x'_1 - \frac{E\{x'_2\, x_2\}}{E\{x'_2\, x'_2\}} x'_2$$

$$\vdots$$

es decir :

$$x'_1 = x_1$$

$$a_{ln} = \frac{E\{x'_l\, x_n\}}{E\{x'_l\, x'_l\}}, \qquad 1 \le l \le n$$

$$x'_n = x_n - \sum_{l=1}^{n-1} a_{ln} x'_l, \qquad 2 \le n \le N$$

(hay otras versiones algorítmicas de mejores características numéricas)

b) Está claro que es inmediata la selección de las x': pero ello no lleva implícita una selección de las x, ya que en las x' seleccionadas pueden entrar cualesquiera combinaciones de las variables originales.

(Sin embargo, sí es cierto que la aplicación de este proceso evita sufrir los efectos de colinealidades entre variables).

No es solución ordenar la ortogonalización según la correlación de los datos con s, ya que no se puede prever el efecto que tiene el proceso sobre dichas variables (ni siquiera es óptimo elegir en cada paso la variable que dé lugar a la variable transformada de mayor efecto).

Pero también es verdad que el efecto favorable de evitar colinealidades permite obtener buenos resultados aplicando estos procedimientos subóptimos: que se conocen como métodos de ***Ortogonalización de Mínimos Cuadrados*** *(OLS, "Orthogonal Least Squares") (el nombre se refiere, estrictamente, a su aplicación en forma* ***muestral****).*

D: ¿Cómo variaría la formulación expuesta si se tratase con vas complejas?

Considerando que hay que manejar $E\{|e|^2\} = E\{ee^*\}$, se llega de inmediato a ver que hay que sustituir $E\{x_i x_j\}$ y $E\{sx_i\}$ por $E\{x_i x_j^*\}$ y $E\{sx_i^*\}$.

Conviene notar que la transposición se convierte en transposición hermítica: y que esta operación conjuga V_{xx}.

*D: ¿Cómo hay que extender la formulación si se estimase una variable multidimensional, **s**?*

Bastaría proceder fila a fila para encontrar como solución

$$\hat{\boldsymbol{s}} = \hat{\boldsymbol{W}}^T \boldsymbol{x}$$

siendo

$$\hat{\boldsymbol{W}} = \boldsymbol{V}_{xx}^{-1} \boldsymbol{V}_{sx}$$

(nótese que $\boldsymbol{V}_{sx} = E\{\boldsymbol{x}\boldsymbol{s}^T\}$)

Versiones analíticas y máquina

Si la física del problema es conocida, V_{xx} y v_{sx} se pueden obtener de dicho conocimiento: se trataría entonces de un aplicación **analítica**.

Si no es así, tendrían que estimarse V_{xx} y v_{sx} a partir de observaciones **muestrales** $\{s^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^{K}$: en cuyo caso se hablaría de una formulación **máquina**.

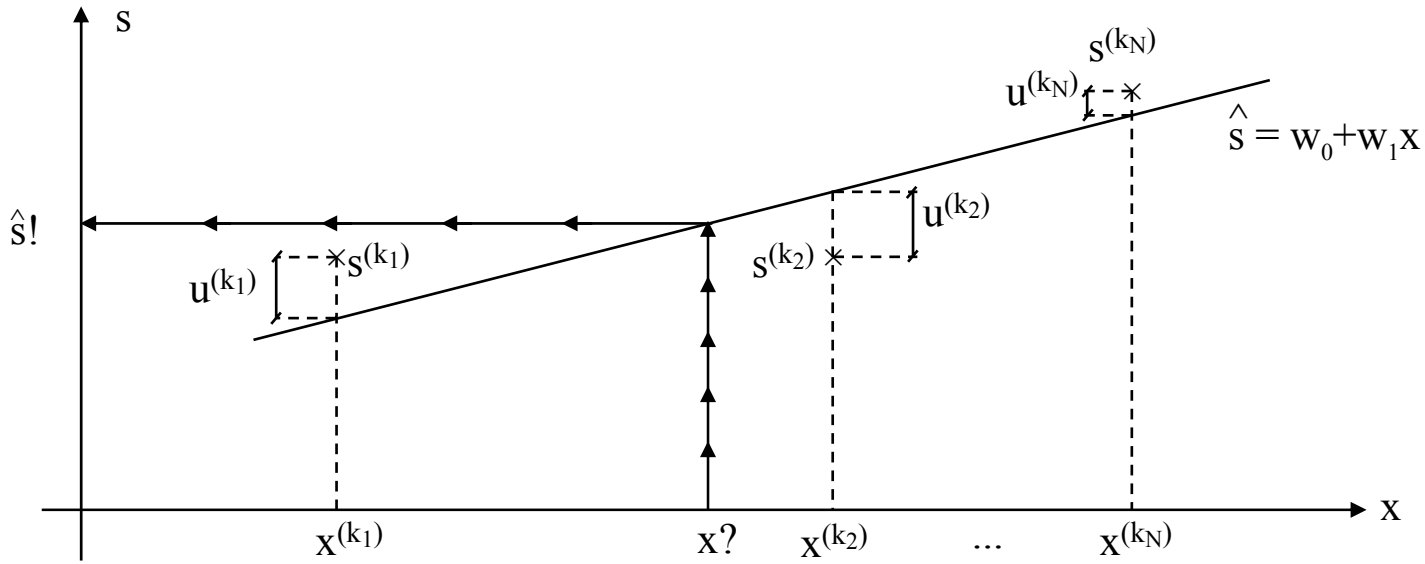
Desde ahora se hace notar que, para comprobar la coherencia de lo que se está suponiendo (el modelo físico, lo razonable de emplear la estimación lms), conviene, cuando sea posible, verificar que los resultados son estables para varias aproximaciones: analítica y máquina si hay modelo físico y datos disponibles, o entre varias aproximaciones máquina.

