

SELECCIÓN DEL MODELO POR CRITERIOS DE INFORMACIÓN PARA ANÁLISIS DE COMPONENTES PRINCIPALES OCULTAS DE MARKOV

Model selection using information criteria for hidden Markov principal components analysis

RESUMEN

El modelo de análisis de componentes principales ocultas de Markov tiene en cuenta la dependencia entre observaciones multivariadas, que en el análisis tradicional de componentes principales se omite por completo. Para la aplicación de este modelo dinámico, es necesario conocer la dimensión de los subespacios principales asociados a cada estado de una cadena de Markov y el número de estados en la cadena. En este artículo, se emplean criterios de información para seleccionar los parámetros del modelo y se presenta una estrategia secuencial de selección de parámetros que reduce el costo computacional. Se muestran resultados utilizando bases de datos sintéticas.

PALABRAS CLAVES: análisis de componentes principales, modelos ocultos de Markov, selección del modelo, criterios de información.

ABSTRACT

Hidden Markov principal component analysis incorporates the dependence between multivariate observations, which in a typical principal component analysis is discarded. For the application of this dynamic model, it is necessary to know beforehand, the dimensionality of the principal subspace associated with each state of a Markov chain, also, the number of states of the chain. In this article, information criteria are employed for model selection. A sequential pruning strategy for model selection is presented as well, in order to reduce computational cost. Results using synthetic data are shown.

KEYWORDS: principal component analysis, hidden Markov models, Gaussian model selection, information criteria.

1. INTRODUCCIÓN

El análisis de componentes principales ocultas de Markov (HMPCA) fue introducido en [1, 2] con el objetivo de incorporar la dependencia entre observaciones en el análisis tradicional de componentes principales (PCA). Empleando la formulación de componentes principales probabilísticas (PPCA) [3] como el modelo de observación en un modelo oculto de Markov (HMM) es posible establecer un modelo dinámico de componentes principales, cuyas ventajas son su capacidad de representación de series de datos multivariadas, con variables que estén correlacionadas en el tiempo, además de posibilitar la reducción localizada de características dinámicas. Los parámetros del modelo pueden estimarse fácilmente empleando un algoritmo EM.

Dos preguntas abiertas en este modelo tienen que ver, primero, con la dimensión de los subespacios de cada modelo de PPCA asociado a cada estado y, segundo, el número de estados de la cadena de Markov. Una forma directa de seleccionar los parámetros de un modelo que emplea criterios de máxima verosimilitud es recurrir a criterios de información como el criterio de información de Bayes (BIC) y el criterio de longitud de descripción

mínima (MDL). El objetivo de estos criterios es penalizar la verosimilitud maximizada, pues como es bien sabido, el criterio de máxima verosimilitud es susceptible al sobreentrenamiento, lo que conduce a estimaciones sesgadas cuando se emplea como criterio de selección de parámetros. En [4], se establece una estrategia secuencial para reducir el costo computacional de aplicar los criterios BIC y MDL, en la selección del número de estados en un HMM con funciones de probabilidad de observación gaussianas.

En este artículo se emplean los criterios BIC y MDL para estimar el número de estados de la cadena de Markov y la dimensión de los subespacios principales asociados a los estados, en el modelo de HMPCA. Igualmente, se presenta la modificación de la estrategia secuencial propuesta por [4], para ajustarla a HMPCA.

En la sección 2, se describe el modelo de análisis de componentes principales ocultas de Markov. En la sección 3, se presentan los criterios de selección de parámetros y la estrategia de selección secuencial. En la sección 4, se muestran resultados usando conjuntos de datos sintéticos y finalmente en la sección 5, la discusión y los resultados.

