마코프 랜덤 필드 하에서 정규혼합모형에 의한 다중 결측값 대체기법: 색조영상 결측 화소값 대체에 응용

김승구^{1,a}

^a상지대학교 컴퓨터데이터정보학

요약

자료의 독립성 가정 하에서 EM 알고리즘에 의한 결측치 대체(imputation of missing values) 기법은 잘 알려져 있다. 그러나 공간자료를 다루는 응용문제에서는 독립성 가정이 확장된 마코프 랜덤 필드(Markov random field; MRF) 하에서 다루어져야 할 것이다. 이에 본 논문에서는 마코프 랜덤 필드 모형 하에서 다변량 자료 중에 다중의 결측치의 대체를 위한 EM 알고리즘을 제공한다. 이 기법은 몇 가지 현실적 가정 하에서 결국 혼합모형에 의한 대체 기법임을 보인다. 그리고 제공된 기법으로 3-변량으로 구성된 색조영상(color image)의 결측화소값 대체문제에 적용하여 그 유용성과 문제점을 밝히며, 문제점의 개선방안에 대해 논의한다.

주요용어: 다중 결측, 결측치 대체, 마코프 랜덤 필드, EM 알고리즘, 색조영상.

1. 서론

관측된 자료를 바탕으로 결측치(missing value)를 대체(imputation)하는 방법은 아주 오랜 역사만큼이나 수많은 방법들이 존재한다. Little과 Rubin (2002)과 Marker 등 (2002)은 이 문제를 다룬 가장 대표적인 저서라 하겠다. 과거 결측치 대체 기법은 모수적 기법보다는 핫-덱(hot-deck) 기법과 같은 비모수적 방법이 대세를 이루었다. 예를 들면, Kalton과 Kasprzyk (1986)의 NND(nearest neighbor donor) 기법이 이러한 종류라 할 것이다. 모수적 기법이 문제를 단순화 시키기는 하지만 결측치의 대체값이 가정된 분포에 절대적으로 의존하기 때문이라 할 것이다. 그러나 최근에는 혼합모형(특히 정규혼합모형)에 기반한 방법이 대세를 이루고 있는 것으로 보인다. 그 이유는 정규혼합모형으로 거의 대부분의 경험분포를 잘 근사시킬 수 있기 때문일 것이다 (McLachlan과 Peel, 2000). 비교적 최근 Hunt와 Jorgensen (2003)은 정규혼합모형에 의한 다중 결측치 대체법을 다루었다. Zio 등 (2007)은 (비록 독립성 가정 하에서이지만) 정규혼합모형(normal mixture model)에 의한 대체법을 모수적 기법과 비모수적 기법이 결합된 방법이라고 언급하며, 다양한 실험을 통해 이 방법에 대해 매우 우호적으로 보고하고 있다.

많은 연구들이 정규혼합모형에 의한 결측치 대체를 다루고 있지만 대부분 관측치들의 독립성을 가정한 것이다. 그러나 공간자료를 다루는 응용문제에서는 독립성 가정이 확장된 마코프 랜덤 필드(Markov random field; MRF) 하에서 다루어져야 할 것이다. MRF 기반 대체기법으로써 유전자 군집을 목적으로 한 Blanchet과 Vignes (2009)와 CCD 카메라 영상의 화질 개선을 위해 ICM(iterative conditional mode) 알고리즘에 기반한 Dass와 Nair (2003)의 연구를 눈여겨 볼 만하다. 본 연구에서는 제공하는 내용은 이들이 추구하는 목적과 크게 다르지는 않다. 다만, 본 연구에선 HMRP(hidden MRF) 하에서의 대체 기법과 다변량 정규혼합모형 기반 대체기법의 관계를 명료하게 정리할 것이다. 결론적

¹ (220-702) 강원도 원주시 우산동 660 상지대학교 컴퓨터데이터정보학과, 교수. E-mail: sgukim@sangji.ac.kr

으로 말해 HMRF 하에서의 대체 기법은 자연스럽게 정규혼합모형에 의한 기법을 의미하게 된다. 그 이유는 HMRF 모형 하에서의 추정 문제는 몇 가지 현실적인 가정 하에서 유한 정규혼합모형의 적합 문제와 동일하기 때문인데, 다음 절에서 자세히 다룰 것이다.

