SELECCIÓN DE DIMENSIONALIDAD EN ANÁLISIS DE COMPONENTES PRINCIPALES UTILIZANDO MODELOS BAYESIANOS

RESUMEN
El gran inconveniente del análisis de componentes principales (PCA) es la correct elección del número componentes que deben ser retenidas, debido a esto, en este artículo se presenta un análisis experimental de varias técnicas de selección de dimensionalidad automáticas para PCA, basadas en dos variantes denominadas Análisis de Componentes Principales Probabilístico (PPCA) y Análisis Variacional de Componentes Principales (VPCA). Los métodos de empleados en este trabajo están fundamentados en la selección bayesiana del modelo y los criterios de información. En los resultados obtenidos, el método que aplica Laplace obtiene mejores resultados globales con un costo computacional satisfactorio.

PALABRAS CLAVES: Análisis de componentes principales, Modelos Bayesianos, Selección del Modelo, Análisis de Componentes Principales Probabilístico.

RICARDO HENAO
Ingeniero Electrónico, Ms.C
Profesor Auxiliar
Grupo de Investigación "LIDER"
Universidad Tecnológica de Pereira
rhenao@utp.edu.co

JORGE HERNÁNDO RIVERA
Ingeniero Electrónico
Profesor Auxiliar
Grupo de Investigación "LIDER"
Universidad Tecnológica de Pereira
j.rivera@utp.edu.co

ABSTRACT
The central inconvenient in Principal Component Analysis (PCA) is to choose correctly the number of principal components to be retained. This article presents an experimental analysis reviewing several techniques of automatic dimensionality selection for PCA based on two variations designated as Probabilistic Principal Component Analysis (PPCA) and Variational Principal Component Analysis (VPCA). The methods used in this work are based on bayesian model selection and information criteria. The obtained results showed that Laplace is better globally speaking and computationally faster.

KEYWORDS: *Principal component analysis, Bayesian models, model selection, Probabilistic Principal Component Analysis.*

1. INTRODUCCIÓN

El análisis de componentes principales (PCA) [4] descompone un conjunto de datos de alta dimensión en un subespacio de baja dimensión. Sin embargo PCA presenta un gran inconveniente que reside en la selección de la cantidad de componentes suficientes para representar un subespacio consistente para una aplicación en particular, debido a esto, en este trabajo se presenta un análisis experimental de varias técnicas de selección de dimensionalidad automática para PCA. Con esto, dicha descomposición puede ser útil para compresión, atenuación de ruido, preprocesamiento de datos, procesamiento de imágenes, visualización, análisis exploratorio, reconocimiento de patrones y predicción de series de tiempo. Recientemente, [2] presenta una formulación probabilística de PCA (PPCA) a partir de un modelo gaussiano de variables latentes el cual está estrechamente relacionado con el análisis de factores estadístico y que además puede ser interpretado como estimación de densidad con máxima verosimilitud (ML) y más tarde, en [5] se expone una formulación alternativa de PCA bayesiano presentada en [1] denominada análisis variacional de componentes principales (VPCA) partiendo de el hecho de que la mayoría de los modelos bayesianos no triviales requieren marginalizaciones analíticamente intratables, para lo cual una aproximación basada en una representación gaussiana local (factorial) de la distribución posterior puede ser empleada.

2. PCA PROBABILÍSTICO

Un modelo variable latente busca relacionar un vector dimensional de observaciones t con un correspondiente vector q-dimensional de variables latentes x, luego t puede ser expresado como una combinación lineal de vectores base y ruido:

$$t = Wx + \mu + \varepsilon \quad (1)$$

donde, la matrix W de tamaño $d \times q$ es la base y el vector μ permite que el modelo tenga media diferente de cero. Convencionalmente, $p(x) \sim N(0, I)$ y $p(e) \sim N(0, \Psi)$, que para el caso de PCA corresponde a un modelo de error o ruido isotrópico con varianzas residuales $\psi_i i = \sigma^2$ de manera que $\Psi = \sigma^2 I$ a diferencia del análisis de factores en el cual Ψ es en general diagonal. El objetivo de PCA es estimar la base W y la varianza del ruido σ^2 a partir del conjunto $D = t_1, ..., t_N$. La ecuación (1) implica que la probabilidad de observar t dado x es:

Fecha de Recepción: 31 Mayo de 2005
Fecha de Aceptación: 16 Agosto de 2005

$$p\left(t\middle|x,W,\mu,\sigma^2\right)\sim \mathrm{N}\left(Wx+\mu,\sigma^2 I\right)$$
$$p\left(x\middle|W,\mu,\sigma^2\right)=\int_W p\left(t\middle|x,W,\mu,\sigma^2\right)p(W)\,dW \qquad (2)$$
$$\sim N\left(\mu,WW'+\sigma^2 I\right)$$

Con esto, la probabilidad del conjunto de datos *D* es:

$$p\left(D\middle|W,\mu,\sigma^2\right)\sim\prod_i p\left(x_i\middle|W,\mu,\sigma^2\right)$$

y la correspondiente verosimilitud logarítmica,

$$L=\log\left(p\left(D\middle|W,\mu,\sigma^2\right)\right)=-\frac{N}{2}\left\{d\log(2\pi)+\log|C|+tr\left(C^{-1}S\right)\right\} \qquad (3)$$

donde

$$C=WW'+\sigma^2 I$$
$$S=\frac{1}{N}\sum_{i=1}^{N}(t_i-\mu)(t_i-\mu)'$$

con *S*, la matriz de covarianza muestal. El estimador ML para μ es la media de *D* ($\mu=1/N\sum_i t_i$). Las estimaciones para *W* y σ^2 pueden ser obtenidas mediante maximización iterativa de ***L***, sin embargo PPCA muestra que en contraste con el análisis de factores el máximo de (3) ocurre cuando:

$$W_{ML}=U_q\left(\Lambda_q-\sigma^2 I\right)^{1/2}R \qquad (4)$$

donde los *q* vectores columna en la matriz *Uq* de tamaño $(d\times q)$ son los vectores propios de *S* con valores propios correspondientes $\lambda_1,\ldots,\lambda_q$ en la matriz diagonal Λ_q y *R* una matriz de rotación ortogonal arbitraria de tamaño ($q\times q$). Además, para $W=W_{ML}$ el estimador ML para σ^2 está dado por:

$$\sigma^2_{ML}=\frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

la cual es el promedio de los valores propios restantes y puede interpretarse como la varianza "perdida" en la proyección promediada sobre el número de dimensiones descartadas.

Con esto último, ***L*** de la ecuación (3) se vuelve:

$$L_{ML}=\log\left(D\middle|W_{ML},\mu,\sigma^2_{ML}\right)$$
$$=-\frac{N}{2}\left\{\sum_{i=1}^{q}\log(\lambda_i)+(d-q)\log(\sigma^2_{ML})+d\log(2\pi)+d\right\} \qquad (5)$$

De la ecuación (5), la matriz de covarianza de *t* es $U_q\tilde{\Lambda}U_q'$, donde U_q contiene todos los vectores propios de *S* y

$$\tilde{\Lambda}=\begin{bmatrix}\Lambda_q & 0\\ 0 & \sigma^2_{ML}I\end{bmatrix} \qquad (6)$$

es decir, la estimación ML de la covarianza pero con los (*d*-*q*) valores propios más pequeños establecidos como su promedio.

3. ANÁLISIS VARIACIONAL DE COMPONENTES PRINCIPALES

A partir de la formulación de PPCA dada en la sección 2 se puede obtener la densidad predictiva a partir de $p(t|W,\mu,\sigma^2)$ y su correspondiente distribución posterior $p(W,\mu,\sigma^2|D)$, marginalizando sobre los parámetros de la siguiente manera:

$$p(t|D)=\int p(t|W,\mu,\sigma^2)\,p(W,\mu,\sigma^2|D)\,dW\,d\mu\,d\sigma^2 \qquad (7)$$

Para poder implementar (7) se deben tener en cuenta dos aspectos: la elección de la distribución previa y una formulación numéricamente tratable para realizar la marginalización. Respecto del primero, en [1] el autor se enfoca en seleccionar las dimensionalidad efectiva del subespacio principal evitando un esquema de selección discreta de modelo para mas bien utilizar un conjunto de hiperparámetros continuos que permitan determinar automáticamente la dimensión del modelo como parte del proceso de inferencia bayesiana. Esto último se puede lograr introduciendo en la formulación una densidad previa jerárquica $p(W|\alpha)$ sobre la matriz *W* donde α es un vector de q dimensiones. Cada α_i controla una de las columnas de *W* a través de una distribución gaussiana de la forma:

$$p(W|\alpha)=\prod_{i=1}^{q}\left(\frac{\alpha_N}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\alpha_i\|w_i\|^2\right)$$

donde w_i es columna de *W*. Cada α_i controla la varianza inversa de su correspondiente w_i de manera que para un α_i en particular con distribución posterior concentrada en valores grandes, poseerá un w_i pequeño de manera que la dirección controlada por este último puede ser efectivamente descartada si α_i es lo suficientemente grande. Para completar el modelo bayesiano se deben elegir densidades previas para el resto de los parámetros. Haciendo $\tau\equiv\sigma^2$ se pueden calcular como:

$$p(\mu)\sim N(\mu|0,\beta^{-1}I)$$
$$p(\alpha)\sim\prod_{i=1}^{q}\Gamma(\alpha_i|a_\alpha,b_\alpha)$$
$$p(\tau)\sim\Gamma(\tau|c_\tau,d_\tau)$$

donde $\Gamma(x|a,b)$ denota una distribución gamma sobre x dada por:

$$\Gamma(x|a,b)=\frac{b^a x^{a-1}e^{-bx}}{\Gamma(a)} \qquad (8)$$

con $\Gamma(a)$ la función gamma. La distribución de la ecuación 8 posee las siguientes propiedades útiles:

$$\langle x\rangle=\frac{a}{b}$$
$$\langle x^2\rangle-\langle x\rangle^2=\frac{a}{b^2}$$

de manera que para obtener densidades previas amplias, se pueden establecer como parámetros $a_\alpha=b_\alpha=a_\tau=b_\tau=10^{-3}$ y $\beta=10^{-3}$ [1].

En cuanto al segundo aspecto a tener en cuenta correspondiente a una formulación numérica tratable para marginalizar los parámetros, [1] propone dos aproximaciones basadas en ML tipo II utilizando aproximación gaussiana local para la distribución

posterior y una cadena de Markov de monte Carlo basada en muestro de Gibbs. En este trabajo se emplea el tratamiento variacional propuesto en [5] que es computacionalmente más eficiente.

Considerando el problema de evaluar la verosimilitud marginal,

$$p(D) = \int p(D,\theta)d\theta$$

donde θ denota todos los parámetros y variables latentes del modelo. Los métodos variacionales pretenden mediante la introducción de una distribución $Q(\theta)$ proveer una aproximación al verdadero posterior del modelo, para lo cual se puede considerar la siguiente transformación:

$$p(D) = \log \int p(D,\theta)d\theta = \log \int Q(\theta)\frac{p(D,\theta)}{Q(\theta)}d\theta \quad (9)$$
$$\geq Q(\theta)\int \log \frac{p(D,\theta)}{Q(\theta)}d\theta = L(\theta)$$

obtenida aplicando la desigualdad de Jensen. De la ecuación 9 se puede ver que $L(\theta)$ es una rigurosa cota inferior de la verdadera verosimilitud logarítmica marginal y la diferencia entre $\log p(D)$ y $L(\theta)$ está dada por:

$$KL(Q \| p) = -\int Q(\theta)\log \frac{p(\theta \mid D)}{Q(\theta)}d\theta$$

la cuál es la divergencia de Kullback-Leibler (KL) entre la distribución de aproximación $Q(\theta)$ y el verdadero posterior $p(\theta \mid D)$. El objetivo principal del tratamiento variacional es elegir una forma apropiada para $Q(\theta)$ de tal manera que $L(Q)$ pueda ser fácilmente evaluada y que sea lo suficientemente flexible para que la cota pueda ajustarse lo suficiente. Para el caso de VPCA se restringe la forma funcional de $Q(\theta)$ asumiendo que se puede factorizar sobre sus componentes θ_i, de manera que

$$Q(\theta) = \prod_i Q_i(\theta_i) \quad (10)$$

La divergencia KL puede ser minimizada sobre todas los posibles factores haciendo una minimización libre sobre $Q(\theta)$, de lo cual se obtiene el siguiente resultado:

$$Q_i(\theta_i) = \frac{\exp\langle \log p(D \mid \theta)\rangle_{k\neq i}}{\int \exp\langle \log p(D \mid \theta)\rangle_{k\neq i} d\theta_j}$$

donde $\langle . \rangle_{k\neq i}$ es la esperanza con respecto a las distribuciones $Q_k(\theta_k)$ para todo $k \neq i$. De la ecuación (10) reemplazando los parámetros se obtiene:

$$Q(X,W,\alpha,\mu,\tau) = Q(X)Q(W)Q(\alpha)Q(\mu)Q(\tau) \quad (11)$$

La forma de calcular cada una de las componentes de (11) es bastante extensa, por lo tanto no se incluye en este trabajo, sin embargo puede ser encontrada en [5].

