UNIVERSIDAD DE JAÉN

# Material del curso "Análisis de datos procedentes de investigaciones mediante programas informáticos"

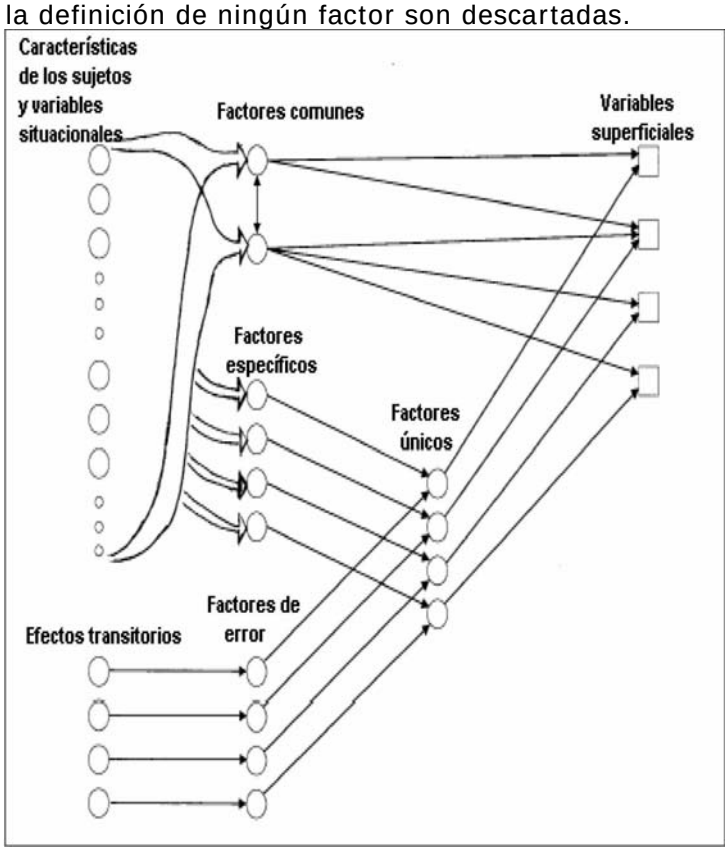
Manuel Miguel Ramos Álvarez

**Índice**

## MATERIAL XIV "DESCRIPCIÓN CON ANÁLISIS FACTORIALES"

# 1. Conceptos fundamentales del análisis factorial y de componentes principales

- Objetivo: reducir un conjunto de variables observadas o medidas –las **superficiales**- a un conjunto menor de variables subyacentes –las **latentes**-.
- Lógica científica. Que los aspectos del comportamiento que se manifiestan en las variables superficiales, se entienden bien cuando se recurre a atributos latentes o constructos, o factores.
- Lógica estadística. La existencia de factores explica, en parte, la variabilidad (varianza) de las variables superficiales y las correlaciones entre ellas.
    - Los individuos puntúan de forma bastante diferente unos de otros, y algunos atributos correlacionan alto entre sí, mientras que entre otros la correlación es baja. Estos patrones de variación-correlación son aprovechados para identificar los factores subyacentes.
- Conclusión: técnicas orientadas al descubrimiento de las variables subyacentes a un conjunto de v. superf., agrupando las que correlacionan alto entre sí en un factor, puesto que dependerán de una misma variable latente. Además, las v. superf. que no contribuyan a la definición de ningún factor son descartadas.




- Tipos de Factores:
    - **Comunes.** Los que influyen sobre más de una v. superf.
    - **Específicos.** Los que influencian solamente una de las v. superf.
    - Los comunes son los responsables de la correlación entre v. superf. mientras que ambos los son de las varianzas.
    - Asociados a **Errores** de medida. Por características transitorias de los sujetos o de las situaciones de prueba (parte de la fiabilidad pruebas).
    - Se combinan los específicos con de error de medida para obtener factores **únicos, cada uno de los cuales** influye en una sola v. superf.