본 논문의 4절에서는 3절에서 유도한 HMRF 기반 정규혼합모형을 색조영상(color image)의 결측 화소값 대체에 응용한다. 결측화소값 대체에 관한 관련연구로서 Ogawa 등 (2006) 및 Kim 등 (2006)을 주목해 볼 만하다. 그러나 이 기법들은 MRF를 기반으로 하지 않은 방법이라 할 수 있다. 본 연구에서는 결측화소의 예측값을 구할 때 주변화소들과의 평활을 제안하였는데, 이것이 색조영상에 잘 적응함을 실험을 통해 보였다. 마지막으로 5절에서는 결론과 논의점을 정리하였다.

2. 마코프 랜덤 필드 모형

n 개의 화소들로 이루어진 주어진 다차원 관측영상 $\mathbf{y}$ 의 각 화소는 p -변량 벡터 $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^T$ ($i = 1, \dots, n$)로서 관측된다 하자. 단, 이후로 윗첨자 T 는 행렬의 전치를 나타낸다. n 개의 화소들은 모두 $\mathcal{G} = \{1, \dots, g\}$ 의 상태공간을 가지고 있으며, 관측치 $\mathbf{y}_i$ 가 상태 $h (\in \mathcal{G})$ 하에서 관측되었을 때 확률변수 $Z_{ih} = (\mathbf{Z}_i)_h = 1$, 그렇지 않으면 0이라 하자. 이때 $\mathbf{Z}$ 는

$$p(\mathbf{z}; \beta) = \frac{1}{C(\beta)} \exp\{-U(\mathbf{z}; \beta)\} \quad (2.1)$$

와 같은 깁스 밀도(Gibbs density)를 가지는 MRF라 하자. 여기서 $U(\cdot)$ 은 어떤 에너지 함수(energy function)를 나타내고, $C(\beta)$ 는 어떤 상수 $\beta (\geq 0)$ 에 대한 깁스 밀도의 정규화 상수항을 나타낸다. 그런데 대다수의 문헌들과 같이 본 연구에서도 관측치들의 마코프 특성(국소적 상관성)을 반영하면서도 추정을 용이하게 하기 위해 식 (2.1)에 대해

$$p(\mathbf{z}; \beta) \approx \prod_{i=1}^n p(\mathbf{z}_i | \mathbf{z}_{\partial_i}; \beta) = \prod_{i=1}^n \prod_{h=1}^g \prod_{i_h} \pi_{ih}^{z_{ih}} \quad (2.2)$$

의 근사를 가정하기로 한다 (Besag, 1975; Qian과 Titterton, 1992). 여기서, $\pi_{ih} = \text{pr}\{Z_{ih} = 1 | \mathbf{z}_{\partial_i}, \beta\}$ 를 나타내며, ∂_i 는 주어진 근방계 ∂ 에 대응한 화소 i 에서의 이웃화소들을 의미하고 $\mathbf{z}_{\partial_i}$ 은 그에 대응한 상태 값들을 나타낸다 (4절의 영상실험에서는 ∂ 로서 1차 근방계를 사용하였다). 그리고 Besag (1986)에서와 같이

$$\pi_{ih} = \pi_h(\mathbf{Z}_i | \mathbf{z}_{\partial_i}; \beta) \stackrel{\text{def}}{=} \frac{\exp(\beta u_{ih})}{\sum_{t=1}^g \exp(\beta u_{it})}, \quad h = 1, \dots, g; i = 1, \dots, n \quad (2.3)$$