4. SELECCIÓN BAYESIANA DEL MODELO PARA PPCA

La selección bayesiana de modelo utiliza las reglas de la teoría de probabilidad para seleccionar uno entre varias hipótesis diferentes lo cual es completamente análogo a la clasificación bayesiana. La probabilidad del conjunto de datos dado un modelo puede ser calculada como la integral de dicho modelo sobre un conjunto de parámetros desconocidos θ que lo definen:

$$p(D \mid M) = \int_\theta p(D \mid \theta)p(\theta \mid M)d\theta \quad (12)$$

La expresión (12) se denomina evidencia del modelo M [4,3]. Una propiedad útil de la selección bayesiana de modelo es que garantiza la elección real del modelo siempre y cuando esté dentro de las hipótesis candidato mientras el tamaño del conjunto de datos tiene a infinito.

Para el modelo de PCA se desea seleccionar la dimensión q del subespacio y para hacer esto se debe calcular la probabilidad de los datos en cada una de las posibles dimensiones lo cual requiere integrar sobre todos los parámetros de PCA, es decir, sobre (μ, W, σ^2). Asumiendo que solo se tiene la información suministrada por D, se deben seleccionar densidades previas para cada parámetro que provean la menor cantidad de información posible. Para el case de μ se puede elegir una densidad previa uniforme debido a que no afecta la selección del modelo en este caso. A diferencia de μ, W debe tener una densidad previa apropiada teniendo en cuenta que varía en dimensión para cada modelo, con esto, W se puede descomponer de igual manera que en la ecuación 4, con la cual se puede construir una densidad previa gaussiana conjugada para $(U_q, \Lambda_q, R, \sigma^2_{ML})$ parametrizado por α como:

$$p(U_q,\Lambda_q,R,\sigma^2) \propto |\Lambda_q|^{-(\alpha+2)/2} (\sigma^2)^{(\alpha+2)(d-q)/2} \exp(-\frac{\alpha}{2}tr(\Lambda_q^{-1}))\exp(-\frac{\alpha(d-q)}{2\sigma^2}) \quad (13)$$

Como las variables son a-priori independientes, la distribución se puede descomponer en factores $p(U_q)p(\Lambda_q)p(R)p(\sigma^2_{ML})$ con $p(\Lambda_q)$ y $p(\sigma^2_{ML})$ de la ecuación (13) y $p(U_q)$ constante y definida por:

$$p(U_q) = 2^{-q}\prod_{i=1}^{q}\Gamma((d-1+i)/2)\pi^{(d-1+i)/2} \quad (14)$$

Además de proveer una densidad previa conveniente, la descomposición en (4) permite remover grados de libertad redundantes dados por R y separar W_{ML} en componentes independientes. Combinando la verosimilitud con la densidad previa de las ecuaciones 3 y 16 se obtiene el posterior como:

$$p(D \mid q) = c_k \int |C_{ML}|^{-n/2} \exp(-\frac{1}{2}tr(C_{ML}^{-1}(S+\alpha I)))dU_q d\Lambda_q d\sigma^2_{ML} \quad (15)$$

con $n = N+1+\alpha$, $C_{ML} = W'_{ML}W_{ML} + \sigma^2_{ML}I$ y c_k los términos de normalización para $p(U_q)$, $p(\Lambda_q)$ y $p(\sigma^2_{ML})$ [7]. En las fórmulas no aparece R debido a que luego de integrar se obtiene una constante que multiplica a $\int_R p(R)dR = 1$.