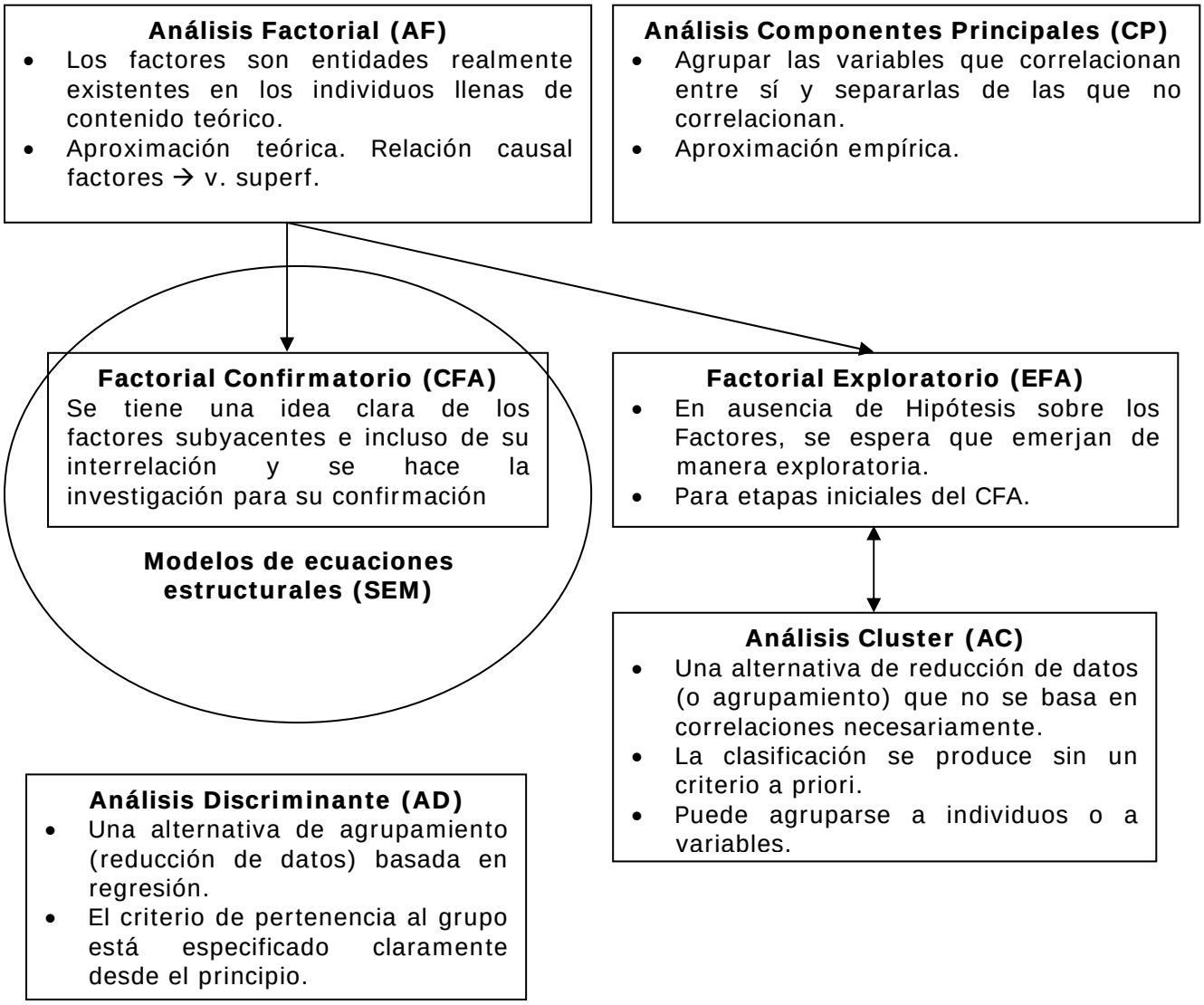
# 2. Planteamiento computacional del análisis factorial y de componentes principales

- La puntuación observada para un sujeto depende de la influencia de ambos, comunes y únicos (específicos + error).

$$\underbrace{Y_{ij}}_{\substack{\text{Puntuac.suj. i}\\ \text{v.superf. j}}} = \underbrace{\underbrace{F_{1,i}}_{\substack{\text{Carga}\\ \text{Factorial}}} \cdot \underbrace{a_{i,1}}_{\substack{\text{Puntuac.suj.}\\ \text{cada Factor}}}}_{p-\textit{factores comunes}} + F_{2,i} \cdot a_{i,2} + \ldots + F_{p,i} \cdot a_{i,p} + \underbrace{U_i}_{\text{factor único}}$$