MAURICIO ALVAREZ

Ingeniero Electrónico, M.Sc.
Profesor Auxiliar
Universidad Tecnológica de Pereira
malvarez@utp.edu.co

RICARDO HENAO

Ingeniero Electrónico, M.Sc.
Profesor Auxiliar
Universidad Tecnológica de Pereira
rhenao@utp.edu.co

2. COMPONENTES PRINCIPALES OCULTAS DE MARKOV

Un modelo oculto de Markov es básicamente una cadena de Markov donde las observaciones a la salida son una variable aleatoria generada de acuerdo a una función probabilística asociada con cada estado [5]. Formalmente, un modelo oculto de Markov de Q estados está definido por π , la distribución inicial de los estados con elementos $\pi_k = p(q_0 = k), 1 \leq k \leq Q$; A , la matriz de probabilidades de transición con elementos a_{ij} que denota la probabilidad de hacer una transición del estado k al estado j , esto es $a_{kj} = p(q_n = j | q_{n-1} = k), n = 1, \dots, N$; Θ , los parámetros de una función de densidad de probabilidad de salida $p(\mathbf{t}_n | q_n, \Theta)$ donde $\mathbf{T} = \{\mathbf{t}_n\}_{n=1}^N$, $\mathbf{t}_n \in \mathbb{R}^d$ es la secuencia de observaciones y $\mathbf{q} = \{q_n\}_{n=1}^N$ es la secuencia no observada de estados.

Por simplicidad, el conjunto total de parámetros se denota como $\tilde{\Lambda} = (\tilde{\pi}, \tilde{A}, \tilde{\Theta})$. En un HMM parametrizado por algún vector $\tilde{\Lambda}$, la variable observada \mathbf{t}_n depende solamente del estado actual q_n (y no de los estados previos) y el estado actual a su vez depende sólo del estado anterior (y no de los anteriores a él). Esto permite que la verosimilitud logarítmica conjunta de una secuencia de observaciones \mathbf{T} y una secuencia de variables ocultas (estados) \mathbf{q} se pueda escribir como

$$\log p_{\tilde{\Lambda}}(\mathbf{T}, \mathbf{q}) = \log p_{\tilde{\Lambda}}(q_1) + \sum_{n=2}^N \log p_{\tilde{\Lambda}}(q_n | q_{n-1}) \quad (1)$$

$$+ \sum_{n=1}^N \log p_{\tilde{\Lambda}}(\mathbf{t}_n | q_n)$$

Un procedimiento estándar para maximizar la ecuación (1) emplea el algoritmo de maximización de la esperanza (EM). Este algoritmo maximiza $Q(\tilde{\Lambda}, \Lambda)$, que es la verosimilitud logarítmica conjunta tomada con respecto a la distribución anterior de variables de estado ocultas $p_{\Lambda}(\mathbf{q} | \mathbf{T})$. De este modo,

$$Q(\tilde{\Lambda}, \Lambda) = \sum_{\mathbf{q}} p_{\Lambda}(\mathbf{q} | \mathbf{T}) \log p_{\Lambda}(q_1)$$

$$+ \sum_{\mathbf{q}} p_{\Lambda}(\mathbf{q} | \mathbf{T}) \sum_{n=2}^N \log p_{\Lambda}(q_n | q_{n-1}) \quad (2)$$

$$+ \sum_{\mathbf{q}} p_{\Lambda}(\mathbf{q} | \mathbf{T}) \sum_{n=1}^N \log p_{\Lambda}(\mathbf{t}_n | q_n)$$

Donde $\sum_{\mathbf{q}}$ denota la suma sobre todas las posibles secuencias de estado ocultas. La ecuación (2) consta de tres términos: el primero para los parámetros de la matriz de probabilidad inicial, π , la segunda para los parámetros de la matriz de transición de estados A y la tercera para los parámetros del modelo de observación, Θ ,

$$Q(\tilde{\Lambda}, \Lambda) = Q(\tilde{\pi}, \pi) + Q(\tilde{A}, A) + Q(\tilde{\Theta}, \Theta) \quad (3)$$

Estos términos pueden maximizarse de forma separada dando origen a ecuaciones de actualización para los diferentes parámetros del modelo. El tercer término en la ecuación (3) puede escribirse como

$$Q(\tilde{\Theta}, \Theta) = \sum_i \sum_n p_{\Lambda}(q_n | \mathbf{t}_n) \log p_{\tilde{\Lambda}}(\mathbf{t}_n | q_n)$$