Para integrar (15) se puede emplear el método de Laplace el cual es bastante utilizado para aproximar integrales en

estadística de Bayes [4]. Utilizando dicho método, la aproximación de $p(D|q)$ está dada por:

$$p(D|q) \approx p(U_q)\left(\prod_{i=1}^{q}\lambda_i\right)^{-N/2}(\sigma^2)^{(-N(d-q)/2)}(2\pi)^{(m+q)/2}|A_Z|^{-1/2}N^{-q/2} \quad (16)$$

con $|A_Z|$ dada por:

$$|A_Z| = \prod_{i=1}^{q}\prod_{j=i+1}^{d}(\widetilde{\lambda}_j^{-1} - \widetilde{\lambda}_i^{-1})(\lambda_i - \lambda_j)N \quad (17)$$

donde $\widetilde{\lambda}_j$ son los elementos de $\tilde{\Lambda}$ de la ecuación (6). El costo de calcular $p(D|q)$ es $O(\min(N,d)q)$ sin contar el cálculo de los valores propios λ_i de S.

Una simplificación del método de Laplace es la aproximación por el criterio bayesiano de información (BiC) [4] que desecha de la ecuación (16) los términos que no incrementan con N, en cuyo caso se obtiene:

$$p(D|q) \approx \left(\prod_{i=1}^{q}\lambda_i\right)^{-N/2}(\sigma^2)^{(-N(d-q)/2)}(2\pi)^{(m+q)/2}N^{-(m-q)/2} \quad (18)$$

que es computacionalmente más eficiente que Laplace.

En [8] los autores hacen selección con un modelo ligeramente diferente al de PPCA. Partiendo de la ecuación (1) restringiendo W a ser ortogonal, esto es $W'W = I$ e incluyendo una densidad previa $p(x|\alpha) \sim N(0, I/\alpha)$ para α, con lo cual se puede obtener analíticamente el valor ML para α como:

$$\alpha_q^{-1} = \frac{\sum_{i=1}^{k}\lambda_j}{q} - (\sigma_{ML}^2)$$

y con W_{ML}, μ y σ_{ML}^2 de la misma manera que se obtuvieron para PPCA, el posterior ML se calcula como:

$$p(D|W_{ML}, \mu, \sigma_{ML}^2, \alpha_q) = (2\pi)^{-Nd/2}\left(\frac{\sum_{i=1}^{k}\lambda_j}{q}\right)(\sigma_{ML}^2)^{N(d-q)/2}\exp(-\frac{Nd}{2}) \quad (19)$$

5. SELECCIÓN DEL MODELO MEDIANTE CRITERIOS DE INFORMACIÓN

En PCA se han utilizado desde hace mucho tiempo los criterios de información para seleccionar la dimensión del modelo ya que son fáciles de implementar y requieren de un mínimo costo computacional. Dentro de los criterios más conocidos se encuentran el criterio de información de Akaike (AIC) y la longitud de descripción mínima (MDL).

Partiendo de un modelo de PCA de la forma de 1 con $\mu = 0$ y asumiendo que D está conformado por vectores gaussianos independiente e idénticamente distribuidos (iid) con media 0, se puede construir una función de densidad de probabilidad parametrizada $p(D|\theta)$ que genera las siguientes verosimilitudes logarítmicas:

$$AIC = -2\log(p(D|\theta_{ML})) + 2k$$

$$MDL = -\log(p(D|\theta_{ML})) + \frac{1}{2}k\log(N)$$

donde θ_{ML} es el estimador ML para θ y k es el número de parámetros libres del modelo. Se puede ver que para $k = q(2d-q)$ y θ_{ML}, los posteriores ML logarítmicos se pueden calcular como [4]:

$$AIC(q) = 2q(2d-q) - 2N(d-q)\log(\varrho(q))$$
$$MDL(q) = \frac{1}{2}q(2d-q)\log(N) - N(d-q)\log(\varrho(q)) \quad (20)$$

donde,

$$\varrho(q) = \frac{(\prod_{i=q+1}^{d}\lambda_i)^{1/(d-q)}}{1/(d-q)\sum_{i=q+1}^{d}\lambda_i}$$

es la razón entre la media geométrica de los $(d-q)$ valores propios más pequeños de S y su media aritmética. Las principales limitaciones que poseen estos dos estimadores es que son bastante sensitivos a las variaciones en la relación señal a ruido (SNR) y a la cantidad de observaciones disponibles.