- Por tanto, parte de la varianza de una variable superficial es debida a factores comunes, es la varianza común o **comunalidad**, y parte es debida al factor único, la varianza única.
- La correlación entre variables sólo puede ser debida a los factores comunes, puesto que los factores únicos afectan solamente a una variable. En otras palabras, dos variables correlacionan si comparten, son influenciadas por, un factor común.
- Objetivo de cálculos es Minimizar la matriz Residual (o diferencia entre la Matriz de Correlaciones observadas y Predicha a partir de los factores).
- Los factores comunes pueden estar a su vez relacionados entre sí (i.e. habilidad numérica espacial); lo que determinará el algoritmo de cómputo.

# 3. Técnicas relacionadas



# 4. El Análisis Factorial y de Componentes Principales

1. Identificar el **dominio** en el cual se inserta la investigación de cara a una adecuada interpretación. Llegar a una idea al menos tentativa sobre cuáles podrían ser los factores que convendría estudiar. Intentar incluir un buen número de factores (5-6) para buscar completitud y relevancia (ver validez de constructo en Ramos, Catena y Trujillo, 2004).

Analizar→Reducción de datos →Análisis factorial → Variables: y1_i1_s0,...

2. **Planificación y estructuración del diseño:**
   a. **Seleccionar** un buen número de variables superficiales que van a ser medidas a un elevado número de sujetos (300)
   b. Más de una variable por cada factor hipotético.
   c. Intentar incluir variables **marcadoras** (las que sepamos con cierto grado de seguridad que dependen del factor) y evitar las que puedan correlacionar con más de un factor.
   d. Seleccionar la muestra de sujetos intentando maximizar su heterogeneidad.
   e. Si hay muestras diferentes de sujetos, o las pruebas se han pasado en momentos temporales diferentes (i.e. test-restest) mejor realizar el análisis factorial por separado y comparar la solución factorial alcanzada en cada una de ellas.
3. Se realiza la investigación y se obtienen los datos.
4. Se computa la **matriz de correlaciones** observada -entre todos los pares de variables superficiales medidas- y se realiza un examen de la misma.
   a. Decidir si es pertinente el análisis factorial, mediante índices que permitan saber si hay correlaciones altas en la matriz que permitan extraer factores.
      i. Test de esfericidad de **Bartlett** a partir del determinante de la matriz. Sí correlaciones (y factibilidad del factorial) si es significativo.
      ii. Prueba de Kaiser-Meyer-Olkin (**KMO**) y su complementario basado en la correlación anti-imagen, más completo pues considera tanto las correlaciones como las correlaciones parciales. Sí correlaciones (y factibilidad del factorial) cuando KMO > 0,60 y proporción baja de anti-imágenes con valores elevados de correlación (la mayoría por encima de 0.60).
      iii. Mejor segunda opción ya que la primera está sesgada por el tamaño muestral.

Analizar→Reducción de datos →Análisis factorial → Variables: y1_i1_s0,... Descriptivos: Determinante, KMO...,Anti-imagen→ Continuar →Aceptar.

5. Se realiza la **extracción** de factores.
   a. Variantes:
      i. **Componentes principales (CP)**. Se buscan factores que sean ortogonales. Reducir la matriz de correlaciones a un conjunto menor de componentes principales (factores); diagonalizándola (matriz ortogonal), obteniendo sus autovectores y autovalores. Entonces las cargas factoriales se obtienen multiplicando cada uno de los elementos del autovector por su autovalor. **El autovalor** asociado al factor es justamente la suma de cuadrados de sus cargas y una medida pues de su importancia para explicar la variabilidad de las var. superf. Las **cargas** indican la importancia del factor en cada variable, y pueden obtenerse correlacionado factores y variables superficiales. Se basa en el algoritmo iterativo de Hotelling (1933): los componentes se van extrayendo de manera sucesiva, la mayor proporción de variabilidad es explicada por el primero, después por el segundo, y así sucesivamente.