Con el objetivo de obtener $p_{\tilde{\Lambda}}(\mathbf{t}_n | q_n)$, se asume, para el modelo de observación, la variable \mathbf{t} está definida por una transformación lineal de la variable latente \mathbf{x} , de dimensión p más ruido gaussiano, de forma que

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

donde $p(\mathbf{x}) \sim N(0, \mathbf{I})$ y $p(\boldsymbol{\varepsilon}) \sim N(0, \sigma^2 \mathbf{I})$. La distribución condicional de las variables observadas dadas las variables latentes es $p(\mathbf{t} | \mathbf{x}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \sim N(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2 \mathbf{I})$ y la distribución marginal está dada como $p(\mathbf{t} | \boldsymbol{\theta}) \sim N(\boldsymbol{\mu}, \mathbf{C})$, donde $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \sigma^2, \mathbf{W}\}$ y $\mathbf{C} = \mathbf{W}\mathbf{W}^* + \sigma^2 \mathbf{I}$. Para un modelo de mezcla, la distribución marginal se modela como $p(\mathbf{t} | \boldsymbol{\theta}) = \sum_{i=1}^M c_i p(\mathbf{t} | \boldsymbol{\theta}_i)$, donde los coeficientes de mezcla siguen las restricciones $\sum_i c_i = 1$ y $0 \leq c_i \leq 1, \forall i$ y se ha definido $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^M$. Usando este modelo de mezcla como la densidad de observación en el contexto de un HMM, se puede escribir

$$p(\mathbf{t}_n | q_n = k, \boldsymbol{\theta}) = \sum_{i=1}^M c_{ik} p(\mathbf{t}_n | \boldsymbol{\theta}_{ik})$$

$$= \sum_{i=1}^M c_{ik} p(\mathbf{t}_n | \boldsymbol{\mu}_{ik}, \sigma_{ik}^2, \mathbf{W}_{ik})$$

Puede demostrarse que las ecuaciones de actualización para $\boldsymbol{\theta}_{ik}$ y c_{ik} están dados por

$$\bar{\mathbf{W}}_{ik} = \left[\sum_{n=1}^N \gamma_{nik} (\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_{ik}) \langle \mathbf{x}_{nik} \rangle^* \right] \left[\sum_{n=1}^N \gamma_{nik} \langle \mathbf{x}_{nik} \mathbf{x}_{nik}^* \rangle \right]^{-1}$$

$$\bar{\sigma}_{ik}^2 = \left[d \sum_{n=1}^N \gamma_{nik} \right]^{-1} \sum_{n=1}^N \gamma_{nik} \left[\|\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_{ik}\|^2 - 2 \langle \mathbf{x}_{nik} \rangle^* \bar{\mathbf{W}}_{ik} (\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_{ik}) + \text{tr}(\bar{\mathbf{W}}_{ik} \bar{\mathbf{W}}_{ik} \langle \mathbf{x}_{nik} \mathbf{x}_{nik}^* \rangle) \right] \quad (4)$$

donde

$$\boldsymbol{\mu}_{ik} = \left[\sum_{n=1}^N \gamma_{nik} \right]^{-1} \left[\sum_{n=1}^N \gamma_{nik} (\mathbf{t}_n - \bar{\mathbf{W}}_{ik} \langle \mathbf{x}_{nik} \rangle) \right]$$

$$\tilde{c}_{ik} = \frac{1}{N} \sum_{n=1}^N \gamma_{nik}$$

$$\langle \mathbf{x}_{nik} \rangle = \mathbf{M}_{ik}^{-1} \bar{\mathbf{W}}_{ik}^* (\mathbf{t}_n - \boldsymbol{\mu}_{ik})$$

$$\langle \mathbf{x}_{nik} \mathbf{x}_{nik}^* \rangle = \sigma_{ik}^2 \mathbf{M}_{ik}^{-1} + \langle \mathbf{x}_{nik} \rangle \langle \mathbf{x}_{nik} \rangle^*$$

$$\gamma_{nik} = p(q_n = k | \mathbf{T}, \tilde{\Lambda}) (c_{ik} p(\mathbf{t}_n | \boldsymbol{\theta}_{ik})) / \left(\sum_{j=1}^M c_{jk} p(\mathbf{t}_n | \boldsymbol{\theta}_{jk}) \right)$$

$$\mathbf{M}_{ik} = \bar{\mathbf{W}}_{ik}^* \bar{\mathbf{W}}_{ik} + \sigma_{ik}^2 \mathbf{I}$$