6. RESULTADOS

Con la finalidad de probar el desempeño de los algoritmos descritos en este trabajo se tomaron conjuntos de datos tanto artificiales como reales conociendo de antemano el modelo que siguen con el objeto de observar que tan a menudo dichos algoritmos seleccionaban la dimensión adecuada. En total se implementaron 6 estimadores a saber: Laplace (16), BIC (18), RR-N (19), AIC (20.1), MDL (20.2) y VPCA. En el caso VPCA se presentan los pesos del vector debido a la naturaleza del método, por lo tanto la dimensionalidad estimada es igual al número de componentes no suprimidas (pesos menores). Las pruebas realizados en este trabajo siguen un modelo similar al utilizado por Minka [7], en su marco experimental.

Suficiencia de observaciones ($N>>d$) Para este primer experimento los datos fueron generados a partir de una distribución gaussiana de 10 dimensiones con varianza en 5 direcciones principales dadas por el vector [10 8 6 4 2] y varianza 1 en las 5 restantes. La figura 1 muestra los valores propios tanto de la matriz de covarianza real como de la observada S para una realización particular de 100 muestras. En la figura 2 se presentan los resultados obtenidos por cada uno de algoritmos implementados, todos ellos entregaron $q = 5$ a diferencia de RR-N que seleccionó $q = 4$. Los resultados acumulados luego de 60 repeticiones se pueden observar en la figura 3 y muestran para BIC y Laplace selección sin fallos, en el caso de AIC y MDL la mayoría de los fallos ocurrieron porque seleccionaron $q = 6$ a diferencia de VPCA y RR-N que presento fallos principalmente eligiendo $q = 4$ lo cual es de alguna manera razonable considerando que la varianza de la dimensión 5 es 2.

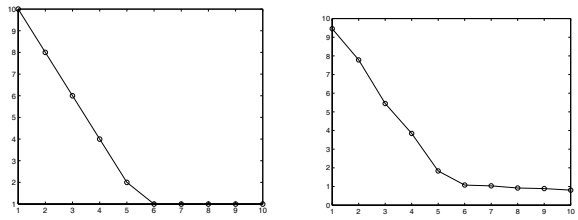
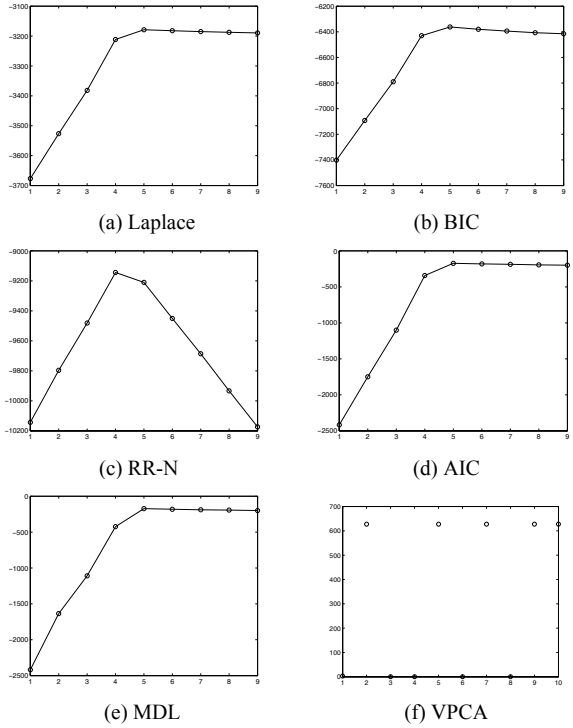


Figura 1.Valores propios a) Reales y b) observados para N=100, d=10 y q=5



(a) Laplace (b) BIC

(c) RR-N (d) AIC

(e) MDL (f) VPCA

Figura 2. Resultados para cada uno de los 6 algoritmos, la dimensión real es 5.

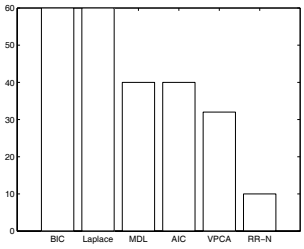


Figura 3. Resultados luego de 60 repeticiones, para *N*=100, *d*=10 y *q*=5

Pocos datos (*N*>*d*) En este experimento se tiene un conjunto de pocos datos *N*=20 con dimensionalidad comparable *d*=15 y condición de ruido bajo. La varianza se encuentra en las 5 primeras dimensiones con 10 componentes adicionales de ruido con varianza 0.1. Los resultados luego de 60 repeticiones se presentan en la figura 4.

