

Estimación PLS de modelos SEM, 2015. UMA.

**Modelos de Ecuaciones Estructurales con
PLS (Partial Least Squares)**

5 de febrero de 2015

José Mondéjar Jiménez

**Área de Estadística Económica y Empresarial
Universidad de Castilla-La Mancha**

Presentación

EXPOSICIÓN:

1. Origen de los SEM.
2. Lógica interna de los modelos SEM y PLS.
3. Escalas de medida.

APLICACIONES PRÁCTICAS:

1. Modelos reflexivos y formativos.
2. Fiabilidad y validez del modelo de medida.
3. Medidas de ajuste del modelo estructural.
4. Bootstrapping, Blindfolding y FIMIX.

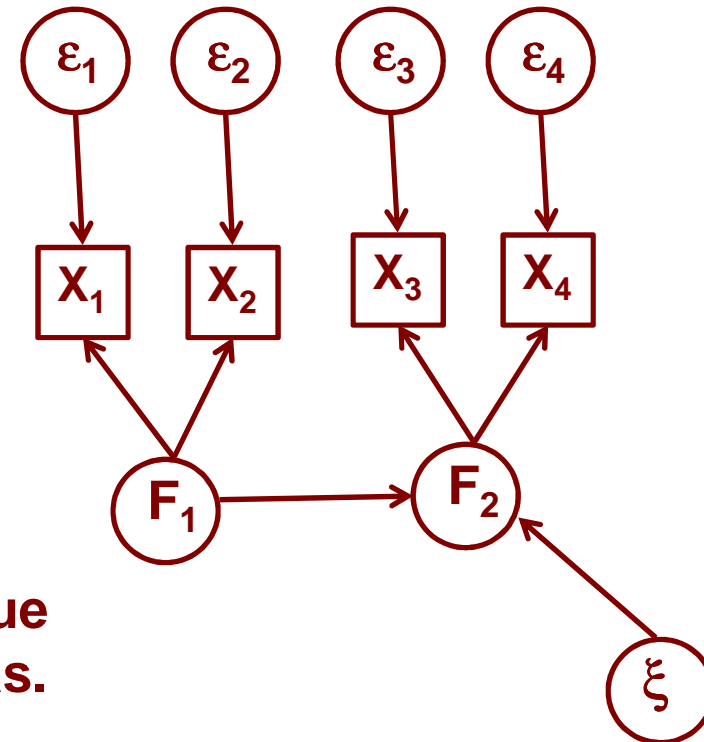
- Los modelos de ecuaciones estructurales pretenden modelar las relaciones estadísticas entre constructos latentes que no son directamente observables.
- En 1960, Karl Jöreskog diseñó un algoritmo de máxima verosimilitud para la estimación de modelos de estructuras de covarianzas, y en 1970, desarrolló el primer software de amplia difusión: LISREL.
- Su director de tesis, Herman Wold, criticó la dependencia de los supuestos distribucionales, que afecta a la validez de los resultados empíricos. Por ello, propuso un enfoque alternativo, Partial Least Squares (PLS), y en 1977 desarrolló el algoritmo NIPALS (Nonlinear Iterative Partial Least Squares).
- Ambos enfoques han seguido desarrollándose y, en la actualidad, son instrumentos cuantitativos complementarios de medición, evaluación y contraste ampliamente utilizados.

2

Lógica interna de los SEM

- Supongamos la matriz de correlación de cuatro variables, que dependen de dos factores relacionados, tal como indica la figura de la derecha:

	X_1	X_2	X_3	X_4
X_1	1.00			
X_2	.087	1.00		
X_3	.140	0.18	1.00	
X_4	.152	.143	.272	1.00



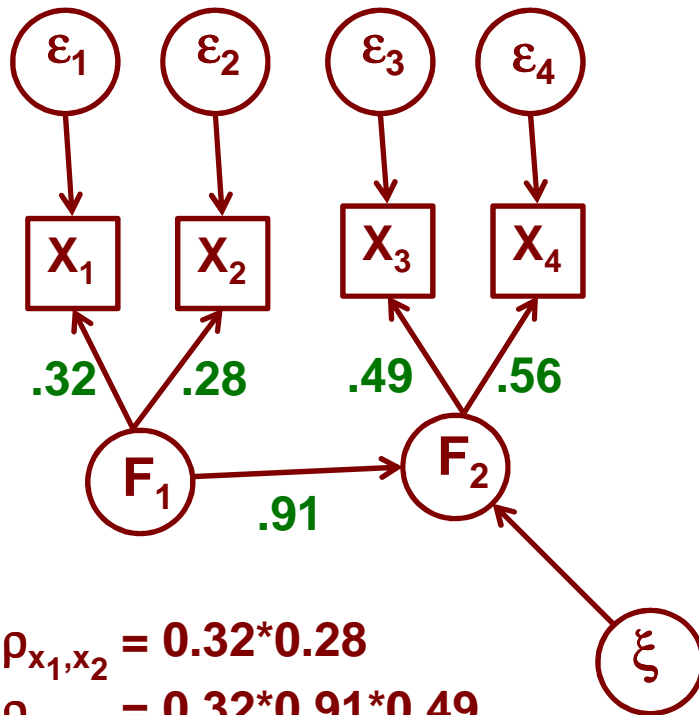
- ε_i son errores de observación.
- F_i son factores no observados, que influyen en las variables observadas.
- ξ es un error de ajuste.

¿Podemos estimar los F_i y todas las relaciones para “reproducir” la matriz de correlación?

2

Lógica interna de los SEM

- **Posible solución:**



$$\begin{aligned} \rho_{x_1, x_2} &= 0.32 * 0.28 \\ \rho_{x_1, x_3} &= 0.32 * 0.91 * 0.49 \\ \rho_{x_1, x_4} &= 0.32 * 0.91 * 0.56 \\ \rho_{x_2, x_3} &= 0.28 * 0.91 * 0.49 \\ \rho_{x_2, x_4} &= 0.28 * 0.91 * 0.56 \\ \rho_{x_3, x_4} &= 0.49 * 0.56 \end{aligned}$$

	X ₁	X ₂	X ₃	X ₄
X ₁	1.00			
X ₂	.087	1.00		
X ₃	.140	0.18	1.00	
X ₄	.152	.143	.272	1.00

	X ₁	X ₂	X ₃
X ₂	.0896 (.0026)		
X ₃	.1427 (.0027)	.1249 (.0551)	
X ₄	.1631 (.0111)	.1427 (.0003)	.2744 (.0024)

Error absoluto medio: 0.0124

¿Por qué dos metodologías? Básicamente por los supuestos y los objetivos buscados:

- A. La metodología SEM está basada en el supuesto de normalidad de las variables y utiliza una estimación máximo verosímil.
 1. Ajusta las varianzas-covarianzas entre las variables observadas.
 2. Necesita una muestra generalmente grande.
 3. Está enfocada a la “reproducción” de la estructura de relaciones entre variables.

¿Por qué dos metodologías? Básicamente por los supuestos y los objetivos buscados:

B. La metodología PLS no asume la normalidad y estima por mínimos cuadrados de forma recursiva.

1. Utiliza un esquema de regresión con supuestos “débiles”. Válido con escalas numéricas, ordinales e incluso nominales (con precaución).
2. Aplicable incluso con muestras pequeñas.
3. Enfocada hacia la predicción, generando combinaciones lineales de indicadores que maximicen la covarianza.

