응용통계연구 (2011)
24(5), 963–977
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.963

수정된 반복 주성분 분석 기법에 대한 연구

김동규[1] · 김아현[2] · 김현중[3]

[1]연세대학교 응용통계학과, [2]연세대학교 응용통계학과, [3]연세대학교 응용통계학과

(2011년 5월 접수, 2011년 9월 채택)

요 약

다변량 자료를 분석함에 있어 자료의 차원을 축소하는데 활용되는 중요한 툴 중 하나인 PCA 분석(주성분 분석, Principal Component Analysis)을 실시간으로 처리해야 하는 적용 분야가 최근 늘고 있다. PCA 분석에서는 표본 공분산 행렬의 고유값과 고유벡터를 도출하는 것이 관건인데, 자료의 양이 방대하며 고차원인 경우 이를 실시간으로 수행하기에는 어려움이 따른다. 이러한 문제점을 해결하기 위해서 Erdogmus 등 (2004)는 일차 섭동 이론(first order perturbation theory)을 활용하여 공분산 행렬의 고유값과 고유벡터를 추정하는 Recursive PCA 방법을 제안했다. 이 방법은 추가된 자료의 양이 많지 않은 경우는 상당히 정확하지만, 추가된 자료의 양이 많아짐에 따라 오차도 커진다는 한계를 가지고 있다. 본 논문은 공분산 행렬의 고유값과 고유벡터가 가지고 있는 수학적 관계를 이용하여 Erdogmus 등 (2004)가 제안한 Recursive PCA 방법을 수정한 Modified Recursive PCA 방법을 제안하였다. 또한, 모의 실험을 통해 Recursive PCA 방법과 Modified Recursive PCA 방법에서의 고유값과 고유벡터 추정값의 정확도를 비교해 보았으며 그 결과 기존 Recursive PCA 방법 보다 정확한 추정이 가능함을 확인할 수 있었다.

주요용어: Recursive PCA, 주요인 분석, 일차 섭동 이론.

1. 서론

PCA 분석은 상관관계가 존재하는 다변량 자료에 대해 정보 손실을 가능한 줄이면서 자료의 차원을 축소하기 위한 방법으로 신호 처리(signal-processing)를 비롯한 다양한 분야에서 활용되고 있는 중요한 통계적 분석 기법이다 (Duda와 Hard, 1973; Kung 등, 1994). PCA 분석을 위해서는 주어진 자료의 공분산 행렬에 대한 고유값과 고유벡터 행렬을 찾는 것이 핵심이며, 이를 찾기 위한 다양한 기법들이 연구된 바 있다 (Golub과 Loan, 1993). 그런데 대부분의 경우 최초에 자료가 모두 주어진 경우에 적용할 수 있는 방법일 뿐만 아니라 방대한 수준의 행렬 연산을 필요로 하기 때문에, 실시간으로 대용량의 데이터가 업데이트 되는 분야에서는 이들 기법을 적용하는 것은 쉽지 않다. 또한 DOA(Direction of Arrival) 추적의 경우와 같이 신호 통계량(signal statistics)이 시간에 따라 변하는 분야에 적용함에 있어서 과거의 모든 데이터를 블록화하여 접근하는 방식은 한계가 있다.

따라서 좀 더 빠르고, 적응적이며, 실시간 분석에 용이한 PCA 분석 기법에 대한 요구가 증가하고 있는데, 이에 부응하여 Erdogmus 등 (2004)는 섭동 이론(perturbation theory)를 활용한 RPCA(Recursive

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2010-0008769).
[2]교신저자: (120-749) 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 박사과정.
E-mail: ahyun1003@yonsei.ac.kr

PCA) 방법을 제안하였다. 이 방법은 자료가 하나씩 새롭게 추가됨에 따라 그 자료의 공분산 행렬에 대한 고유값과 고유벡터 행렬도 업데이트 되는데, 업데이트 되는 정도를 일차 섭동 이론을 이용하여 반복적으로 추정한다는 것으로, 추가된 자료의 양이 많지 않은 경우에는 상당히 정확한 것으로 알려졌다. 그런데 추가된 자료의 양이 늘어남에 따라 고유값과 고유벡터에 대한 추정치는 정확도가 떨어져, 현실 적용에 있어서는 문제가 될 소지가 있다. 대게 PCA 분석의 결과는 사후 분석에 대한 자료로 활용하는 경우가 많기 때문에 PCA 분석의 핵심이 되는 고유값과 고유벡터에서 발생한 오차가 전체 모형에 있어서 중요한 결함이 될 수 있기 때문이다.

본 연구는 Erdogmus 등 (2004)의 RPCA 방법에서 정확도를 좀 더 개선할 수 있도록 수정한 MRPCA(Modified Recursive PCA) 방법을 제안하고자 한다. RPCA 방법의 경우 새로운 데이터가 추가될 때마다 이 효과가 반영된 고유값과 고유벡터를 동시에 추정하는 방식이지만, MRPCA 방법에서는 고유벡터만 RPCA 방법과 동일한 방식으로 먼저 추정한 뒤, 공분산 행렬과 고유값 및 고유벡터 간에 존재하는 이론적인 관계에 의해 고유값을 도출해 낸다는 것이다. 이렇게 함으로써 고유값에 대한 추정치는 기존의 RPCA 방법에 비해 정확도가 크게 향상되는 것을 확인할 수 있었다.

본 연구의 구성은 다음과 같다. 먼저 2장에서는 RPCA 방법에 대해 개괄하고 본 연구에서 제안하는 MRPCA의 방법론과 알고리즘에 대해 논의하였다. 3장에서는 RPCA 방법과 MRPCA 방법을 각각 활용하였을 때 추정치의 정확도를 비교하기 위한 모의 실험 결과를 정리하였는데, 자료가 다변량 정규 분포를 따르는 경우와 그렇지 않은 경우로 나누어서 각각 실험하였다. 4장에서는 결론을 맺고 향후 연구 과제를 밝혔다.

2. RPCA(Recursive PCA)와 MRPCA(Modified Recursive PCA)

2.1. RPCA(Recursive PCA)방법론 고찰

시점 t에서 관찰된 기대값이 영벡터인 n차원의 정상(stationary) 데이터를 x_t라고 하자. 시점 $t = 1 \sim k$까지 관찰된 데이터에 대한 표본 공분산 행렬 R_k는 아래와 같다.

$$R_k = \frac{1}{k}\sum_{i=1}^{k} x_t x_t^T = \frac{k-1}{k} R_{k-1} + \frac{1}{k} x_t x_t^T. \tag{2.1}$$

공분산 행렬 R_k에 대한 정규직교 고유벡터 행렬을 Q_k, 고유값을 대각원소로 하는 행렬을 Λ_k라고 한다면, 공분산 행렬 R_k와 R_{k-1}은 $R_k = Q_k \Lambda_k Q_k^T$와 $R_{k-1} = Q_{k-1}\Lambda_{k-1}Q_{k-1}^T$로 나타낼 수 있다. 이를 위 식 (2.1)에 대입하여 정리하면, 아래와 같이 Q_k와 Λ_k에 대한 반복 산출 공식을 도출할 수 있다.

$$\begin{aligned} Q_k(k\Lambda_k)Q_k^T &= Q_{k-1}\left((k-1)\Lambda_{k-1}\right)Q_{k-1}^T + x_k x_k^T & (2.2)\\ &= Q_{k-1}\left((k-1)\Lambda_{k-1} + Q_{k-1}^T x_k x_k^T Q_{k-1}\right)Q_{k-1}^T & (2.3)\\ &= Q_{k-1}\left((k-1)\Lambda_{k-1} + \alpha_k \alpha_k^T\right)Q_{k-1}^T, & (2.4) \end{aligned}$$

단, $\alpha_k = Q_{k-1}^T x_k$.