Analizar→Reducción de datos →Análisis factorial Variables: y1_i1_s0,...→ Extracción→Método: Componentes principales → Analizar: matriz de correlaciones →Extraer: Autovalores mayores que 1 →Mostrar: Solución factorial sin rotar. Gráfico de sedimentación → Continuar→Aceptar.

      ii. **Factores principales.** Método recursivo que mejora el de componentes principales pues considera también las comunalidades en las estimaciones, de manera que los

...-> Extracción -> Método: Ejes Principales -> ...

factores se obtiene por convergencia sucesiva hasta que las comunalidades ya no cambian significativamente.

iii. **Mínimos cuadrados.** Evita el uso la diagonal positiva de manera más tajante que la anterior. Los factores se extraen minimizando la suma de los cuadrados de las diferencias entre la matriz predicha y la matriz original. Una variante del mismo (generalizada) permite ponderar las correlaciones según las variables con varianza compartida con las demás.

iv. **Máxima verosimilitud.** Trata de extraer factores de forma sucesiva de manera que cada uno explique tanta varianza como sea posible de una matriz poblacional, no de la muestral. Su eficiencia disminuye notablemente cuando el número de sujetos es bajo, por lo que en este caso es recomendable usar otros métodos.

v. **Método alfa.** Enfatiza la generalización (dada por el coeficiente de fiabilidad o consistencia interna alfa) de los resultados a un universo de contenido en una batería que consideran una muestra no aleatoria de ese universo. No está claro si la ganancia en fiabilidad resulta relevante respecto de la extracción de factores.

vi. **Método de Imagen.** Es un complemento de los anteriores, partiendo de una matriz de covarianzas imagen, que pretende la predicción de la parte común (la imagen parcial) de una variable (con otras) desde las demás variables mediante procedimientos de regresión múltiple. El problema más importante de esta aproximación reside en cómo interpretar las cargas factoriales imagen, puesto que no son correlaciones de las variables con los factores.

vii. Los métodos alternativos al de CP se concentran más en la extracción de factores comunes y separan su contribución de los factores únicos (factores específicos + factores de error) incorporando de manera expresa el error (complementario de las comunalidades). Luego, el AF podría ganar más con los otros métodos no CP cuando se tiene a priori una idea clara de la significación de las entidades hipotéticas.

b. Calibrar **el número de factores** que es necesario retener según los resultados de cómputo y de la interpretación de los mismos así como de su **importancia (calidad)**.

i. Dependerá de las cargas (o pesos) de cada variable en cada factor y del patrón de correlaciones que observemos (i.e. si hay dos patrones entonces dos factores).

ii. De la cantidad de variabilidad que se puede explicar. Los métodos de extracción nos los proporcionan de manera jerarquizada según peso o importancia. Si por ejemplo omitimos factores con menor carga, ¿Cambiamos de manera significativa el porcentaje de variabilidad?

iii. En términos generales, descartar factores con un **autovalor inferior a la unidad** y complementar con el método "**scree plot**" (o gráfico de sedimentación), buscando en él el Punto de Inflexión.

iv. Comparar diferentes soluciones para distintos algoritmos o según distinto número de factores mediante las **comunalidades** y poder así calibrar cuál es el que permite explicar mejor todas las variables en su conjunto. La comunalidad ($h^2$) es la varianza de cada variable explicada por los componentes principales o los factores comunes (y el error es su complementario, 1- $h^2$).