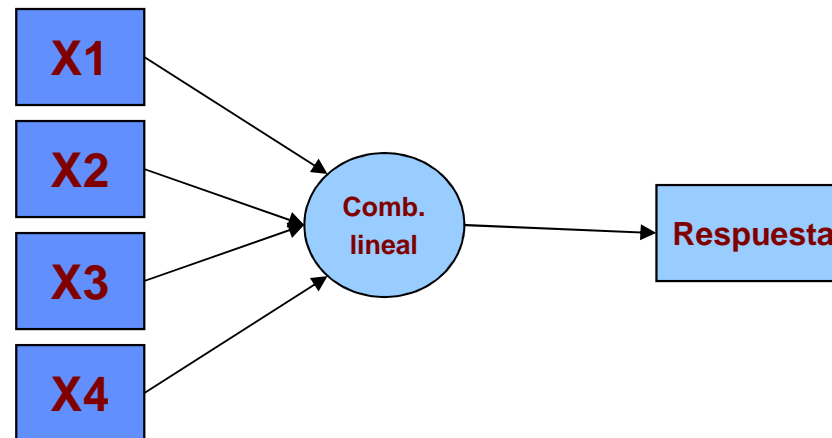
2

Lógica interna de la metodología PLS

ANTECEDENTES:

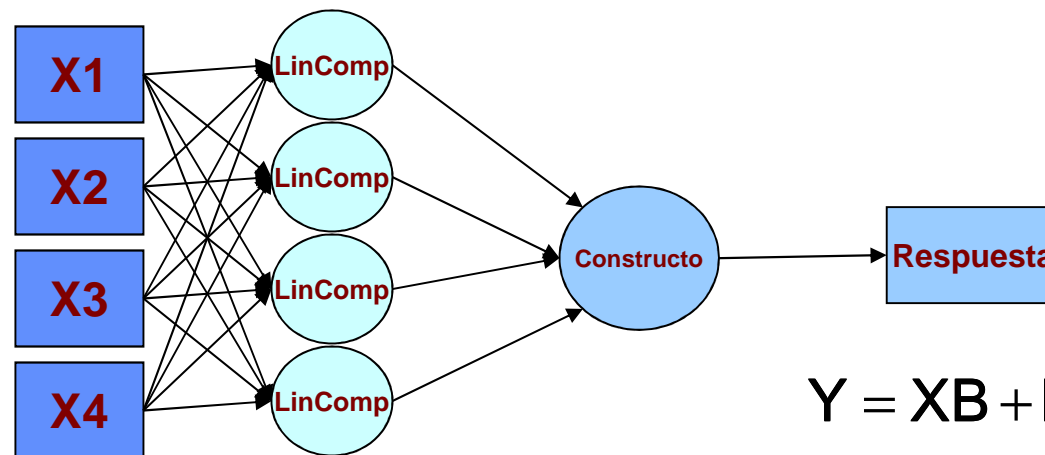
Regresión Múltiple:

$$Y = XB + E$$



Regresión por Componentes Principales:

$$T = XW \quad Y = TQ + E$$



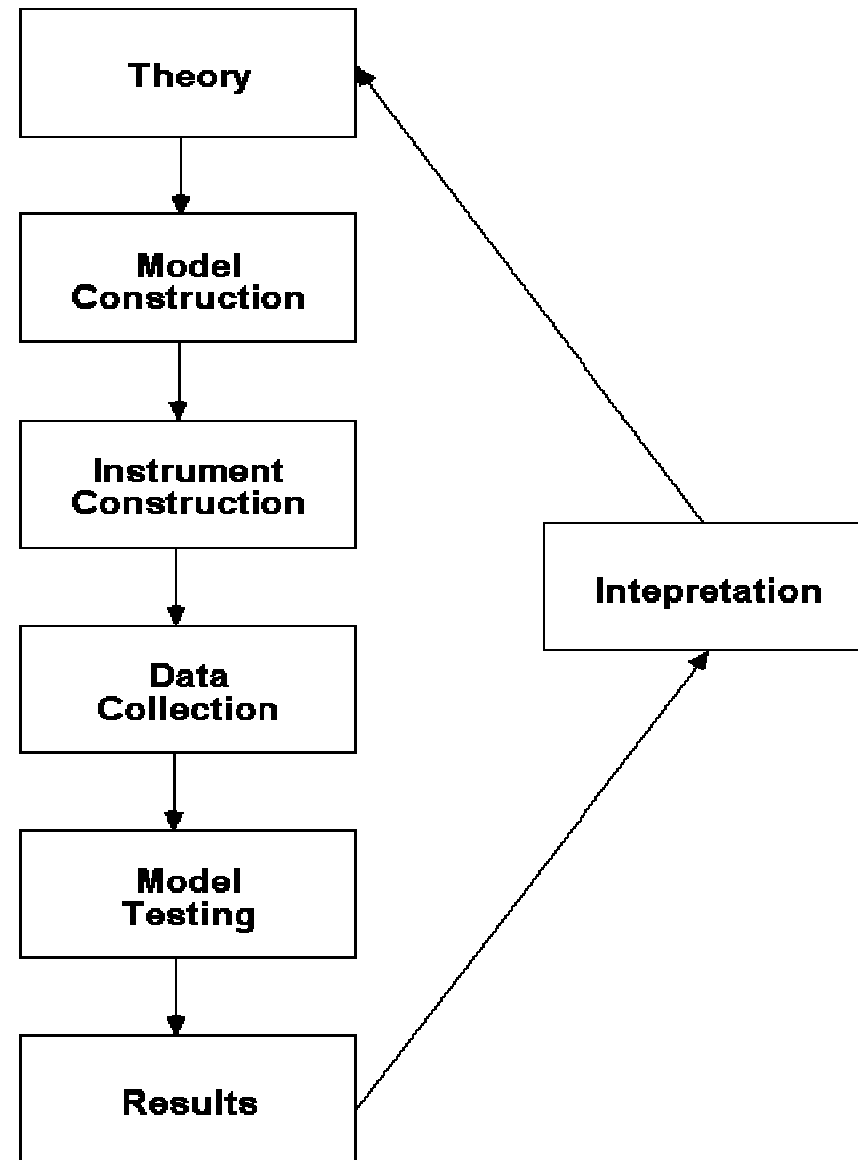
$$Y = XB + E \quad \text{donde } B = WQ$$

1. **NIPALS (Nonlinear Iterative Partial Least Squares)**
2. **PLS Regression (Partial Least Squares Regression)**
3. **PLS Discriminant Analysis**
4. **SIMCA (Soft Independent Modeling by Class Analogy)**
5. **PLS Approach to Structural Equation Modeling**
6. **N-way PLS**
7. **PLS Logistic Regression**
8. **PLS Generalized Linear Model**

2

Lógica interna de la metodología PLS

DIAGRAMA DE PASOS:



1. **Especificación del modelo teórico.** *Cuidado: Un modelo puede estar teóricamente bien especificado pero ser estadística o empíricamente no identificable por efecto de los datos.*
2. **Contrastar el ajuste del modelo de medida.**
3. **Re-especificar y re-estimar el modelo si los datos indican que es necesario.**
4. **Cuando el modelo de medida tenga un buen ajuste, contrastar el modelo estructural.**
5. **Re-especificar y reestimar el modelo estructural, si fuese necesario.**

1. La interpretación “clásica” de la regresión está referida a variables medidas en escalas continuas.
2. Cuando se utilizan escalas discretas (tipo Likert) se debe tener en cuenta la limitación de la varianza.
3. Este tipo de escalas se adaptan mejor a la metodología PLS que las escalas Guttman o las escalas Mokken, por lo que se ha generalizado su uso.
4. Cuando la escala es de pocos puntos, se debe cuidar especialmente la interpretación del estadístico alfa de Cronbach y la estructura de correlación.

**Veamos algunos ejemplos
prácticos
(learning by doing)**

2

Fiabilidad y validez del modelo de medida

FIABILIDAD: consistencia entre los diversos indicadores de cada variable latente, indicando unidimensionalidad de la misma. En escalas Likert, se suele utilizar la correlación entre el ítem y el total de la escala.