위 식 (2.4)의 우변 마지막 식에서 $((k-1)\Lambda_{k-1} + \alpha_k\alpha_k^T)$를 $V_k D_k V_k^T$로 스펙트럼 분해(Spectral Decompostion)할 수 있다면, 식 (2.4)는 아래와 같이 정리된다.

$$Q_k(k\Lambda_k)Q_k^T = Q_{k-1}\left(V_k D_k V_k^T\right)Q_{k-1}^T. \tag{2.5}$$

이제 식 (2.5)를 활용하면 시점 k에서의 공분산 행렬 R_k에 대한 고유벡터 행렬 Q_k와 고유값 대각행렬 Λ_k를 시점 $k-1$에서의 결과인 $R_{k-1}, Q_{k-1}, \Lambda_{k-1}$ 등을 이용하여 반복적으로 도출할 수 있게 된다. 즉, 시점 $k-1$에서의 공분산 행렬 R_{k-1}에 대한 고유벡터 행렬 Q_{k-1}과 고유값 대각행렬 Λ_{k-1}, 그리고 시점 k에서의 새로운 데이터 x_k가 주어지면, 아래 식 (2.6)과 식 (2.7)를 이용하여 Q_k와 Λ_k를 간단하게 구할 수 있는 것이다.

$$Q_k = Q_{k-1}V_k, \tag{2.6}$$

$$\Lambda_k = \frac{D_k}{k}, \tag{2.7}$$

단, $(k-1)\Lambda_{k-1} + \alpha_k\alpha_k^T = V_kD_kV_k^T,\ \alpha_k = Q_{k-1}^Tx_k$.

일반적으로 $((k-1)\Lambda_{k-1} + \alpha_k\alpha_k^T)$에 대한 고유값과 고유벡터를 직접 도출하는 것은 쉬운 일이 아니다. 그런데 Erdogmus 등 (2004)의 연구에 따르면 k가 상당히 큰 경우에는 일차 섭동 이론을 이용하여 $((k-1)\Lambda_{k-1} + \alpha_k\alpha_k^T)$의 고유값과 고유벡터를 반복적으로 추정할 수 있음을 밝혔다.

우선 편의상 $((k-1)\Lambda_{k-1}+\alpha_k\alpha_k^T)$에서 대각행렬인 $(k-1)\Lambda_{k-1}$을 Λ_{k-1}^*로 두자. k가 충분히 큰 경우에, $(\Lambda_{k-1}^* + \alpha_k\alpha_k^T)$는 대각행렬인 Λ_{k-1}^*의 영향이 지배적(diagonally dominant)이라 말할 수 있으며, 따라서 $(\Lambda_{k-1}^*+\alpha_k\alpha_k^T)$의 고유값 행렬 D_k는 Λ_{k-1}^*로, $(\Lambda_{k-1}^*+\alpha_k\alpha_k^T)$의 고유벡터 행렬 V_k는 Λ_{k-1}^*의 고유벡터에 해당하는 항등행렬(I)로 각각 근사하게 된다. 고유값과 고유벡터 행렬에 대한 미세한 변동을 반영할 수 있는 섭동행렬(perturbation matrix)를 P_Λ와 P_V를 도입하여 $D_k = \Lambda_{k-1}^* + P_\Lambda$로 $V_k = I + P_V$로 두면, $(\Lambda_{k-1}^* + \alpha_k\alpha_k^T)$에 대한 고유값 행렬 D_k과 고유벡터 행렬 V_k는 아래와 같이 근사할 수 있다.

$$V_kD_kV_k^T = (I+P_V)(\Lambda_{k-1}^* + P_\Lambda)(I+P_\Lambda)^T \tag{2.8}$$

$$= \Lambda_{k-1}^* + \Lambda_{k-1}^*P_\Lambda^T + P_\Lambda + P_\Lambda P_V^T + P_V\Lambda_{k-1}^* + P_V\Lambda_{k-1}^*P_V^T + P_VP_\Lambda + P_VP_\Lambda P_\Lambda^T \tag{2.9}$$

$$= \Lambda_{k-1}^* + P_\Lambda + DP_V^T + P_VD + P_V\Lambda_{k-1}^*P_V^T + P_VP_\Lambda P_V^T. \tag{2.10}$$

위 식 (2.10)에서 $P_V\Lambda_{k-1}^*P_V^T$와 $P_VP_\Lambda P_V^T$는 무시할 만큼 작다고 가정하고, 식 (2.5)에서와 같이 $(\Lambda_{k-1}^* + \alpha_k\alpha_k^T)$가 $V_kD_kV_k^T$로 근사한다는 것을 이용하면 아래를 보일 수 있다.

$$\alpha_k\alpha_k^T = P_\Lambda + D_kP_V^T + P_VD_k. \tag{2.11}$$

V_k는 정규직교 행렬이므로, $V_k = I + P_V$는 $V_kV_k^T = I$를 만족해야 하며, 추가적으로 $P_VP_V^T \approx 0$라고 가정하면 $P_V = -P_V^T$가 됨을 알 수 있다. 한편, P_Λ와 D_k는 태생적으로 대각 행렬이라는 것을 감안하면, 섭동행렬 P_Λ와 P_V는 아래와 같은 해를 갖게 된다.

$$P_\Lambda\text{의 } (i,i) \text{ 원소} = \alpha_i^2 \tag{2.12}$$

$$P_V\text{의 } (i,j) \text{ 원소} = \begin{cases} \dfrac{\alpha_i\alpha_j}{\lambda_j + \alpha_j^2 - \lambda_i - \alpha_i^2}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{2.13}$$

단, α_i는 α_k의 i번째 원소, λ_i는 $\Lambda_{k-1}^* = (k-1)\Lambda_{k-1}$의 i번째 대각 원소에 해당함.

위 방법에 의해 P_Λ와 P_V를 도출하고 나면, 시점 k의 고유값 행렬 Q_k와 고유벡터 행렬 Λ_k는 식 (2.6)과 식 (2.7)에 의해 아래와 같이 근사할 수 있다.

$$Q_k = Q_{k-1}(I+P_V), \tag{2.14}$$

$$\Lambda_k = \frac{1}{k}\left((k-1)\Lambda_{k-1} + P_\Lambda\right). \tag{2.15}$$

2.2. 수정된 RPCA(MRPCA; Modified Recursive PCA) 방법론 고찰

앞서 소개한 Erdogmus 등 (2004)의 RPCA 방법은 데이터가 추가될 때마다 표본 공분산 행렬의 고유값과 고유벡터를 반복적으로 추정할 수 있는 방법으로써, 실시간으로 대용량의 데이터를 처리해야 하는 분야에서 신속한 처리가 가능하며 행렬 연산에 따른 시스템의 과부하를 피할 수 있는 방법이라는 점에서 실용적이다. 그런데 이 RPCA 방법은 추가된 데이터의 수가 그리 많지 않은 경우에는 데이터가 추가되어 업데이트 될 때마다 표본 공분산 행렬을 매번 스펙트럼 분해해서 도출한 경우와 거의 차이가 나지 않지만, 추가된 데이터의 수가 많아짐에 따라 그 정확도가 떨어지는 것으로 알려져 있다. 본 연구에서는 이를 보완하기 위해 데이터가 상당히 많이 추가된다고 하더라도 공분산 행렬의 고유값과 고유벡터를 좀 더 안정적으로 정확하게 추정할 수 있는 방법인 MRPCA 방법을 제안하고자 한다.