6. **Rotar los factores** con el objetivo de facilitar su interpretación (casi siempre necesario).
    a. La rotación no altera la estructura de la solución (i.e. cantidad de varianza explicada), sino solamente la cercanía de cada variable superficial a cada factor. Es como cambiar nuestro punto de vista sobre los mismos para facilitar la interpretación de los factores.
    b. Debido a que los factores iniciales con más carga no son puros y se "comen" parte de la varianza de los inferiores en la jerarquía de extracción. La rotación consigue obtener los pesos esperados de cada factor extraído.
    c. **Lógica**: Una transformación de la matriz de cargas factoriales original de forma que los ejes factoriales (cada uno representa un factor) se aproximen lo máximo posible a las variables en las que tienen alta saturación (alto peso). Se realizan por parejas de factores, con más de dos se requieren varias iteraciones.
    d. Para seleccionar una matriz de transformación usar el principio de estructura simple (principio parsimonia):
        - Cada factor cargas altas y próximas a cero.
        - Cada variable debe ser explicada por un solo factor.
        - Factores diferentes deben tener distribución de cargas distinta.
    e. Variantes:
        i. **Ortogonal**. Cuando los factores no se espera que estén correlacionados. Ambos ejes son rotados el mismo número de grados en la misma dirección.
            - **Varimax**. De Kaiser (1958). *max*imización de la *var*ianza de los factores, lograr la mayor diversidad posible en el patrón de cargas en cada factor. Se hace respecto a las cargas a través de la variables para cada factor. Reduce (simplifica) el número de variables.
            - **Cuartimax**. Se centra sólo en una parte de la Varianza. Se hace respecto a las cargas a través de los factores para cada variable (un poco al contrario que el anterior). Reduce (simplifica) el número de factores.
            - **Bicuartimax**. Se aplican simultáneamente los dos tipos anteriores.
            - **Ecuamax y Ortomax**. Como el anterior pero de manera que hay una ponderación, varimax se pondera como el nº factores dividido por 2. simplifica la interpretación debido a que reduce el número de variables que saturan alto en un factor y reduce también el número de factores necesarios para explicar una variable.
            - **Normalizados**. Cargas normalizadas al dividir por la Raíz de la comunalidades.
        ii. **Oblicua** para factores que correlacionan. Los ejes se rotan de forma individual, con ángulos y direcciones específicas para cada eje. Proporcionan dos matrices de pesos que tienen uso diferente.
            - **La de estructura**, que indica la correlación entre factores y variables. La de **configuración** indica la aportación de cada factor a la v. superf. y más bien para calcular las puntuaciones factoriales de los sujetos.
            - Variantes: **Oblimin directo** (mínima la suma productos de los cuadrados de los coeficientes de la matriz de configuración) y **Promax** (maximizar la razón entre las cargas altas bajas). En ambos casos se

Analizar→Reducción de datos →Análisis factorial Variables: y1_i1_s0,...→ Extracción→Método: Componentes principales → Analizar: matriz de correlaciones →Extraer: Autovalores mayores que 1 →Mostrar: Solución factorial sin rotar. Gráfico de sedimentación →Continuar →Rotación →Método: Varimax -> Mostrar: Solución factorial rotada; Gráficos de saturaciones -> Continuar -> Aceptar (análisis).

...→Rotación →Método: Oblimin Directo → Delta: 0 →...

puede controlar el ángulo y por ende el grado de oblicuidad (parámetros Delta y Kappa).