로서 정의한다. 이때 u_{ih} 는 i 번째 화소의 이웃 중에 상태가 h 인 화소의 개수를 나타낸다. 또한 관측치 $\mathbf{y}_1, \dots, \mathbf{y}_n$ 들은 다변량 정규분포를 따르며, $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$ 이 주어졌을 때 조건부 독립 임을 가정한다. 즉,

$$f(\mathbf{y} | \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{z}_i; \boldsymbol{\theta}) \quad (2.4)$$

단,

$$f(\mathbf{y}_i | \mathbf{z}_i; \boldsymbol{\theta}) = \prod_{h=1}^g \phi(\mathbf{y}_i; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)^{z_{ih}} \quad (2.5)$$

이며, μ_h 은 $(p \times 1)$ 성분-평균벡터 그리고 Σ_h 은 $(p \times p)$ 성분-공분산 행렬을 나타낸다. 그리고 $\phi(\cdot)$ 은 다변량 정규분포를 나타내고, θ 는 $\{\mu_h\}$ 들과 $\{\Sigma_h\}$ 들을 포함하는 모수벡터이다.

이상의 가정 하에서 본 논문에서는 β 를 주어진 상수로서 취급할 것이며, 이때 결합 실현치 $(\mathbf{y}^T, \mathbf{z}^T)^T$ 에 대한 로그-우도는

$$\begin{aligned} \ln L(\theta; \mathbf{y}, \mathbf{z}) &= \ln p(\mathbf{z}; \beta) f(\mathbf{y}|\mathbf{z}; \theta) = \ln \prod_{i=1}^n \prod_{h=1}^g p(\mathbf{z}_i | \mathbf{z}_{\partial_i}; \beta) f(\mathbf{y}_i | \mathbf{z}_i; \theta) \\ &\propto \sum_{i=1}^n \sum_{h=1}^g z_{ih} \{ \ln \pi_{ih} + \ln \phi(\mathbf{y}_i; \mu_h, \Sigma_h) \} \end{aligned} \quad (2.6)$$

과 같이 나타낼 수 있다.

그런데 정규혼합모형

$$f(\mathbf{y}_i; \Theta) = \sum_{h=1}^g \pi_{ih} \phi(\mathbf{y}_i; \mu_h, \Sigma_h), \quad i = 1, \dots, n \quad (2.7)$$

의 EM 알고리즘에 의한 적합을 위해, 관측치 $\mathbf{y}_i$ 가 h 번째 성분으로부터의 표본이면 $(\mathbf{z})_i = \mathbf{z}_{hi} = 1$ 그렇지 않으면 0을 나타내는 지시변수라 놓고 $\mathbf{z}$ 를 결측자료 그리고 $(\mathbf{y}^T, \mathbf{z}^T)^T$ 를 완비자료로 설정하는데, 이때 완비자료의 로그-우도가 바로 식 (2.6)이다. 여기서 Θ 는 θ 에 $\{\pi_h\}$ 들을 추가하여 포함한 모수벡터를 나타낸다. 결국 MRF $\mathbf{Z}$ 하에서 모수의 추정 과정은 곧 EM 알고리즘에 의한 g -성분 정규혼합모형 적합의 문제로 볼 수 있는 것이다.

다음 절에서 자세히 소개되겠지만, 관측치 $\mathbf{y}_i$ 의 일부분의 원소에 미관측 결측치가 포함되어 있을 때 식 (2.7)의 혼합모형 적합을 통해 결측치의 대체가 이루어지게 된다. 따라서 앞서 언급한 가정들을 고려한다면, MRF 모형 하에서 미관측 결측치 대체 기법은 “자체적으로” 정규혼합모형에 의한 대체 기법이라 할 수 있을 것이다.