RPCA 방법에 따르면 시점 $k-1$에서 공분산 행렬 R_{k-1}의 고유값 행렬 Λ_{k-1}과 고유벡터 행렬 Q_{k-1}이 주어졌을 때, 새로운 데이터가 추가되는 시점 k가 되면 위 식 (2.14)와 식 (2.15)에 의해 고유값 행렬과 고유벡터 행렬 모두 Λ_k와 Q_k로 업데이트된다. 그런데 공분산 행렬과 고유값, 고유벡터 행렬 간에는 아래 식 (2.16)과 같은 이론적인 관계가 존재하므로, RPCA 방법에 의해 고유값과 고유벡터 중 하나를 업데이트 하면 나머지 하나는 아래의 관계식을 이용하여 바로 도출할 수 있다는 것이 본 연구에서 제안하는 MRPCA 방법의 출발점이다. 이 방법 역시 RPCA 방법과 마찬가지로 데이터가 추가될 때마다 업데이트 된 공분산 행렬을 매번 스펙트럼 분해 하는 것 보다 효율적이며 RPCA 방법의 문제점이라고 할 수 있는 정확성까지 개선할 수 있는 방법이다.

$$Q_k\Lambda_k = R_kQ_k. \tag{2.16}$$

구체적으로 MRPCA 방법은 시점 $k-1$에서 공분산 행렬 R_{k-1}의 고유값 행렬 Λ_{k-1}과 고유벡터 행렬 Q_{k-1}이 주어졌을 때, 새로운 데이터 x_k가 추가되는 시점 k가 되면 기존의 RPCA 방법을 이용하여 고유벡터 행렬을 먼저 Q_k로 업데이트 한 뒤, 위 식 (2.16)의 이론적인 관계에 의해 Λ_k를 도출하자는 것이다. 기존의 RPCA 방법으로 고유값, 고유벡터 행렬을 업데이트 하는 경우 그 추정치들은 위 식 (2.16)을 만족한다는 것이 보장되지 않지만, MRPCA 방법의 경우는 업데이트 된 고유값, 고유벡터 행렬이 식 (2.16)을 항상 만족하므로 그 정확도가 향상될 것이라 예상할 수 있다. 한편, 이와 반대로 시점 k에서 공분산 행렬 R_k의 고유값 행렬 Λ_k를 RPCA 방법으로 먼저 업데이트 한 뒤 식 (2.16)의 관계에 의해 Q_k를 찾는 것도 생각할 수 있지만, 이 경우에는 식 (2.16)을 만족하는 Q_k의 해를 찾는 문제가 쉽지 않다는 단점이 있어 본 연구에서는 다루지 않기로 한다.

2.3. 수정된 RPCA(MRPCA; Modified Recursive PCA) 알고리즘

MRPCA 알고리즘은 표 2.1과 같이 요약할 수 있다. MRPCA 알고리즘은 표본 공분산 행렬의 고유벡터 행렬을 업데이트 하는 부분(표 2.1에서 2-4)에 해당함)을 제외하고 Erdogmus 등 (2004)의 RPCA 알고리즘과 동일하다.

3. 모의 실험 분석 및 결과

공분산 행렬의 고유값과 고유벡터 추정에 관한 RPCA 방법과 MRPCA 방법의 성과를 모의 실험을 통해 비교하고자 한다. 본 실험에서는 데이터의 확률적 속성이 정상적(stationary)인 경우를 다루며, 특히 분포의 종류에 따라 결과에 차이를 가져올 수 있으므로, 데이터가 다변량 정규 분포를 따르는 경우와 각 변수 별 주변 확률 분포가 제각기 다른 연속형 분포인 경우로 나누어 각각 살펴본다.

표 **2.1.** MRPCA(Modified Recursive PCA) 알고리즘 요약

1. 시점 1에서 $k-1$까지의 데이터로 도출된 공분산 행렬 R_{k-1}의 고유값, 고유벡터 행렬 Λ_{k-1}, Q_{k-1}이 주어짐.
2. 시점 k에 대해 아래의 알고리즘을 수행.
   1) 새로운 데이터 x_k를 얻음.
   2) $\alpha_k = Q_{k-1}^T x_k$를 계산.
   3) P_V와 P_Λ를 계산

$$P_\Lambda\text{의 } (i,i) \text{ 원소} = \alpha_i^2,$$

$$P_V\text{의 } (i,j) \text{ 원소} = \begin{cases} \dfrac{\alpha_i \alpha_j}{\lambda_j + \alpha_j^2 - \lambda_i - \alpha_i^2}, & i \neq j, \\ 0, & i = j, \end{cases}$$

   단, α_i는 α_k의 i번째 원소, λ_i는 $\Lambda^* = (k-1)\Lambda_{k-1}$의 i번째 대각 원소에 해당함.
   4) 고유벡터 행렬 Q_k를 추정

$$\tilde{Q}_k = Q_{k-1}(I + P_V).$$

   5) $\tilde{Q}_k$를 정규화

$$\hat{Q}_k = \tilde{Q}_k T_k,$$

   단, T_k : $\tilde{Q}_k$의 각 열 별 노름(norm)의 역행렬을 원소로 하는 대각행렬.
   6) 고유값 행렬 Λ_k을 추정

$$\hat{\Lambda}_k = \hat{Q}_k^T R_k \hat{Q}_k.$$

3.1. 다변량 정규 데이터에 대한 모의 실험

어떤 임의의 공분산 구조를 가지며 모평균이 영 벡터인 다변량 정규 분포로부터 데이터가 추출된다고 하자. 이 때 데이터가 하나씩 새로 추가될 때마다 공분산 행렬의 고유값과 고유벡터도 새롭게 추정되는데, 그 추정 방식을 RPCA 방법과 MRPCA 방법으로 각각 수행할 수 있을 것이다. 이렇게 추정된 고유값, 고유벡터 값들과 공분산 행렬을 직접 스펙트럼 분해해서 구한 추정값과 비교해 봄으로써 RPCA 방법과 MRPCA 방법 중 어느 것이 보다 정확한지에 대해 알아보고자 한다.

구체적으로 p차원의 모평균이 영 벡터, 모분산이 R인 다변량 정규분포로부터 10,000개의 난수를 생성한 뒤 이를 초기 데이터 셋으로 두자. 동일한 분포로부터 새로운 난수를 J개 더 생성하는데, 데이터가 하나씩 추가될 때마다 고유값 행렬 $L^j\,(j = 1, 2, \ldots, J)$과 고유벡터 행렬 $E^j\,(j = 1, 2, \ldots, J)$을 업그레이드 해 나간다. 이 때 추정하는 방식은 Erdogmus 등 (2004)의 RPCA 알고리즘과 본 연구에서 제안하는 MRPCA 알고리즘을 이용하는 경우, 그리고 이 두 알고리즘의 결과 비교를 위해 스펙트럼 분해를 이용하는 경우까지 세 방식으로 각각 수행한다. RPCA 방법인 경우에 추정된 고유값, 고유벡터 행렬을 $L_1^j, E_1^j\,(j = 1, 2, \ldots, J)$, MRPCA 방법인 경우에는 $L_2^j, E_2^j\,(j = 1, 2, \ldots, J)$, 스펙트럼 분해인 경우에는 $L_0^j, E_0^j\,(j = 1, 2, \ldots, J)$라고 두자.

L_0^j, L_1^j, L_2^j의 대각원소는 세 방식으로 추정된 고유값에 해당하므로, L_1^j과 L_0^j 또는 L_2^j과 L_0^j의 각 대각원소의 차이를 이용하면 RPCA 방법과 MRPCA 방법 중 어느 것이 더욱 정확한지를 말할 수 있을 것이다. 예를 들어 그림 3.1은 차원이 $p = 4$이고 공분산 구조가 사전에 정의되어 있는 어느 다변량 정규 데이터를 대상으로 이상의 과정을 실험한 결과이다. 10,000개의 데이터가 주어진 상황에서, $J = 500$개의 데이터를 추가해 가면서 4개의 고유값을 세 가지 방식으로 각각 추정하였을 때, RPCA 방