Alta dimensionalidad del ruido Este conjunto es generado a partir de una distribución gaussiana de 100 dimensiones con varianza en las 5 primeras direcciones dada por el vector $[10\ 8\ 6\ 4\ 2]$ y $1/4$ en las 95 restantes.

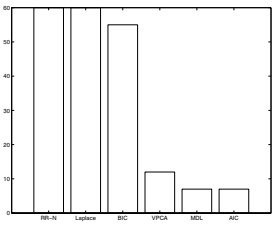
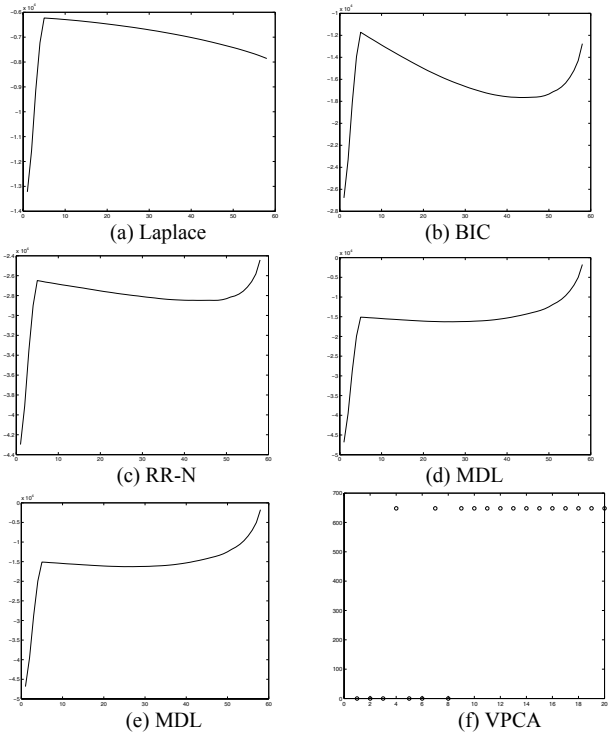


Figura 4. Resultados luego de 60 repeticiones, para *N*=20, *d*=15 y *q*=5

En la figura 6 se muestran los resultados para cada método independientemente, para el caso de BIC, AIC y MDL, debido a que son aproximaciones que parten de el hecho de que $N >> d$, cuando N es comparable con d las estimaciones pierden bastante confiabilidad lo cual se refleja en los picos en el extremo derecho de las figuras, sinembargo parece no ser tan acentuado en el caso de BIC. De otro lado, tales picos pueden ser rechazados siempre y cuando se presente un pico claro en otro lado con dimensionalidad menor. En la figura 7 se presentan los resultados para 60 repeticiones de cada algoritmo, se debe anotar que para este caso el costo computacional de VPCA es considerablemente alto cuando d es grande.



(a) Laplace (b) BIC

(c) RR-N (d) MDL

(e) MDL (f) VPCA

Figura 5. Resultados de cada uno de los 6 algoritmos, la dimensión real es 5

Distribución uniforme (*N*>>*d*) En este experimento es similar al primero con la diferencia que la distribución que genera los datos es uniforme. En la figura 6 se presentan los resultados obtenidos luego de 60 repeticiones. En comparación a la primera prueba, los resultados son similares para ambas distribuciones pero con un desempeño sustancialmente menor para RR-N.

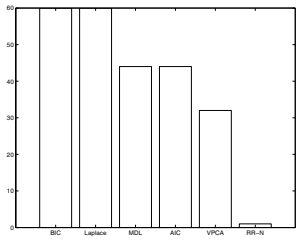
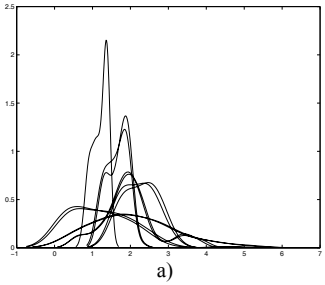


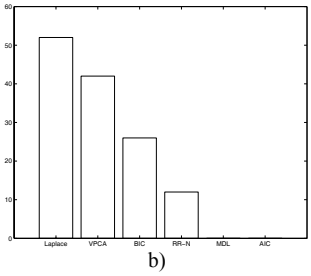
Figura 6: Resultados luego de 60 repeticiones, para N=100, d=10 y q=5

Densidades arbitrarias Este último experimento pretende probar que tan robustos son los algoritmos implementados en este trabajo considerando 11 dimensiones con funciones de densidad arbitrarias para conformar una distribución que es en general no gaussiana. Para complicar las cosas un poco más los datos en cada dimensión no poseen media cero ni varianza unitaria. En la figura 7(a) se muestran las distribuciones estimadas con kernel de cada una de las 11 dimensiones del subespacio principal.