La cantidad γ_{nik} puede calcularse eficientemente usando el algoritmo Forward-Backward [6]. Las ecuaciones de actualización para $\tilde{\pi}$ y \tilde{A} son las mismas que en las de un HMM convencional [5]. El modelo anterior se denomina Análisis de Componentes Principales Ocultas

de Markov (HMPCA) y fue introducido por primera vez en [1].

3. SELECCIÓN DEL NÚMERO DE ESTADOS Y LA DIMENSIÓN DEL SUBESPACIO PRINCIPAL

Un problema fundamental cuando se usa HMPCA en aplicaciones prácticas es la determinación del número de estados Q . Adicionalmente, para el modelo de observación de PPCA asociado a cada estado, es necesario conocer la dimensión del subespacio principal p . Para una cadena oculta de Markov es necesario igualmente determinar la topología, por ejemplo, si es un modelo ergódico o izquierda- derecha. La determinación del número de estados, la dimensión del subespacio y la topología de modelo se conoce conjuntamente como el problema de *selección del modelo*. En este artículo, se pretende determinar el número de estados y la dimensión del subespacio. La determinación de la topología depende de la aplicación específica y no se considera en este trabajo.

El criterio de máxima verosimilitud no es apropiado para la selección del número de estados y componentes, debido a que los modelos se encuentran anidados, es decir, un HMPCA con menor número de estados puede verse siempre como un caso particular de un modelo con más estados. Existe otro problema asociado a la identificabilidad del modelo. Se puede demostrar [4] que una HMM con más de una gaussiana por estado y un HMM con varios estados, pero sólo una gaussiana por estado son equivalentes. Debido a lo anterior y reconociendo que PPCA es un modelo gaussiano, se considerarán HMPCAs con un único modelo de PPCA por estado.

Criterio de información de Bayes (BIC)

En el criterio de información de Bayes, la verosimilitud maximizada se penaliza por la complejidad del modelo, medido por el número de parámetros libre en Λ_k , donde k es el número de estados en el modelo. Sea O un conjunto L de secuencias de observación $O = \{\mathbf{T}_l\}_{l=1}^L$ y N_o el número total de observaciones en O , es decir, $N_o = \sum_{l=1}^L N_l$, donde N_l es la longitud de la secuencia de observación l . El número óptimo de estados bajo el criterio BIC es el que maximiza [4]

$$BIC(Q) = \log p(O | \hat{\mathbf{A}}_Q) - \frac{M_Q}{2} \log(N_o) \quad (5)$$

En la ecuación (5) $\log p(O | \hat{\mathbf{A}}_Q)$ denota la estimación de máxima verosimilitud del modelo con Q estados y M_Q es el número total de parámetros libres de $\hat{\mathbf{A}}_Q$.

Criterio de longitud de descripción mínima para mezclas (MMDL)

En [7] se propone una modificación del criterio MDL para aplicarse en modelos de mezcla. Esta modificación se conoce como el criterio de longitud de descripción mínima para mezclas (MMDL). Para el caso de un HMM y bajo el criterio MMDL, el número óptimo de estados es el que maximiza [4]

$$MMDL(Q) = \log p(O | \hat{\mathbf{A}}_Q) - \frac{M_Q^A + M_Q^\pi}{2} \log(N_o) - \frac{M_1^\theta}{2} \sum_{k=1}^Q \log(N_o p_\infty(k)) \quad (6)$$