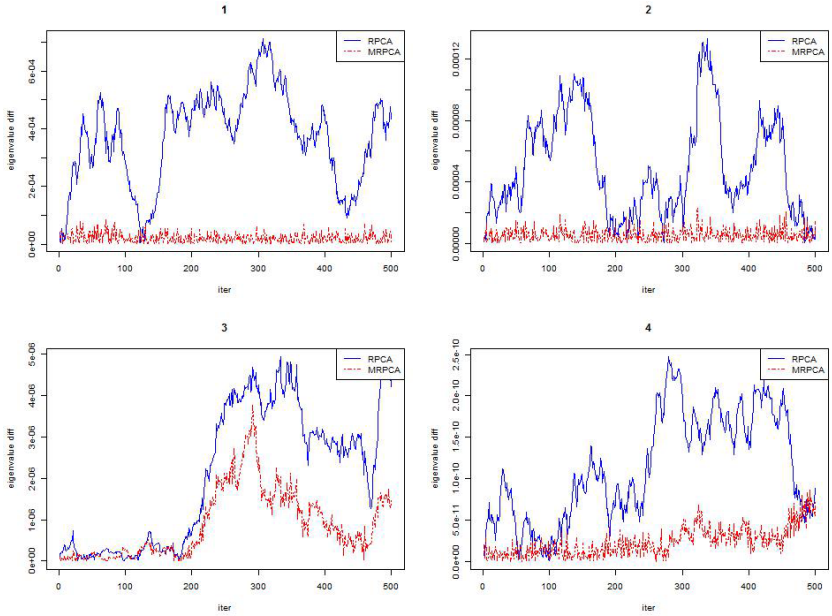

그림 3.1. 추정방식에 따른 고유값 추정치의 차이 비교

식과 스펙트럼 분해 방식 간에 추정된 고유값의 차이 (L_1^j의 d번째 대각 원소 – L_0^j의 d번째 대각 원소)와 MRPCA 방식과 스펙트럼 분해 방식 간에 추정된 고유값의 차이(L_2^j의 d번째 대각 원소 – L_0^j의 d번째 대각 원소)를 $j = 1, 2, \ldots, J(= 500)$, $d = 1, 2, \ldots, p(= 4)$에 대해 모두 도출한 것으로, 그림에 있는 1∼4의 라벨은 차원을 의미한다. MRPCA 방법의 경우는 500개의 데이터가 추가될 때까지 스펙트럼 분해로 구한 추정값과 의 차이가 0 근처에서 머물고 있는 반면 RPCA 방식은 점점 증가하는 추세라서 스펙트럼 분해로 구한 추정값과의 차이가 점점 커지고 있다는 것을 확인할 수 있다.

각 차원($d = 1, 2, \ldots, p$) 별 고유값 마다 아래와 같은 차이제곱합을 정의할 수 있는데, 이러한 정의에 의하면 각 차원 별로 SSE(RPCA)_d – SSE(MRPCA)_d의 값이 양인 경우에는 고유값을 추정함에 있어서 MRPCA 방법이 RPCA 방법에 비해 좀 더 스펙트럼 분해의 결과에 근사함을 의미하므로, 더 정확한 방법이라는 것을 의미한다. 또한 앞서 살펴본 과정을 한 회의 실험이라고 본다면, 이를 다양한 공분산 행렬 R_i, $i = 1, 2, \ldots, I$에 대하여 실험할 수 있으므로 이렇게 도출된 I개의 SSE(RPCA)_d^i – SSE(MRPCA)_d^i값으로 보다 일반적인 결론을 이끌어 낼 수 있을 것이다.

$$\text{SSE(RPCA)}_d^i = \sum_{j=1}^{J} \frac{\left(L_{1,i}^j\text{의 } d\text{번째 대각 원소} - L_{0,i}^j\text{의 } d\text{번째 대각 원소}\right)^2}{i^2}, \tag{3.1}$$

$$\text{SSE(MRPCA)}_d^i = \sum_{j=1}^{J} \frac{\left(L_{1,i}^j\text{의 } d\text{번째 대각 원소} - L_{0,i}^j\text{의 } d\text{번째 대각 원소}\right)^2}{i^2}, \tag{3.2}$$

단, 여기서 $d = 1, 2, \ldots, p$이다. 일반적으로 어떤 행렬에 대한 고유값, 고유벡터의 추정은 데이터의 차원이 얼마인지, 고유값 스프레드(eigenspread, 가장 큰 고유값과 가장 작은 고유값의 비율)가 얼마인지에 따라 그 정확도가 좌우되는 것으로 알려져 있다. 따라서 본 실험에서는 데이터의 차원이 2, 4, 8인 경우 각각에 대해 고유값 스프레드가 10^2정도로 작은 경우와 10^6정도로 큰 경우로 나눠서 그 결과에 차이

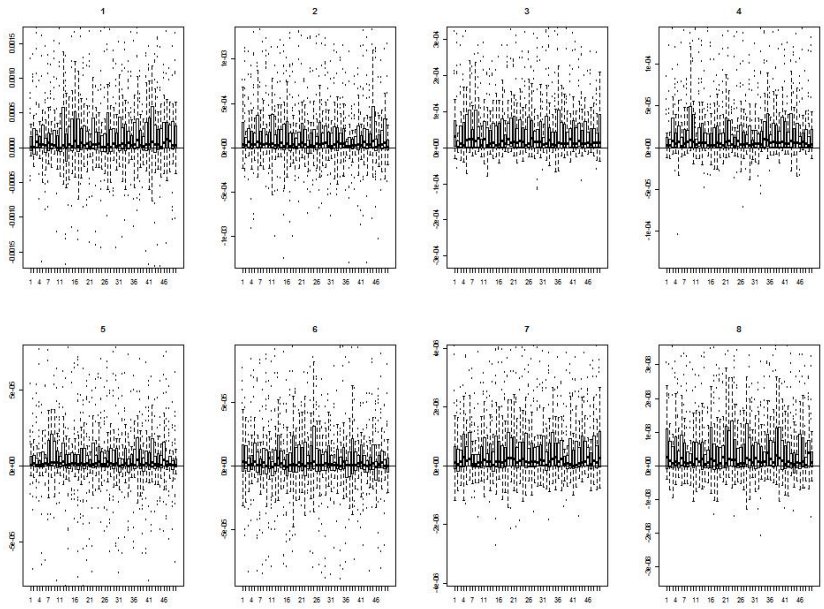

그림 **3.2.** $p = 8$, $\lambda = 10^2$인 경우의 각 차원 별 $\mathrm{SSE(RPCA)}_d - \mathrm{SSE(MRPCA)}_d$ 상자그림

가 있는가를 확인해 보았다.

또한 앞서 언급한 바와 같이 I회의 반복 실험을 위해서는 주어진 데이터 차원과 주어진 고유값 스프레드에 대하여 I개의 서로 다른 공분산 구조 R_i, $i = 1, 2, \ldots, I$를 만들어야 하는데 그 과정은 아래와 같다. 먼저 p차원의 데이터에 공분산에 대한 고유값 스프레드가 γ가 되는 공분산 행렬을 하나 찾은 뒤 이를 R_0로 두고, R_0에 대한 고유값 행렬을 L_0, 고유벡터 행렬을 E_0라고 두자. 아래 식에 의하면 고유값 스프레드가 γ로 일정하게 유지되는 I개의 공분산 행렬 R_i, $i = 1, 2, \ldots, I$를 만들 수 있다.

$$R_i = E_0(i \cdot \Lambda_0)E_0, \quad i = 1, 2, \ldots, I. \tag{3.3}$$

위 방법에 의해 p차원의 공분산에 대한 고유값 스프레드가 γ가 되는 공분산 행렬을 I개 만들면, 각각의 R_i, $i = 1, 2, \ldots, I$를 공분산 구조로 하는 다변량 정규 데이터를 만들어 낼 수 있을 것이다.

그림 3.2와 그림 3.3은 $p = 8$이면서 $\gamma = 10^2, 10^6$인 경우 각각에 대하여 먼저 R_0를 찾은 뒤, 이를 이용해서 생성된 $R_1, \ldots, R_{I(=50)}$으로 이상의 모의 실험(주어진 공분산 구조 R_i를 갖는 다변량 정규 분포로부터 10,000개의 초기 데이터 셋이 주어졌을 때 추가로 $J(= 500)$개의 데이터를 하나씩 더해 가면서 RPCA, MRPCA, 스펙트럼 분해 방식으로 각각 고유값을 업데이트한 뒤, RPCA 및 MRPCA 방법과 스펙트럼 분해 방법의 추정치 차이 제곱합 $\mathrm{SSE(RPCA)}_d^i$, $\mathrm{SSE(MRPCA)}_d^i$을 도출하는 과정)을 $i = 1, \ldots, I(= 50)$에 대해 모두 도출한 것이다.