3. EM 알고리즘

i 번째 화소의 p -변량 관측치 $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^T (i = 1, \dots, n)$ 에서 원소들의 배치 순서는 매우 중요하다. 그러나 본 논문의 표기의 편의성을 위해 처음 p_i 개 원소는 관측치이며, 그 다음 $p - p_i$ 개의 원소는 미관측된 결측치라 하고, 각각 $\mathbf{y}_i^{\text{obs}}$ 와 $\mathbf{y}_i^{\text{mis}}$ 으로 나타내자. 즉, j 번째 화소의 관측벡터는

$$\mathbf{y}_i = (\overbrace{y_{i1}, \dots, y_{ip_i}}^{\text{obs}}, \overbrace{y_{ip_i+1}, \dots, y_{ip}}^{\text{mis}})^T = (\mathbf{y}_i^{\text{obs}T}, \mathbf{y}_i^{\text{mis}T})^T, \quad i = 1, \dots, n$$

로 나타내자. 이렇게 해도 본 논문에서 일반성을 잃지 않는다.

이제 우리는 $(\mathbf{y}^{\text{mis}T}, \mathbf{z}^T)$ 를 결측자료 그리고 $(\mathbf{y}^{\text{obs}T}, \mathbf{y}^{\text{mis}T}, \mathbf{z}^T)$ 를 완비자료로 하여 식 (2.7)의 혼합모형을 적합하는 문제를 다루어야 한다. 식 (2.6)로부터 완비자료의 로그-우도는

$$\begin{aligned} \ln L_c(\Theta) &= \sum_{i=1}^n \sum_{h=1}^g z_{ih} \{ \ln \pi_{ih} + \ln \phi(\mathbf{y}_i; \mu_h, \Sigma_h) \} \\ &\propto \sum_{i=1}^n \sum_{h=1}^g z_{ih} \left\{ \ln \pi_{ih} - \frac{1}{2} \ln |\Sigma_h| - \frac{1}{2} (\mathbf{y}_i - \mu_h)^T \Sigma_h^{-1} (\mathbf{y}_i - \mu_h) \right\} \\ &= \sum_{i=1}^n \sum_{h=1}^g z_{ih} \left\{ \ln \pi_{ih} - \frac{1}{2} \ln |\Sigma_h| \right\} - \frac{1}{2} \sum_{i=1}^n \sum_{h=1}^g \left\{ z_{ij} \mathbf{y}_i^T \Sigma_h^{-1} \mathbf{y}_i - z_{ij} \mathbf{y}_i^T \Sigma_h^{-1} \mu_i - z_{ij} \mu_i^T \Sigma_h^{-1} \mathbf{y}_i + z_{ij} \mu_i^T \Sigma_h^{-1} \mu_i \right\} \end{aligned} \quad (3.1)$$

와 같이 주어진다.

EM 알고리즘은 $(k+1)$ 번째 반복단계에서 불완비자료 $\mathbf{y}^{\text{obs}}$ 에 대한 완비자료의 로그-우도의 조건부 기대값

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}) = E[\ln L_c(\boldsymbol{\Theta})|\mathbf{y}^{\text{obs}}, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] \quad (3.2)$$

의 최대화를 반복함으로써 $\boldsymbol{\Theta}$ 의 (근사적인) MAP(maximum a posteriori) 추정치를 유도하게 된다.

앞으로 주어진 h 번째 성분-평균 벡터 $\boldsymbol{\mu}_h$ 와 성분-공분산 행렬 $\boldsymbol{\Sigma}_h$ 의 $\mathbf{y}_i^{\text{obs}}$ 와 $\mathbf{y}_i^{\text{mis}}$ 의 크기에 따른 분할을

$$\boldsymbol{\mu}_h = \begin{pmatrix} \boldsymbol{\mu}_{ih}^{\text{obs}} \\ \boldsymbol{\mu}_{ih}^{\text{mis}} \end{pmatrix}, \quad \boldsymbol{\Sigma}_h = \begin{pmatrix} \boldsymbol{\Sigma}_{ih}^{\text{oo}} & \boldsymbol{\Sigma}_{ih}^{\text{om}} \\ \boldsymbol{\Sigma}_{ih}^{\text{mo}} & \boldsymbol{\Sigma}_{ih}^{\text{mm}} \end{pmatrix}$$

과 같이 표시하자 ($\boldsymbol{\mu}_h$ 와 $\boldsymbol{\Sigma}_h$ 이 관측치 i 에 따라 서로 다를 수 있음을 나타내는 것이 아님을 유의하자).