La Regresión Lineal

Este problema se formula análogamente a la estimación de la media de una va gaussiana: se admite que cada observación de s es una combinación lineal de las variables deterministas $\{x_n\}$ y una componente aleatoria, llamada **residuo** o **innovación**, u, que se supone gaussiana de media cero: además, con la misma varianza para cualesquiera valores de las $\{x_n\}$ (condición de **homocedasticidad**).

La va. se estima **por su media** (lo que minimiza el error cuadrático medio de la estimación); y ésta se obtiene a partir de un conjunto de observaciones $\{s^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^{K}$ (típicamente, K>>N), ajustando los parámetros libres vía **mínimos cuadrados** (LS).

De modo que la Regresión Lineal pretende lo que se ilustra en la figura: para cada valor de x (**x**), establecer qué parte de s se debe al mismo, minimizando el efecto de u.



El ajuste LS de los parámetros minimiza el error cuadrático total observado: suma de los cuadrados de

$$u^{(k)} = s^{(k)} - \mathbf{w}_e^T \mathbf{x}_e^{(k)} \quad , \qquad k = 1,...,K$$

es decir, el módulo del vector

$$\bar{u} = \bar{s} - \mathbf{X}_e \mathbf{w}_e$$

con

$$\mathbf{X}_e = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)} \\ \vdots & & & & \\ 1 & x_1^{(K)} & x_2^{(K)} & \cdots & x_N^{(K)} \end{bmatrix} \qquad (K \times (N+1))$$

lo que no es otra cosa que aproximar $\bar{s}$ mediante una combinación lineal de los (N+1) vectores columna de $\mathbf{X}_e$, minimizando el error cuadrático: se aplicará el Principio de Ortogonalidad

$$\hat{\bar{u}} \perp \text{col}\{\mathbf{X}_e\}$$

es decir, en forma bloque

$$\mathbf{X}_e^T \hat{\bar{u}} = \mathbf{0}$$

o sea

$$\mathbf{X}_e^T \bar{s} - \mathbf{X}_e^T \mathbf{X}_e \hat{\mathbf{w}}_e = \mathbf{0}$$

de donde, supuesta $\mathbf{X}_e^T \mathbf{X}_e$ invertible,

$$\hat{\mathbf{w}}_e = \left(\mathbf{X}_e^T \mathbf{X}_e\right)^{-1} \mathbf{X}_e^T \bar{s}$$

La matriz factor de $\bar{s}$ se denomina **seudoinversa de Moore-Penrose** de $\mathbf{X}_e$, $\mathbf{X}_e^{\#}$ y es la solución del problema LS que aquí ha surgido. Proyecta $\bar{s}$ sobre el subespacio formado por las columnas de $\mathbf{X}_e$.

Discusión

Q: ¿Cómo procedería en una situación de heterocedasticidad, en que u tiene una varianza dependiente de ***x****?*

La solución pasaría por convertir el problema en homocedástico: equidistribuyendo u.

Si la forma de la heterocedasticidad es conocida, se puede aplicar en la formulación, y se llega a una solución LS ponderada; si no es así, han de realizarse aproximaciones sucesivas asumiendo homocedasticidad local, y estimando las varianzas a partir de los residuos.