또한 매 회 (i번째 모의실험, $i = 1, 2, \ldots, I(= 50)$) 마다 공분산 행렬은 R_i로 고정되어 있지만, 다변량 정규 분포로부터의 난수 추출 과정에서 오차가 발생하여 결과에 영향을 미칠 수 있다. 따라서 RPCA와 MRPCA 방법론에 따른 성능 차이를 좀 더 명확히 비교하기 위해 모의 실험을 매 회 (i번째 모의실험, $i = 1, 2, \ldots, I(= 50)$)마다 50번씩 다시 반복했으며, 그 결과인 $\mathrm{SSE(RPCA)}_d^i$, $\mathrm{SSE(MRPCA)}_d^i$도 매 회마다 50번씩 반복 계산하였다. 즉 아래 결과에서 하나의 상자 그림은 공분산 R_i가 주어졌을 때

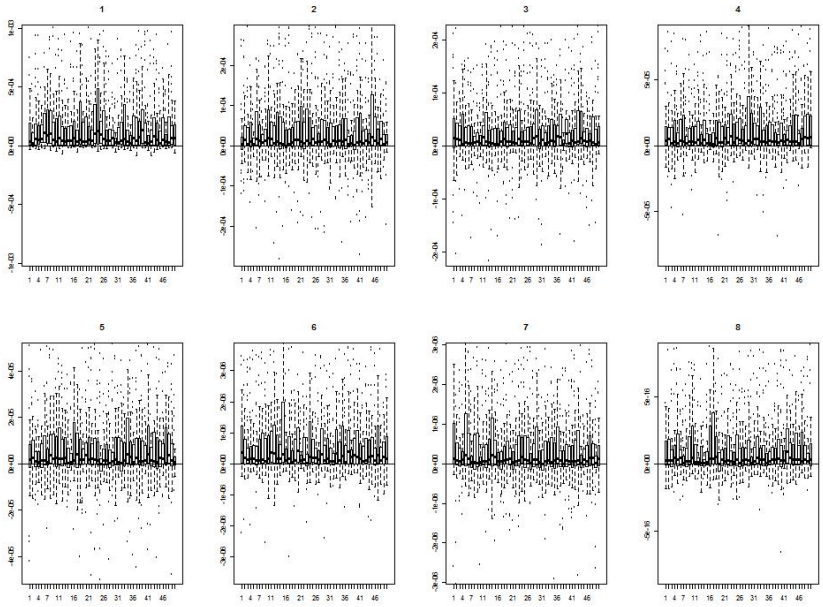

그림 **3.3.** $p = 8$, $\lambda = 10^6$인 경우의 각 차원 별 $\text{SSE(RPCA)}_d - \text{SSE(MRPCA)}_d$ 상자그림

의 모의실험을 반복한 결과인 50개의 $\text{SSE(RPCA)}_d^i - \text{SSE(MRPCA)}_d^i$를 의미하는 것이며, 이를 각 차원($d = 1, 2, \ldots, p$) 별로 매 회 ($i$번째 모의실험, $i = 1, 2, \ldots, I(= 50)$) 마다 도출한 것이다.

결과를 살펴보면 $p = 8$, $\gamma = 10^2$과 $p = 8$, $\gamma = 10^6$인 경우 모두 $\text{SSE(RPCA)}_d^i - \text{SSE(MRPCA)}_d^i$는 0보다 큰 값에 주로 분포하고 있는 것을 확인할 수 있으며, 이는 새로운 데이터가 추가됨에 따라 MRPCA 방법을 이용하여 고유값을 업데이트 하는 것이 RPCA 방법을 이용하는 것 보다 더욱 정확함을 의미한다. 또한 모의실험이 몇 번째 횟수(i)인지와 관계없이 일관된 결과를 보이고 있어서 공분산 행렬 R_i가 어떻게 주어지든지(추정해야 하는 고유값의 절대적인 수치가 얼마든지 관계없이), MRPCA 방법이 RPCA 방법에 비해 일관적으로 우월하다고 말할 수 있다. 한편 데이터의 차원 p가 2인 경우와 4인 경우에도 고유값 스프레드 $\gamma = 10^2, 10^6$인 각각에 대해 같은 실험을 수행해 보았는데 유사한 결과를 얻을 수 있었다.

E_0^j, E_1^j, E_2^j의 각 열은 세 방식으로 추정된 각 차원 별 고유벡터에 해당하는데, 고유벡터의 경우 RPCA 또는 MRPCA 방법으로 추정된 고유벡터와 비교 기준인 스펙트럼 분해로 추정된 고유벡터와의 각도를 이용하여 그 차이가 어느 정도인지를 측정할 수 있다. 그림 3.4는 앞서 살펴본 그림 3.1과 동일한 데이터로 실험한 결과를 고유벡터에 대해 살펴본 것이다. 즉 차원이 $p = 4$이고 공분산 구조가 사전에 정의되어 있는 어느 다변량 정규 데이터 10,000개가 주어진 상황에서, $J = 500$개의 데이터를 추가해 가면서 4개의 고유벡터를 세 가지 방식으로 각각 추정하였을 때, RPCA 방법과 스펙트럼 분해 방법 간에 추정된 고유벡터 간의 각도 $D_R(= E_1^j$의 d번째 열과 E_0^j의 d번째 열 간의 각도)와 MRPCA 방식과 스펙트럼 분해 방식 간에 추정된 고유벡터 간의 각도 차이 $D_{MR}(= E_{1,i}^j$의 d번째 열과 $E_{0,i}^j$의 d번째 열 간의 각도)를 $j = 1, 2, \ldots, J(= 500)$, $d = 1, 2, \ldots, p(= 4)$에 대해 모두 도출한 뒤, 이 둘의 차이 $D_R - D_{MR}$를 그린 것이다. 첫 번째와 두 번째 고유벡터의 경우는 특히 $D_R - D_{MR}$의 값이 전체적으로 0보다 큰 값을 가지는데, 이는 RPCA 방법으로 추정된 고유벡터보다 MRPCA 방법으로 추정한 경우가 스펙트럼 분해

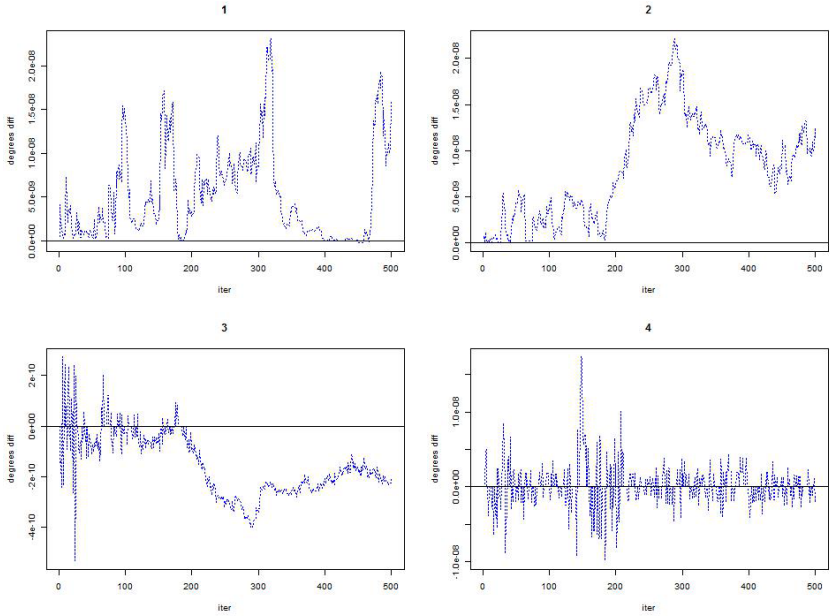

그림 3.4. 추정방식에 따른 고유벡터 추정치의 차이 비교

로 추정된 값과 이로는 각도가 작다는 것이므로, MRPCA 방법이 RPCA 방법보다 고유벡터를 좀 더 정확하게 추정한다는 것을 의미한다.