이때 EM 알고리즘은 M-step에서의 모수 추정을 위해 E-step에서 모수가 $\boldsymbol{\Theta}^{(k)}$ 로 주어졌을 때, $(\mathbf{y}^{\text{obs}}, \mathbf{z}_{\partial_i})$ 에 대한 완비자료의 변량 $z_{ih}, z_{ih}\mathbf{y}_i$ 및 $z_{ij}\mathbf{y}_i\mathbf{y}_i^T$ 의 조건부 기대값을 필요로 한다. 이들은 각각 다음과 같다.

$$\begin{aligned} \tau_{ih}^{(k)} &= E[Z_{ih}|\mathbf{y}_i^{\text{obs}}, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] \\ &= \text{pr}_{\boldsymbol{\Theta}^{(k)}}\{Z_{ih} = 1|\mathbf{y}_i^{\text{obs}}, \mathbf{z}_{\partial_i}\} = \frac{\pi_{ih}^{(k)} \phi(\mathbf{y}_i^{\text{obs}}; \boldsymbol{\mu}_{ih}^{\text{obs}^{(k)}}, \boldsymbol{\Sigma}_{ih}^{\text{oo}^{(k)}})}{\sum_{t=1}^g \pi_{it}^{(k)} \phi(\mathbf{y}_i^{\text{obs}}; \boldsymbol{\mu}_{it}^{\text{obs}^{(k)}}, \boldsymbol{\Sigma}_{it}^{\text{oo}^{(k)}})} \end{aligned} \quad (3.3)$$

이고 이것은 i 번째 관측치가 h 번째 상태에서부터 발생하였을 사후확률 추정치이다. 그리고

$$\begin{aligned} E_1^{(k)} &= E[Z_{ih}\mathbf{Y}_i|\mathbf{y}_i^{\text{obs}}, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] \\ &= \text{pr}_{\boldsymbol{\Theta}^{(k)}}\{Z_{ih} = 1|\mathbf{y}_i^{\text{obs}}, \mathbf{z}_{\partial_i}\} E[\mathbf{Y}_i|\mathbf{y}_i^{\text{obs}}, z_{ih} = 1, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] \\ &= \tau_{ih}^{(k)} \mathbf{y}_{ih}^{(k)} \end{aligned} \quad (3.4)$$

이다. 단, $\mathbf{y}_{ih}^{(k)} = (\mathbf{y}_i^{\text{obs}T}, \mathbf{y}_{ih}^{\text{mis}^{(k)T}})^T$ 이며,

$$\begin{aligned} \mathbf{y}_{ih}^{\text{mis}^{(k)}} &= E[\mathbf{Y}_i^{\text{mis}}|\mathbf{y}_i^{\text{obs}}, z_{ih} = 1, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] \\ &= \boldsymbol{\mu}_{ih}^{\text{mis}^{(k)}} + \boldsymbol{\Sigma}_{ih}^{\text{mo}^{(k)}} \boldsymbol{\Sigma}_{ih}^{\text{oo}^{(k)-1}} (\mathbf{y}_i^{\text{obs}} - \boldsymbol{\mu}_{ih}^{\text{obs}^{(k)}}) \end{aligned} \quad (3.5)$$

를 나타낸다. 여기서 만약 관측치 i 에서 모든 원소가 결측이면 $\mathbf{y}_{ih}^{\text{mis}^{(k)}} = \boldsymbol{\mu}_{ih}^{\text{mis}^{(k)}}$ 을 의미한다. 그리고 비슷한 방법으로

$$E_2^{(k)} = E[Z_{ih}\mathbf{Y}_i\mathbf{Y}_i^T|\mathbf{y}_i^{\text{obs}}, \mathbf{z}_{\partial_i}, \boldsymbol{\Theta}^{(k)}] = \tau_{ih}^{(k)} \mathbf{S}_{ih}^{(k)}$$