En la ecuación anterior, M_Q^A es el número de parámetros libres asociados a la matriz \mathbf{A} , M_Q^π es el número de parámetros libres asociados al vector $\boldsymbol{\pi}$, M_1^θ es el número de parámetros libres en la función de probabilidad asociada al modelo de observación en un HMPCA de un único estado y \mathbf{p}_∞ es la distribución de probabilidad estacionaria. La distribución de probabilidad estacionaria $\mathbf{p}_\infty = [p_\infty(1), \dots, p_\infty(Q)]$ se calcula resolviendo el problema espectral $\mathbf{p}_\infty = \mathbf{p}_\infty \mathbf{A}$. Es decir, \mathbf{p}_∞ es el vector propio izquierdo que corresponde al valor propio uno, asociado a la matriz estocástica \mathbf{A} . Este vector se estima usando el método "reemplazar una ecuación" [8] [9].

BIC y MMDL para HMPCA

Para el modelo de PPCA relacionado con cada estado en HMPCA, el número de parámetros libres asociados a la matriz de covarianza está dado por $dp + 1 - p(p-1)/2$ [10] y el número de parámetros libres de la media es d . Luego, el valor de $M_1^\theta = dp + 1 - p(p-1)/2 + d$, es una función de p . Para enfatizar este hecho, se escribe $M_1^\theta(p)$. Lo anterior hace que $BIC(Q)$ y $MMDL(Q)$ sean funciones, no sólo del parámetro Q , sino del parámetro p . De esta forma, las ecuaciones (5) y (6) deben modificarse como

$$BIC(Q, p) = \log p(O | \hat{\mathbf{A}}_Q) - \frac{(Q+1)(Q-1) + Q M_1^\theta(p)}{2} \log(N_o) \quad (7)$$

$$MMDL(Q, p) = \log p(O | \hat{\mathbf{A}}_Q) - \frac{(Q+1)(Q-1)}{2} \log(N_o) - \frac{M_1^\theta(p)}{2} \sum_{k=1}^Q \log(N_o p_\infty(k)) \quad (8)$$

donde $M_Q = M_Q^\pi + M_Q^A + M_Q^\theta$, $M_Q^A = Q(Q-1)$ y $M_Q^\pi = Q-1$.

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

Estrategia secuencial para la reducción de estados

En [4] se propone una estrategia de reducción secuencial de estados, cuyo aporte fundamental consiste en reducir el número de iteraciones necesarias para encontrar el valor óptimo de Q . La estrategia que se expone a continuación es una modificación de la estrategia de [4], para tener en cuenta el parámetro p ,

1. Escoger algún criterio de selección del modelo, como BIC (ecuación (7)) o MMDL (ecuación (8)). Establecer Q_{\min} y Q_{\max} , que son el número mínimo y máximo de estados permitidos.
2. Para $2 \leq p \leq d$, donde p es la dimensionalidad del subespacio,
 - a. Inicializar un modelo de HMPCA usando Q_{\max} . Se denota Λ_Q^l al modelo inicial empleado en el entrenamiento de un HMPCA con Q estados.
 - b. Mientras $Q \geq Q_{\min}$,
 - i. Utilizar el algoritmo EM para encontrar $\hat{\Lambda}_Q$.
 - ii. Calcular y guardar el valor del criterio de selección de modelo, que se denota como $C(Q, p)$.
 - iii. Encontrar el estado menos probable (el menor valor de p_∞)
 - iv. Podar el estado menos probable y eliminar los correspondientes elementos de \mathbf{A} , Θ para obtener un modelo reducido $\bar{\Lambda}$.
 - v. Establecer $\Lambda_{Q-1}^l \leftarrow \bar{\Lambda}$ y $Q \leftarrow Q-1$
3. El modelo seleccionado finalmente, Λ^* , es aquel para el cual el criterio de selección, $C(Q, p)$, es máximo.