고유값과 마찬가지로 고유벡터에 대해서도 아래와 같이 각 차원($d = 1, 2, \ldots, p$) 별로 J개의 고유벡터 각도 D_R, D_{MR}를 합한 값을 $\mathrm{SD(RPCA)}_d^i$, $\mathrm{SD(MRPCA)}_d^i$로 각각 정의할 수 있는데, 이러한 정의에 의하면 각 차원 별로 $\mathrm{SD(RPCA)}_d^i - \mathrm{SD(MRPCA)}_d^i$의 값이 양인 경우에는 고유벡터를 추정함에 있어서 MRPCA 방법이 RPCA 방법에 비해 좀 더 스펙트럼 분해에 근사하여 더 정확한 방법이라는 것을 의미한다. 또한 이를 R_i, $i = 1, 2, \ldots, I$에 대해 각각 실험한 결과인 I개의 $\mathrm{SD(RPCA)}_d^i - \mathrm{SD(MRPCA)}_d^i$값으로 보다 일반적인 결론을 이끌어 낼 수 있을 것이다.

$$\mathrm{SD(RPCA)}_d^i = \sum_{j=1}^{J} \left(E_{1,i}^j\text{의 } d\text{번째 열과 } E_{0,i}^j\text{의 } d\text{번째 열 간의 각도} \right), \tag{3.4}$$

$$\mathrm{SSD(MRPCA)}_d^i = \sum_{j=1}^{J} \left(E_{2,i}^j\text{의 } d\text{번째 열과 } E_{0,i}^j\text{의 } d\text{번째 열 간의 각도} \right), \tag{3.5}$$

단, 여기서 $d = 1, 2, \ldots, p$이다. 그림 3.5와 그림 3.6은 $p = 8$이면서 $\gamma = 10^2, 10^6$인 경우 각각에 대하여 이상의 모의 실험을 $I = 50$회 반복 수행하되, 매 회($i(= 1, 2, \ldots, I(= 50))$번째 모의실험)마다 난수 추출에 따른 오차를 고려하여 다시 50번씩 반복 실험한 결과인 $\mathrm{SD(RPCA)}_d^i$, $\mathrm{SD(MRPCA)}_d^i$를 도출한 것이다. 즉 각 차원($d = 1, 2, \ldots, p$) 별로 매 회($i(= 1, 2, \ldots, I(= 50))$번째 모의실험)마다 50개의 $\mathrm{SD(RPCA)}_d^i$, $\mathrm{SD(MRPCA)}_d^i$에 대한 상자그림을 그린 것이라고 볼 수 있다. 고유값에 대한 실험과 마찬가지로 각 회당 추가된 데이터의 수는 $J = 500$으로 두었다.

결과를 살펴보면, $p = 8$, $\gamma = 10^2$과 $p = 8$, $\gamma = 10^6$인 경우 모두 $\mathrm{SD(RPCA)}_d^i - \mathrm{SD(MRPCA)}_d^i$는 거의 0에 가까운 값을 가지는 것을 알 수 있다. 이는 새로운 데이터가 추가됨에 따라 고유벡터를 업데

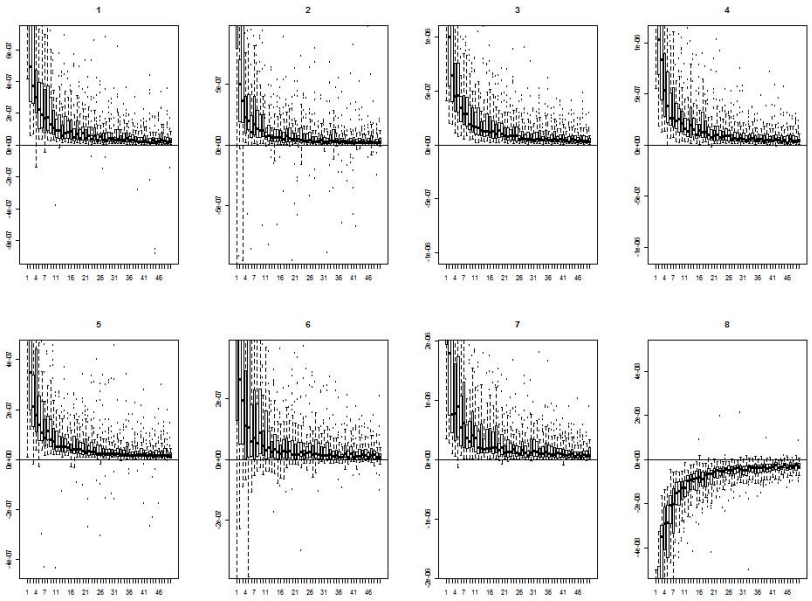

그림 **3.5.** $p = 8$, $\lambda = 10^2$인 경우의 각 차원 별 $\mathrm{SD(RPCA)}_d - \mathrm{SD(MRPCA)}_d$ 상자그림

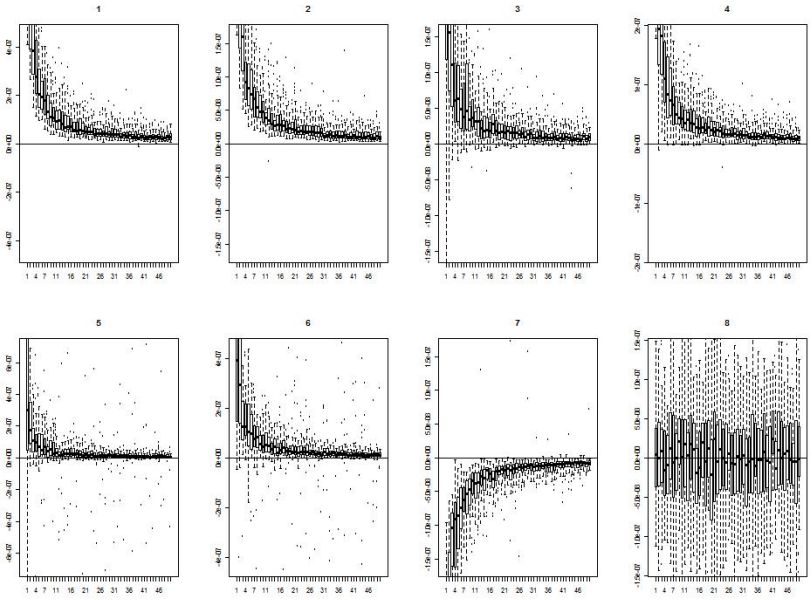

그림 **3.6.** $p = 8$, $\lambda = 10^6$인 경우의 각 차원 별 $\mathrm{SD(RPCA)}_d - \mathrm{SD(MRPCA)}_d$ 상자그림

이트함에 있어서 RPCA 방법과 MRPCA 방법이 크게 차이가 없음을 나타낸다. MRPCA 방법도 고유벡터를 추정하는 방법은 RPCA 방법과 동일하기 때문에 그 성능도 크게 차이가 나지 않는 것으로 해석할 수 있다. 약간의 차이가 생기는 이유는, MRPCA 방법의 경우 RPCA 방법과 고유값 추정치가 다른

표 **3.1.** 각 변수 별 주변(marginal) 확률 분포에 대한 정의

$p = 2$	chi-square (5), uniform (0, 10)
$p = 4$	chi-square (5), uniform (0, 10), exponential (10), normal (0, 5)
$p = 8$	chi-square (5), uniform (0, 10), exponential (10), normal (0, 5), uniform (0, 5), exponential (5), chi-square (10), uniform (0, 2)

데 이러한 고유값 추정치들이 다음 회차 추정에 대한 초기값이 되기 때문인 것으로 보인다. 고유벡터의 경우 한가지 특이한 점은 모의실험이 몇 번째 횟수(i)인지에 따라, 즉 공분산 행렬 R_i가 어떻게 주어지는가에 따라 정확도에 차이를 보이는데, 특히 공분산 행렬에 대한 고유값의 절대적인 수치가 작을수록 MRPCA 방법이 RPCA 방법보다 더욱 정확해진다는 것을 알 수 있다. 또한 예외적으로 마지막 고유벡터의 경우 RPCA 방법이 MRPCA 방법에 비해 더 정확한 것을 알 수 있지만, 차이의 절대적인 정도는 그래프 상의 한 눈금이 10^{-8}단위인 것을 감안하면 아주 작은 편이라고 볼 수 있겠다. 한편, 데이터의 차원 p가 2인 경우와 4인 경우에도 $\gamma = 10^2, 10^6$일 때 각각에 대해 모두 동일한 실험을 수행해 보았는데 유사한 결과를 얻을 수 있었다.