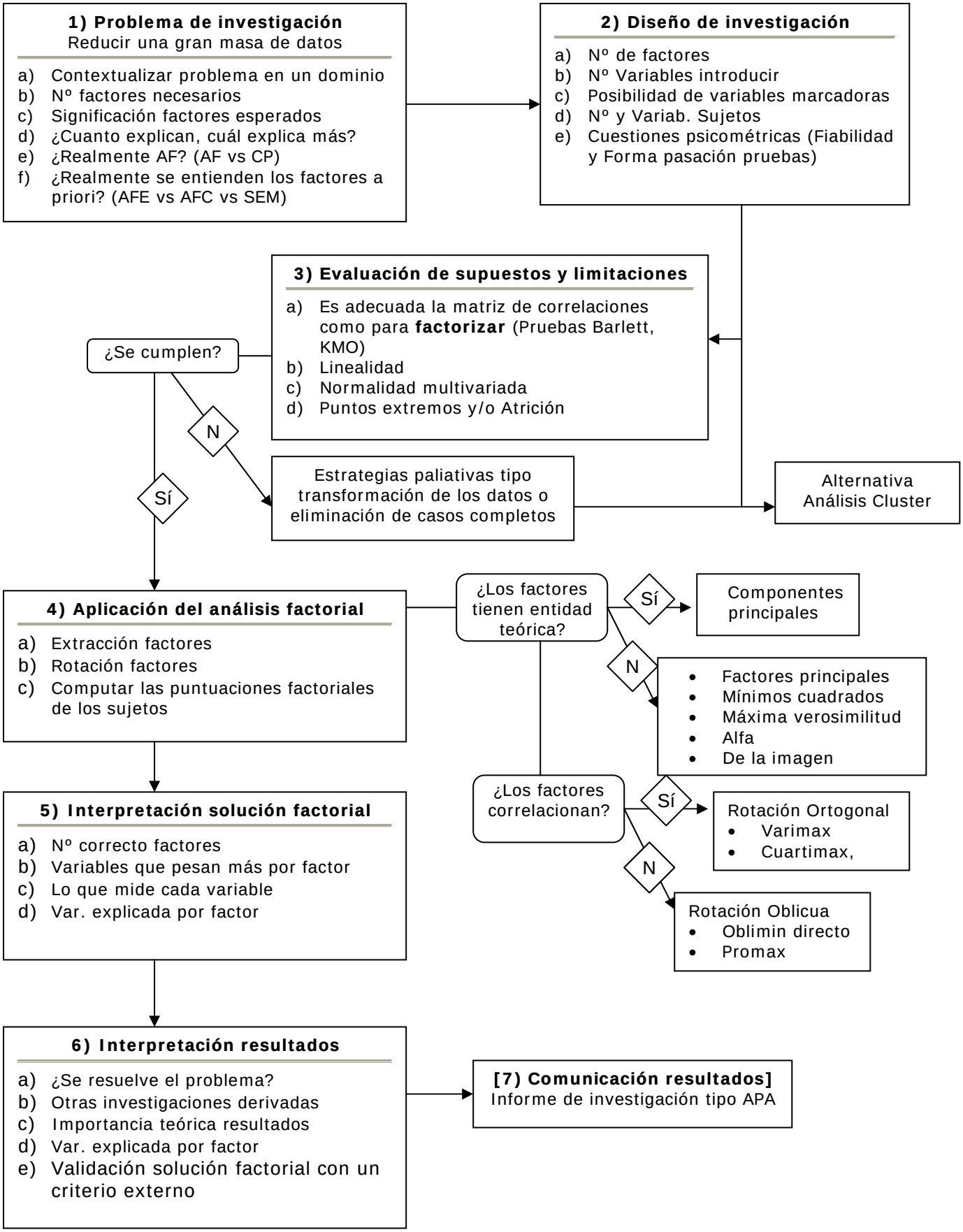
iii. Varimax la más utilizada y las oblicuas las que menos por su dificultad de interpretación. Probar primero la oblicua y si las correlaciones entre los factores son menores de 0.35, optar por la ortogonal. En caso contrario, el tipo Oblimin directo. Otros programas incluyen algoritmos de compromiso para aislar la parte común de los factores frente a la parte independiente de los mismos (i.e. Statistica).

7. **Interpretación** de la solución obtenida. La estimación última viene dada por las matrices de pesos y sobre éstos es donde se interpretarían de manera teórica los factores en caso oportuno. Las cargas factoriales de las variables son el indicador más adecuado para realizar la interpretación. En el caso de una rotación oblicua, es mejor interpretar los pesos de configuración que los de estructura, puesto que los últimos incluyen interacción entre factores.
8. Estimación de los coeficientes de cargas factoriales para obtener las **puntuaciones factoriales** de cada sujeto por ejemplo para compararlos entre sí. Fase opcional. Es una mera estimación de regresión y opera en puntuaciones típicas.
   a. Para extraer factores de segundo orden o agrupar los sujetos en categorías mediante procedimientos como el análisis de cluster. E incluso para introducir los factores como predictores en futuros análisis de regresión.
9. La solución factorial puede ser contrastada, validada, con respecto a un criterio externo (i.e. emplear el rendimiento académico para validar una batería de pruebas de habilidades intelectuales).

.....-> Puntuaciones -> Guardar como variables -> ...

- Un elevado número de sujetos asegura la convergencia de los diferentes métodos.
- El análisis factorial es una técnica que corre un riesgo importante, su mal uso, desembocando en soluciones “vacías de contenido conceptual”.

# 5. Secuencia de investigación en el Análisis Factorial y de Componentes Principales



Volver Principio