4. RESULTADOS

Con el objetivo de probar el funcionamiento del método propuesto para la selección del número de estados y dimensión del subespacio principal en HMPCA, se comparan dos estrategias diferentes. (i) Entrenar un modelo de HMPCA para cada Q entre $Q_{\min}=2$ y $Q_{\max}=8$, y p para $2 \leq p \leq d$, guardando para cada ejecución el valor correspondiente a $BIC(Q, p)$ o $MMDL(Q, p)$ según sea el caso, para finalmente elegir los parámetros del modelo (Q, p) que maximicen el valor del criterio empleado. (ii) Utilizar la estrategia secuencial descrita en la sección anterior.

Debido a que el método propuesto está orientado a seleccionar solamente el número de estados y la dimensión del subespacio principal, para los experimentos sólo se considera el caso en que cada estado posee un único modelo de PPCA, es decir, el

número de mezclas por estado es igual a 1. La inicialización del modelo de HMPCA se realiza con el algoritmo de K -medias y PPCA. El criterio de parada consiste en terminar el algoritmo de entrenamiento cuando el incremento relativo de verosimilitud logarítmica entre iteraciones sucesivas esté por debajo de un umbral preestablecido ($10e-4$ en este caso). Los criterios de comparación para las estrategias consideradas son: (i) la precisión en la selección de los parámetros reales del modelo y (ii) el costo computacional medido como el tiempo requerido por la estrategia de selección para ejecutarse. Considerando que es necesario conocer el número real de estados y la dimensión del subespacio principal para realizar una comparación acertada, los experimentos fueron realizados con conjuntos de entrenamiento generados artificialmente. De otro lado, para incrementar la significancia estadística de las pruebas realizadas, cada experimento fue repetido 50 veces.

Experimento 1

Como primer experimento se ha empleado un modelo relativamente simple. El conjunto de datos consiste en 30 observaciones de longitud 100 y dimensión 10 tomadas de una cadena oculta de Markov de 3 estados con matriz de transición de probabilidades uniforme, distribución inicial de los estados uniforme y modelos de observación con medias $\mu = [-1 \ 0 \ 1]$ y matrices de covarianza diagonales $C_1 = C_2 = C_3 = \text{diag}(2 \times \mathbf{1}_5 \ 0.2 \times \mathbf{1}_5)$, donde C_Q es la matriz de covarianza asociada al estado Q y $\mathbf{1}_m$ es un vector de unos de m dimensiones. En la tabla 1 se presentan los resultados obtenidos. En este caso, BIC y MMDL se comportan mejor aunque con costo computacional ligeramente mayor ($\sim 100s$), sin embargo, BICsec obtuvo una precisión de 98%, correspondiente solamente a un error en la selección de los parámetros del modelo.

Criterio	Precisión	Tiempo (s)
BIC	100% (50/50)	386.59±9.08
MMDL	100% (50/50)	386.59±9.08
BICsec	98% (49/50)	262.28±4.91
MMDLsec	90% (45/50)	262.28±4.91

Tabla 1. Resultados obtenidos para el experimento 1. BICsec y MMDLsec indican que se ha empleado la estrategia secuencial. La precisión es número de veces que el criterio seleccionó correctamente los parámetros del modelo y el tiempo, la medida del costo computacional

En la figura 1 se muestran las superficies resultantes para BICseq y MMDLseq para una ejecución en particular de las 50 totales. Cabe anotar que en ambos casos, a medida que aumenta el número de estados y dimensiones del subespacio principal, el valor del criterio empleado aumenta substancialmente hasta acercarse a los parámetros seleccionados ($Q=3$ y $p=5$) y luego

empieza a decrecer, pero a una tasa mucho menor, lo cual tiene sentido debido a que la penalización en BIC y MMDL se empieza a hacer evidente cuando el número de parámetros de modelo empieza a incrementar (ver ecuaciones (5) y (6)).