3.2. 다변량 정규분포를 따르지 않는 데이터에 대한 모의 실험

임의의 정해진 공분산 구조를 가지며 모평균이 영 벡터인 다변량 데이터지만, 그 분포는 다변량 정규분포를 따르지 않고 각 변수 별로 주변 확률 분포가 표 3.1과 같이 다양하게 정의된 경우에 대하여 전과 동일한 과정의 모의 실험을 수행하였다. 즉 표 3.1에서 정의된 확률적 구조에서 데이터가 하나씩 새로 추가될 때마다 공분산 행렬의 고유값과 고유벡터도 RPCA 방법과 MRPCA 방법으로 각각 추정한 뒤, 이들 값과 공분산 행렬을 직접 스펙트럼 분해해서 구한 고유값, 고유벡터 추정값과 비교해 봄으로써 RPCA 방법과 MRPCA 방법 중 어느 것이 보다 정확한지에 대해 알아보고자 한다. 이 때 공분산 구조가 R_i이면서 조건을 만족하는 데이터를 생성하기 위해서는 우선 표 3.1에서 정의된 분포에서 각각 독립적으로 난수를 추출한 뒤, 다변량 표준화 과정(Mahalanobis Transformation)을 거쳐 공분산이 항등행렬(I)이고 모평균이 영벡터인 데이터로 변환해 주어야 한다. 여기서 공분산 행렬 R_i를 $R_i = AA^T$로 콜레스키 분해(Cholesky decompostion)해서 얻은 행렬 A에 변환된 데이터를 곱해주면 모평균이 영벡터이면서 공분산 구조가 R_i인 데이터를 얻을 수 있다.

그림 3.7과 그림 3.8은 다변량 정규 분포 데이터에 대한 실험과 마찬가지로 $p = 8$일 때 $\gamma = 10^2, 10^6$인 경우 각각에 대하여 각 차원($d = 1, 2, \ldots, p$) 별로 모의 실험을 $I = 50$회 반복 수행하되, 매 회(i번째 모의실험, $i = 1, 2, \ldots, I(= 50)$)마다 난수 추출에 따른 오차를 고려하여 다시 50번씩 반복하여 실험한 결과인 50개의 $\mathrm{SSE(RPCA)}_d^i - \mathrm{SSE(MRPCA)}_d^i$를 상자 그림으로 도출한 것이다. 각 회당 추가된 데이터의 수도 전과 동일하게 $J = 500$으로 두었다.

데이터의 확률적인 구조가 다변량 정규가 아니라고 하더라도 $p = 8$, $\gamma = 10^2$인 경우와 $p = 8$, $\gamma = 10^6$인 경우 모두 $\mathrm{SSE(RPCA)}_d^i - \mathrm{SSE(MRPCA)}_d^i$는 0보다 큰 값에 대부분 분포하고 있으며, 이를 통해 MRPCA 방법으로 고유값을 업데이트 하는 것이 RPCA 방법을 이용하는 것 보다 정확하다는 것을 알 수 있다. 또한 공분산 행렬 R_i가 어떻게 주어지든 관계없이 MRPCA 방법이 RPCA 방법에 비해 일관적으로 정확도가 높은 것으로 보인다. 데이터의 차원 p가 2인 경우와 4인 경우에도 $\gamma = 10^2, 10^6$일 때 각각에 대해 같은 실험을 수행해 보았는데, 유사한 결과를 얻을 수 있었다.

그림 3.9와 그림 3.10도 $p = 8$일 때 $\gamma = 10^2, 10^6$인 경우 각각에 대하여 각 차원($d = 1, 2, \ldots, p$) 별로 동일한 모의 실험을 $I = 50$회 반복 수행하되, 매 회(i번째 모의실험, $i = 1, 2, \ldots, I(= 50)$)마다 난수 추출에 따른 오차를 고려하여 다시 50번씩 반복 실험한 결과인 50개의 $\mathrm{SD(RPCA}_d^i - \mathrm{SD(MRPCA)}_d^i$에 대

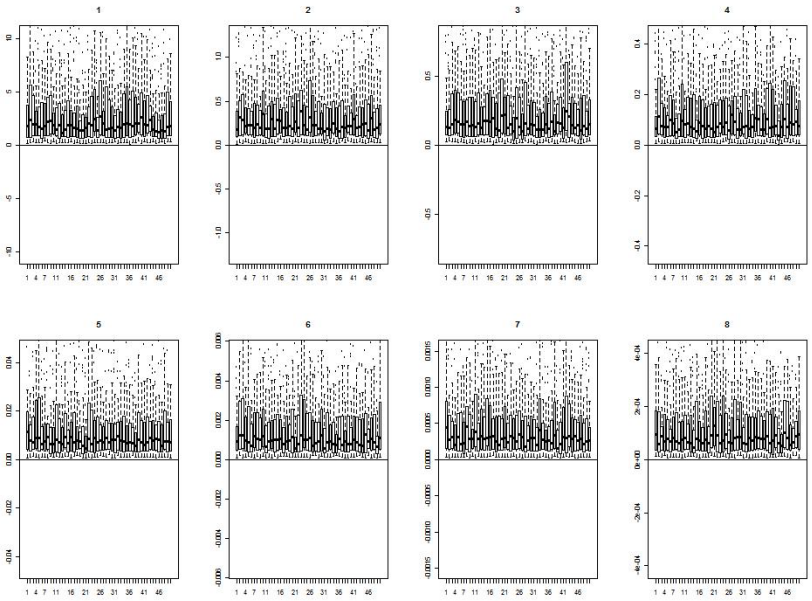

그림 **3.7.** $p = 8$, $\lambda = 10^2$인 경우의 각 차원 별 $\text{SSE(RPCA)}_d - \text{SSE(MRPCA)}_d$ 상자그림

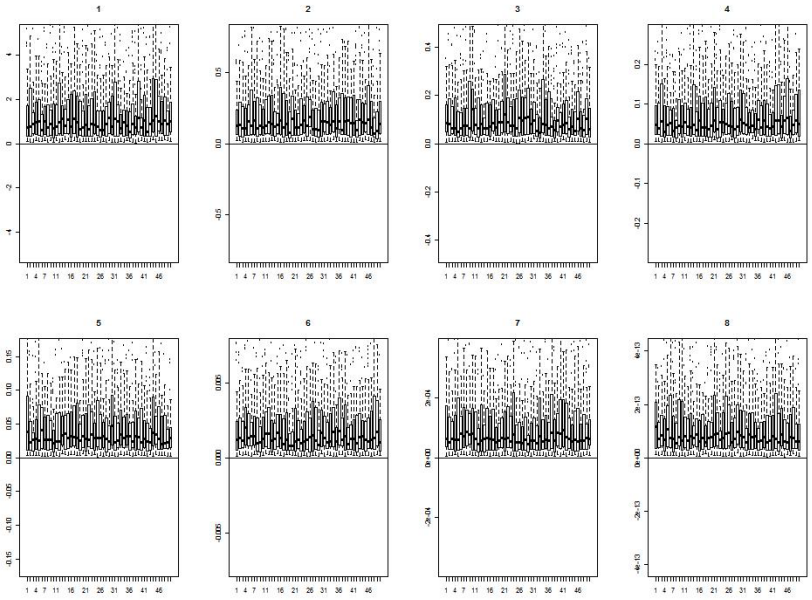

그림 **3.8.** $p = 8$, $\lambda = 10^6$인 경우의 각 차원 별 $\text{SSE(RPCA)}_d - \text{SSE(MRPCA)}_d$ 상자그림

한 상자 그림을 그린 것이다. $p = 8$, $\gamma = 10^2$인 경우와 $p = 8$, $\gamma = 10^6$인 경우 모두 $\text{SD(RPCA)}_d^i - \text{SD(MRPCA)}_d^i$는 거의 0에 가까운 값을 가지는 것을 알 수 있다. 또한 다변량 정규인 데이터에서와 마찬가지로, 공분산 행렬에 대한 고유값의 절대적인 수치가 작을수록 MRPCA 방법이 RPCA 방법보