A el conjunto inicial de 11 dimensiones se le adicionan otras 20 direcciones de ruido gaussiano con media 0 y varianza 1/2. En la figura 7(b) se muestran los resultados obtenidos para 60 repeticiones de cada uno de los 6 algoritmos con las 11 primeras dimensiones constantes y diferentes realizaciones de ruido.



a)



b)

Figura 7: Resultados para cada el conjunto de datos con densidades arbitrarias, N=91, d=31 y q=11

7. DISCUSIÓN

En un experimento en el que los valores propios no caigan hasta mantenerse en un punto mas o menos estable y más bien mantengan un comportamiento decreciente, se debe esperar que todos los estimadores seleccionen la máxima dimensión posible a excepción de RR-N debido a que su modelo es restrictivo, lo cual se sustenta en el hecho que dichos estimadores parten del concepto de estimación de densidad por lo tanto, pueden no ser apropiados para otros propósitos como reducir costo computacional o resaltar o seleccionar características "efectivas". Por ejemplo en una tarea de clasificación de registros ECG, con 1100 observaciones y 25 características extraídas con wavelets, los estimadores implementados para este trabajo seleccionaron en promedio entre 21 y 24 dimensiones con un vector de valores propios observados claramente decreciente, sin embargo un clasificador lineal óptimo implementado con máquinas de soporte vectorial revela que tomando solo 15 de las 25 componentes principales para construir el subespacio principal, el error de validación incrementa aproximadamente 1% lo cual no es significativo si se busca un balance entre costo computacional y desempeño del sistema de clasificación.

Cualquiera de estos estimadores podrían tener mejor desempeño si estuvieran ajustados a diferentes modelos de PCA para diferentes modelos en los datos de manera análoga a lo que presenta [9] para procesos gaussianos sobre el esquema de transporte de mensajes variacional o [10] con su variante de PCA llamada análisis generalizado de componentes (GCA) el cual supone que los datos siguen una distribución super gaussiana, aunque al ajustar la selección a un modelo específico de PCA o distribución modelo limita su utilización a aplicaciones particulares perdiéndose capacidad de generalización en la estimación.

De los estimadores que se implementaron, para los diferentes casos Laplace obtuvo en general mejores resultados tanto en selección apropiada como en costo computacional de lo cual se puede concluir que es un estimador robusto. El resto de los estimadores en alguno o varios de los casos presentaron buenos resultados de modo que su utilización está limitada a la naturaleza o razón dimensión/obsrevaciones del conjunto de datos.

8. BIBLIOGRAFÍA

[1] Bishop, C. M. Bayesian PCA *In M. S. Kearns, S. A. Solla, and D. A. Cohn (Eds.), Advances in Neural Information Processing Systems*. 11(3):382–388, 1998.

[2] Tipping, M. E. and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B., ICANN 99.* Volume 1, page 509–514, 1999.

[3] T. Jollife. *Principal Component Analysis.* Springer Verlag, 1986.

[4] R. E Kass and A. E Raftery, Bayes Models and Model Uncertainty, *The University of Washington, Department of Computer Science,* Technical Report no 254, 2003.

[5] C. M. Bishop. Variational principal components. In IEE, editor, *In Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99, Volume 1,* page 509-514, 1999.

[6] MacKay, D. J. C, Probable Networks and Plausible Predictions - A Review on Practical Bayesian Methods for Supervised Neural Networks, *Network: Computation in Neural Systems*, 6(1): 469-505, 1995.

[7] Thomas P. Minka. Automatic choice of dimensionality for PCA. In MIT Press, editor, *Neural Information Processing Systems, NIPS'00, 2000.*

[8] M.Wax and T. Kailath. Detection of signal by information theoretic criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing, 33(1):387-392, 1985.*

[9] Christopher M. Bishop John Winn. Variational message passing. Journal of Machine Learning, *6(1):661-694, 2005.*

[10] N. D. Lawrence and M.E Tipping. Generalised component analysis. Technical report no cs-03-10, The University of Sheffield, Department of Computer Science, Sheffield, 2003.