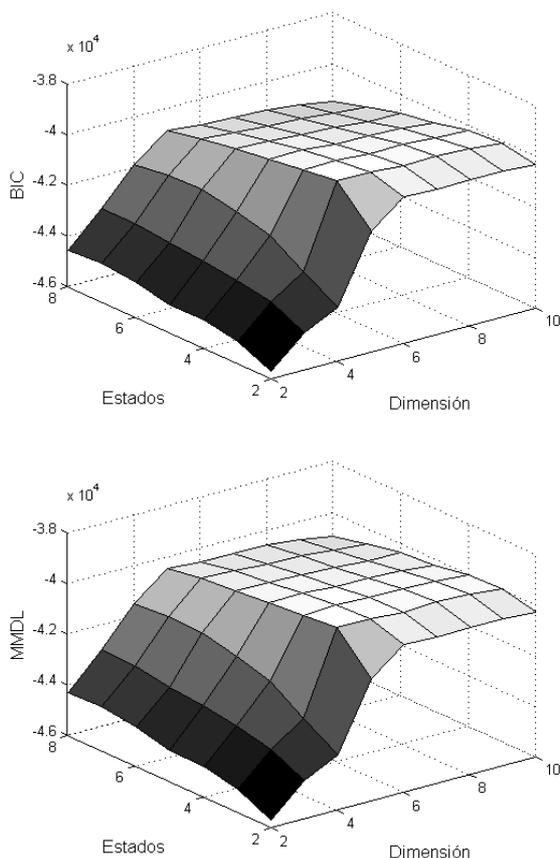


Figura 1. Superficies resultantes para BICsec y MMDLsec. En ambas gráficas, el máximo resultante es $Q=3$ y $p=5$

Experimento 2

Para este experimento, se han modificado las condiciones del primer experimento para obtener un modelo relativamente más complejo, en particular, se ha adicionado un estado a la cadena oculta de Markov y los parámetros del modelo de observación se han modificado de modo que ahora son $\mu = [-1 \ -1 \ 1 \ 1]$, $C_1 = \text{diag}(2 \times \mathbf{1}_5 \ 0.2 \times \mathbf{1}_5)$, $C_2 = \text{diag}(6 \times \mathbf{1}_5 \ 0.6 \times \mathbf{1}_5)$, $C_3 = \text{diag}(1 \times \mathbf{1}_5 \ 0.1 \times \mathbf{1}_5)$ y $C_4 = \text{diag}(5 \times \mathbf{1}_5 \ 0.5 \times \mathbf{1}_5)$. Como se muestra en la tabla 2 y a diferencia del experimento 1, en este caso, las estrategias secuenciales proveen en general un mejor desempeño, siendo BICsec el mejor criterio seleccionando el modelo correctamente 48 de las 50 ejecuciones de la estrategia. Los resultados obtenidos parecen indicar que BIC y MMDL se comportan satisfactoriamente siempre y cuando la complejidad del modelo sea relativamente baja.

Criterio	Precisión	Tiempo (s)
BIC	32% (16/50)	411.75±14.52
MMDL	30% (15/50)	411.75±14.52
BICsec	96% (48/50)	282.12±24.61
MMDLsec	84% (42/50)	282.12±24.61

Tabla 2. Resultados obtenidos para el experimento 2

Experimento 3

Finalmente, se ha partido del mismo modelo del experimento 2 pero cambiando la matriz de transición de estados de la siguiente forma:

$$A = \begin{bmatrix} 0.85 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.85 \end{bmatrix}$$

De esta manera, obtenemos un modelo mucho más complejo debido a que la matriz de transición es ahora informativa. En la tabla 3 se presentan los resultados correspondientes.

Criterio	Precisión	Tiempo (s)
BIC	18% (36/50)	543.84±18.43
MMDL	18% (36/50)	543.84±18.43
BICsec	94% (47/50)	315.21±18.31
MMDLsec	86% (43/50)	315.21±18.31

Tabla 3. Resultados obtenidos para el experimento 3

Los resultados confirman que las estrategias secuenciales se comportan mejor cuando la complejidad del modelo aumenta, siendo BICsec de nuevo el mejor criterio con sólo tres errores de predicción y con un tiempo de ejecución de aproximadamente el 58% del requerido por BIC y MMDL estándar.