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2} \right) = \frac{k\rho}{1 + (k-1)\rho}$$

FIABILIDAD COMPUESTA: no es individual para cada constructo, sino que considera el modelo en su conjunto. Sería recomendable analizarla después de comprobar las cargas cruzadas.

$$IFC_i = \frac{(\sum_j L_{ij})^2}{(\sum_j L_{ij})^2 + \sum_j (1 - L_{ij}^2)}$$

VARIANZA EXTRAÍDA MEDIA (AVE): Proporción de varianza del constructo “explicada” por los indicadores.

$$AVE_i = \frac{\sum_j L_{ij}^2}{\sum_j L_{ij}^2 + \sum_j (1 - L_{ij}^2)}$$

VALIDEZ: de contenido (teórica) y de construcción, tanto convergente (tamaño de las cargas L_{ij} y AVE, cargas cruzadas) y divergente:

$$\min(AVE_i, AVE_j) > \rho_{ij}^2$$

La fiabilidad está relacionada con la consistencia interna de las escalas, es decir, que todos los indicadores “miden la misma dimensión”. La validez hace referencia a que se “mide lo que se quiere medir”, por lo que se debe evaluar en los tres ámbitos comentados: contenido, convergente (capacidad de captar aspectos del mismo constructo) y divergente (discrimina entre constructos, de forma que no hay indicadores “generales” para todos).

Tras comprobar que el modelo de medida está correctamente especificado (y sólo tras este paso), podemos evaluar el modelo estructural.

3

Medidas de ajuste del modelo estructural

Se analizan, básicamente, tres aspectos: el grado de ajuste, la significación de la estimación de los parámetros y la capacidad predictiva del modelo.

GRADO DE AJUSTE: se mide mediante el coeficiente R^2 de Nagelkerke, de interpretación similar al de Pearson.

$$R_N^2 = \frac{1 - \left[\frac{L(M_{intercep})}{L(M_{full})} \right]^{\frac{N}{2}}}{1 - L(M_{intercep})^{\frac{N}{2}}}$$

SIGNIFICACIÓN DE LOS PARÁMETROS: al disponer sólo de una muestra para estimar los parámetros, es preciso simular otras para poder estimar la variabilidad de las estimaciones y construir un estadístico t-Student.

BOOTSTRAPPING

CAPACIDAD PREDICTIVA: capacidad del modelo para “reconstruir” las observaciones iniciales a partir de la información contenida en el resto de la muestra.

BLINDFOLDING

SEGMENTACIÓN DE LA MUESTRA: en Ciencias Sociales es difícil garantizar que toda la muestra se ajusta a la misma distribución probabilística. Con PLS, es posible abordar la segmentación en función del modelo estructural, es decir, agrupando a los individuos de la muestra de forma que presenten parámetros lo más diferentes posibles.

FIMIX

Heterogeneidad no observable: FIMIX

En la práctica, es usual encontrar cierto grado de heterogeneidad, que puede invalidar los resultados. Cuando se utilizan variables latentes, se puede producir una heterogeneidad no observable en los indicadores, sino que se produce en las relaciones entre constructos.

En PLS, existen distintos algoritmos para tratarla:

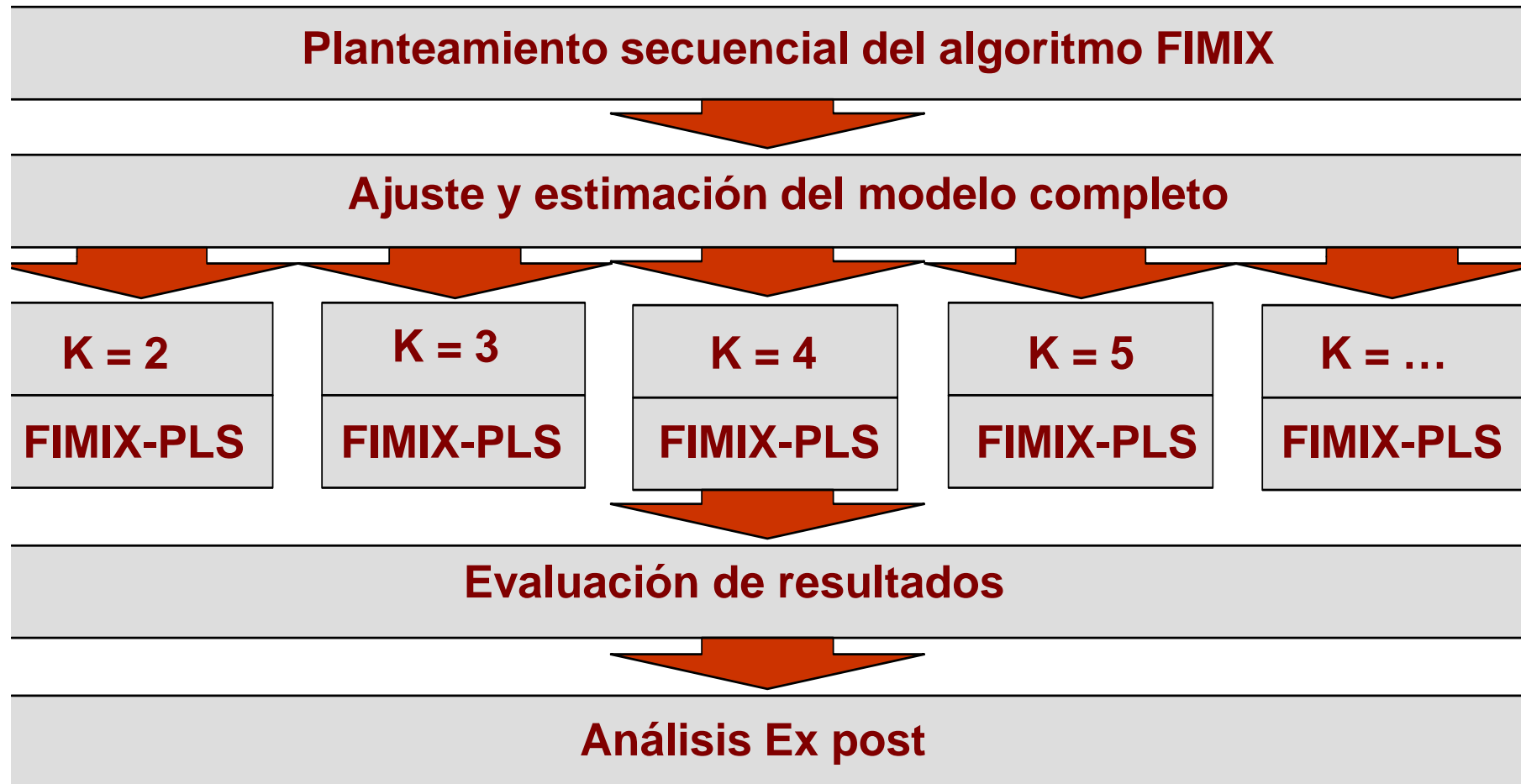
- PLS-TPM
- REBUS-PLS
- FIMIX-PLS

La heterogeneidad no observable afecta a los resultados del modelo:

- Resultados difíciles de explicar, no esperados
- Ajuste pobre (R^2 , AVE, IFC, etc.)
- Coeficientes no significativos por el efecto de la composición

El enfoque FIMIX-PLS es una técnica de segmentación basada en el modelo ajustado, que produce grupos en función de las respuestas obtenidas (y los valores de los parámetros estimados)

Heterogeneidad no observable: FIMIX



Heterogeneidad no observable: FIMIX

La decisión de cuántos grupos se eligen está basada en medidas heurísticas de ajuste (LnL, AIC, BIC, CAIC o Entropía).

La entropía estadística está acotada entre 0 y 1 y “mide” el grado de “desorden” de la distribución muestral:

$$EN_k = 1 - \frac{\sum \sum -P_{ij} \ln(P_{ij})}{\ln(K)}$$

El algoritmo parte de dos grupos y se va aumentando la división hasta que la entropía comienza a descender.

VENTAJAS:

- Es robusto a la distribución de los datos muestrales.
- Obtiene buenos resultados en cuanto a tamaño de grupos y separación.
- Tiende a optimizar el valor de la entropía.

DESVENTAJAS:

- No identifica heterogeneidad en el modelo de medida.
- Es sensible a extremos locales en la distribución.

Heterogeneidad no observable: FIMIX

EJEMPLOS:

Ejemplo 1: Motivaciones del turista en Cuenca.

Ejemplo 2: Eco-emprendedores turísticos.