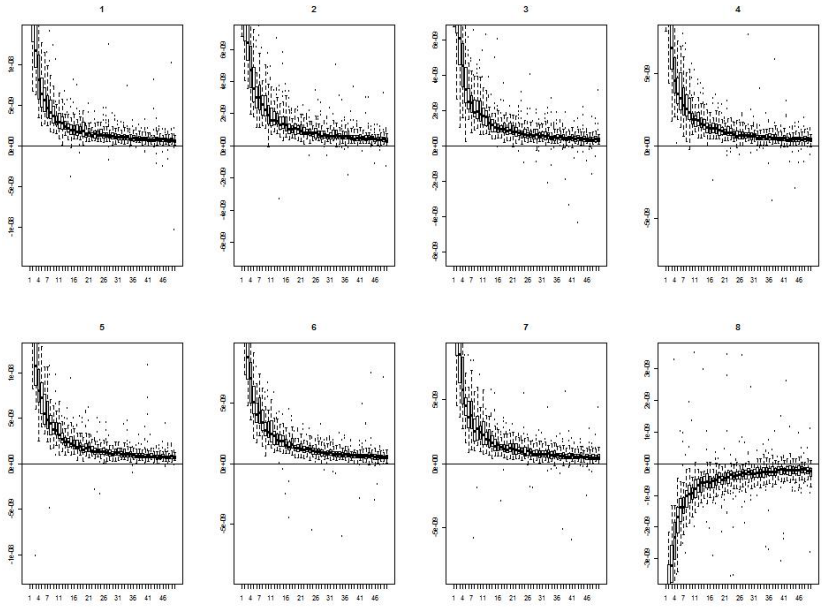

그림 **3.9.** $p = 8$, $\lambda = 10^2$인 경우의 각 차원 별 $\mathrm{SD(RPCA)}_d - \mathrm{SD(MRPCA)}_d$ 상자그림

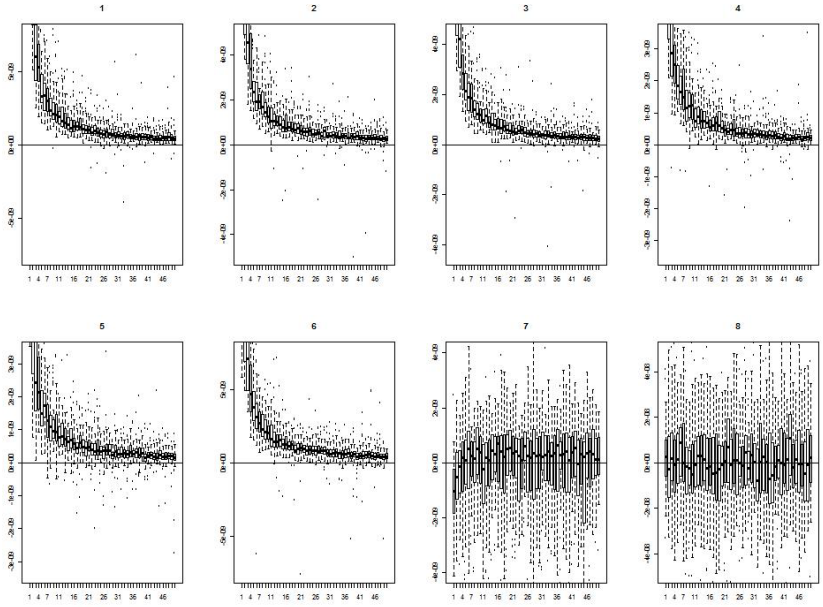

그림 **3.10.** $p = 8$, $\lambda = 10^6$인 경우의 각 차원 별 $\mathrm{SD(RPCA)}_d - \mathrm{SD(MRPCA)}_d$ 상자그림

다 더욱 정확해진다는 사실을 다시 한번 확인할 수 있다. 데이터의 차원 p가 2인 경우와 4인 경우에도 $\gamma = 10^2, 10^6$일 때 각각에 대해 모두 같은 실험을 수행해 보았는데 유사한 결과를 얻을 수 있었다.

4. 결론

본 연구에서는 PCA 분석에서 공분산 행렬의 고유값과 고유벡터를 추정하기 위한 방법론의 하나인 Erdogmus 등 (2004)의 RPCA 방법에 대해 고찰하고, 그 정확성을 좀 더 개선할 수 있는 MRPCA 방법을 제안하였다. 또한 두 방법론에 대한 비교를 위해 모의 실험을 수행하였는데, 데이터가 다변량 정규분포로부터 추출된 경우와 그렇지 않은 경우 모두 MRPCA 방법을 활용한 경우가 추정치가 더욱 정확해졌음을 확인할 수 있었다. RPCA 방법은 특히 대용량의 데이터가 실시간으로 업데이트되어 매 순간 PCA 분석을 시행해야 하는 경우에 고유값과 고유벡터를 신속하게 추정하지만, 추가된 데이터의 양이 많아짐에 따라 정확도가 떨어질 수 밖에 없다는 한계가 있다. 본 연구에서 제안한 MRPCA 방법은 고유값과 고유벡터 추정에 있어 연산 과정이 짧고 신속할 뿐만 아니라, 추가된 데이터가 많아져도 RPCA 방법에 비해 정확하다는 점에서 그 의미를 찾을 수 있다.

한편, RPCA 방법과 마찬가지로 MRPCA 방법도 데이터의 변화가 반영된 고유값과 고유벡터 값을 일차 섭동 이론을 활용하여 추정하는 방법이기 때문에, 추가된 데이터의 비중이 늘어감에 따라 오차가 필연적으로 커질 수밖에 없다. 따라서 어느 시점에서 추정을 멈추는 것이 좋은 것인가에 대한 연구가 필요하다. 또한 앞서 밝힌 바와 같이 본 연구는 데이터가 정상적인 경우에 대해서만 논하였는데, 현실적으로 많은 경우 데이터의 확률적인 속성이나 공분산 구조가 시간이 지남에 따라서 변하므로 이에 대한 추가적인 분석도 필요할 것으로 보인다.

참고문헌

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, Wiley, New York.

Erdogmus, D., Rao, Y. N., Peddaneni, H., Hegde, A. and Principe, J. C. (2004). Recursive principal components analysis using eigenvector matrix perturbation, *EURASIP Journal on Applied Signal Processing*, **2004**, 2034–2041.

Golub, G. and Loan, C. V. (1993). *Matrix Computation*, Johns Hopkins University Press, Baltimore, MD.

Kung, S. Y., Diamantaras, K. I. and Taur, J. S. (1994). Adaptive principal component extraction(APEX) and applications, *IEEE Transaction Signal Processing*, **42**, 296–317.

Modified Recursive PCA

Donggyu Kim[1] · Ahhyoun Kim[2] · HyunJoong Kim[3]

[1]Department of Applied Statistics, Yonsei University
[2]Department of Applied Statistics, Yonsei University
[3]Department of Applied Statistics, Yonsei University

(Received May 2011; accepted September 2011)

Abstract

PCA(Principal Component Analysis) is a well-studied statistical technique and an important tool for handling multivariate data. Although many algorithms exist for PCA, most of them are unsuitable for real time applications or high dimensional problems. Since it is desirable to avoid extensive matrix operations in such cases, alternative solutions are required to calculate the eigenvalues and eigenvectors of the sample covariance matrix. Erdogmus *et al.* (2004) proposed Recursive PCA(RPCA), which is a fast adaptive on-line solution for PCA, based on the first order perturbation theory. It facilitates the real-time implementation of PCA by recursively approximating updated eigenvalues and eigenvectors. However, the performance of the RPCA method becomes questionable as the size of newly-added data increases. In this paper, we modified the RPCA method by taking advantage of the mathematical relation of eigenvalues and eigenvectors of sample covariance matrix. We compared the performance of the proposed algorithm with that of RPCA, and found that the accuracy of the proposed method remarkably improved.

Keywords: Recursive PCA, Principal Component Analysis, first order perturbation theory.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No.2010-0008769).
[2]Corresponding author: Doctoral student, Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. E-mail: ahyun1003@yonsei.ac.kr