5. DISCUSIÓN Y CONCLUSIONES

En este artículo se ha mostrado que los criterios de información empleados para la selección de los parámetros del modelo, i.e. el número de estados y la dimensión del subespacio principal en HMPCA, se comportan en términos generales de manera satisfactoria. Además, se ha encontrado que las estrategias secuenciales aparte de reducir el costo computacional permiten capturar la estructura del modelo un mayor número de veces. En particular, BICsec presenta un mejor desempeño en dos de los tres experimentos realizados con precisiones en todos los casos cercanas al valor óptimo.

Si bien parece que MMDL es más robusto que BIC partiendo sólo de la estructura de la formulación, los

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

resultados indican que BIC se comporta mejor. Dado que la formulación de MMDL se concentra en la información provista por la matriz de transición de estados y no considera directamente que pueden existir diferentes dimensionalidades en el modelo, es posible que para el caso de HMM, el criterio exhiba un mejor comportamiento, sin embargo, dados los resultados este no es el caso de HMPCA, por lo cual, con miras a mejorar el desempeño del criterio sería necesario modificarlo para acomodarlo a la estructura particular del modelo de HMPCA.

En términos de costo computacional, aunque la diferencia de tiempos de ejecución entre las estrategias secuenciales y los criterios estándar no es considerablemente grande, para aplicaciones donde se requiera un modelo más complejo y la cantidad de observaciones empleadas para su entrenamiento sea mucho más grande, dicha diferencia se incrementaría substancialmente, con lo cual, podría convertirse en un aspecto crítico para el diseño de la aplicación.

La metodología propuesta en este trabajo funciona de manera satisfactoria, sin embargo queda abierto el problema de selección de número de mezclas por estado y la modificación de MMDL para el caso particular del modelo de HMPCA. De otro lado, sería mucho más apropiado adoptar un esquema bayesiano para la selección del modelo, lo cual sería mucho más eficiente desde el punto de vista teórico debido a que el modelo resultante no sería tan dependiente del conjunto de datos utilizados para el entrenamiento, que es tal vez uno de los mayores inconvenientes de los enfoques basados en máxima verosimilitud.

6. AGRADECIMIENTOS

Este trabajo es financiado por el Centro de Investigación y Extensión (CIE) mediante el proyecto *Clasificación Automatizada de Eventos Sísmicos*, contrato 9-07-4 y el proyecto *Análisis de la variabilidad estocástica en la detección de patologías sobre registros electrocardiográficos y de voz*, contrato 6-07-3.

7. BIBLIOGRAFÍA

[1] M. Alvarez, R. Henao. Hidden Markov Bayesian Principal Component Analysis. Lecture Notes in Computer Science, Neural Information Processing/ICONIP 2007 (Aceptado para publicación).
 [2] M. Alvarez, R. Henao. Maximum Likelihood for Training Hidden Markov Principal Component Analyzers. XII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial, 2007 (Aceptado para publicación).

[3] M.E. Tipping y C.M. Bishop, Probabilistic Principal Component Analysis, Journal of the Royal Statistical Society, Series B, vol. 21, no. 3, pp. 611–622, 1999.

[4] M. Bicego, V. Murino, M. Figueiredo. A sequential pruning strategy for the selection of the number of states in hidden Markov models. Pattern Recognition Letters. Vol. 24, pp. 1395-1407. 2003

[5] X. Huang, A. Acero, H. W. Hon. Spoken Language Processing. Prentice Hall. Upper Saddle River, New Jersey. 2001.

[6] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of The IEEE, vol. 77, no. 2, 1989.

[7] M. Figueiredo, J. Leitão, y A.K.Jain, On Fitting Mixture Models, en E. Hancock and M. Pellilo (Editors), Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 54 - 69, Lecture Notes in Computer Science, vol. 1654, Springer Verlag, 1999.

[8] R. D. Yates y D. J. Goodman, Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers, Wiley & Sons. 2nd Edition. 2005.

[9] W. J. Stewart, Introduction to the Numerical Solution of Markov Chains. Princeton University Press. 1994.

[10] C. Bishop. Pattern Recognition and Machine Learning. Springer. 2006.