Regresión Lineal vs Estimación Lineal

Cabe notar que las expresiones obtenidas para las soluciones son análogas:

* $\mathbf{X}_e^T \mathbf{X}_e$ hace el papel de $\mathbf{V}_{xx}$

* $\mathbf{X}_e^T \bar{s}$ hace el papel de v_{sx}

y, de hecho, es inmediato comprobar que la solución de la Estimación Lineal se hace formalmente idéntica a la de la Regresión si se sigue la vía muestral y se estiman correlaciones de forma muestral para las ecuaciones extendidas.

Pero el planteamiento es conceptualmente distinto:

* en E. L., las **x** son **va**

* en R.L., las **x** son **deterministas**

teniendo, por tanto, diferentes situaciones de aplicación.

Si cabe la opción de modelar las **x** como deterministas o como aleatorias, está claro que la segunda opción será tanto más ventajosa cuanto **más conocimiento** estadístico se tenga de las **x**; la primera, cuanto menos: por ejemplo, cuando los valores de las **x** puedan ser "controladas" externamente.

Modelos semilineales

Reciben este nombre todos los modelos construidos

- transformando no linealmente las variables **x** en otras variables **y**
- utilizando para la estimación/regresión una combinación lineal de las **y**;

con lo que, fijada la transformación, podemos proceder con las **y** de los modos vistos: y se tendrán métodos **no lineales** de diseño sencillo, cuya eficacia dependerá de lo adecuado de la **transformación** para el problema bajo estudio.

Son muy típicos los métodos semilineales basados en un previo **agrupamiento** de las muestras $\mathbf{x}^{(k)}$, empleando después como $y^{(k)}$ las similitudes de cada $\mathbf{x}^{(k)}$ con los representantes de los grupos, $\mathbf{m}_j$. Estudiaremos las técnicas de agrupamiento en la Unidad 10.

También se incluyen aquí los modelos **polinómicos**.

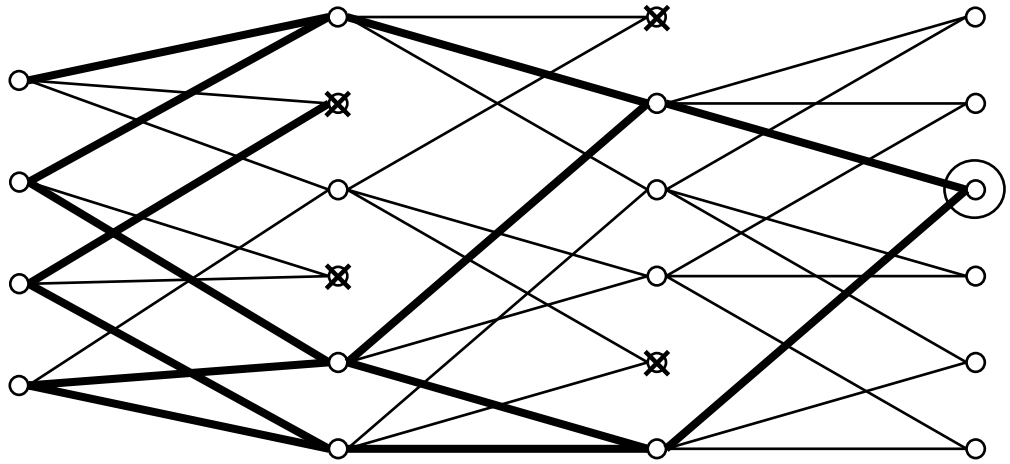
$$\hat{s} = w_o + \sum_{n=1}^{N} w_n x_n + \sum_{n=1}^{N}\sum_{n'=1}^{N} w_{nn'} x_n x_{n'} + \ldots$$

pusto que los términos del polinomio son transformaciones de las $\{x_n\}$. Como se sabe, la dificultad aquí radica en que hay una **explosión dimensional** con el grado del polinomio.

Una técnica (subóptima) para construir modelos polinómicos es el llamado Método de Manejo de Datos en Grupo ("Group Method Data Handling", **GMDH**), debido a Ivakhnenko. Se basa en que multiplicando polinomios se suman sus grados, y que los polinomios de grado 2 en dos variables tienen sólo 6 coeficientes; así,

- se emplean todos los posibles términos de dos variables de orden 2 a partir de las variables (hay $\binom{N+1}{2}$ términos así);
- se seleccionan los N "mejores";
- se itera el proceso, construyendo polinomios de dos variables y orden 2 a partir de las salidas del paso anterior (el orden sube a 4, 6, 8, ...), hasta que se saturan los resultados;
- Se elige como solución la correspondiente a la mejor salida del último paso, y se procede a un diseño final.

Se puede ilustrar el proceso como sigue:



El GMDH tiene mecanismos que aseguran una buena generalización (entrena iterativamente con una parte de las muestras y verifica sobre el resto), y hay una gran cantidad de variantes.