이다. 단,

$$\mathbf{S}_{ih}^{(k)} = \begin{pmatrix} \mathbf{S}_{ih}^{\text{oo}^{(k)}} & \mathbf{S}_{ih}^{\text{om}^{(k)}} \\ \mathbf{S}_{ih}^{\text{mo}^{(k)}} & \mathbf{S}_{ih}^{\text{mm}^{(k)}} \end{pmatrix}$$

이며,

$$\mathbf{S}_{ih}^{\text{oo}^{(k)}} = \mathbf{y}_i^{\text{obs}} \mathbf{y}_i^{\text{obs}T}, \quad \mathbf{S}_{ih}^{\text{om}^{(k)}} = \mathbf{y}_i^{\text{obs}} \mathbf{y}_{ih}^{\text{mis}^{(k)T}}, \quad \mathbf{S}_{ih}^{\text{mo}^{(k)}} = \mathbf{S}_{ih}^{\text{om}^{(k)T}}$$

및

$$\mathbf{S}_{ih}^{mm(k)} = \mathbf{S}_{ih}^{mm(k)} - \mathbf{S}_{ih}^{mo(k)} \mathbf{S}_{ih}^{oo(k)-1} \mathbf{S}_{ih}^{om(k)} + \mathbf{y}_i^{\text{mis}(k)} \mathbf{y}_i^{\text{mis}(k)T}$$

이다.

한편 M-step에서는 π_h , μ_h 및 Σ_h 들을 추정해야 한다($h = 1, \dots, g$). 먼저, 식 (2.3)으로부터

$$\pi_{ih}^{(k+1)} = \frac{\exp(\beta u_{ih}^{(k)})}{\sum_{t=1}^g \exp(\beta u_{it}^{(k)})}, \quad h = 1, \dots, g; i = 1, \dots, n \quad (3.6)$$

와 같이 얻을 수 있다. 그런데 $u_{ih}^{(k)}$ 에 대해서는 Owen (1986)에서 처럼

$$u_{ih}^{(k)} = \sum_{s \in \partial_i} \tau_{hs}^{(k)} \quad (3.7)$$

와 같이하여 사용할 것이다. 이것은 미관측 상태값 z_{hi} 대신 사후확률 추정치를 사용한 것이다.

이제, 성분-평균 벡터 μ_h 를 추정하기 위해, $\mathbf{0} = \partial Q / \partial \mu_h$ 으로부터

$$\mu_h^{(k+1)} = \begin{pmatrix} \mu_{ih}^{\text{obs}(k+1)} \\ \mu_{ih}^{\text{mis}(k+1)} \end{pmatrix} = \frac{\sum_{i=1}^n \tau_{ih}^{(k)} \mathbf{y}_{ih}^{(k)}}{\sum_{i=1}^n \tau_{ih}^{(k)}}, \quad h = 1, \dots, g \quad (3.8)$$

를 쉽게 구할 수 있으며, 공분산 행렬 추정을 위해 $\mathbf{0} = \partial Q / \partial \Sigma_h$ 으로부터

$$\Sigma_h^{(k+1)} = \frac{\sum_{i=1}^n \tau_{ih}^{(k)} \mathbf{S}_{ih}^{(k)}}{\sum_{i=1}^n \tau_{ih}^{(k)}} - \mu_h^{(k)} \mu_h^{(k)T}, \quad h = 1, \dots, g \quad (3.9)$$

를 얻을 수 있다.

적절한 초기 추정치 $\Theta^{(0)}$ 하에서 모수 추정치 $\Theta^{(k)}$ 가 수렴할 때까지 E-step과 M-step을 충분히 반복 하면 로그-우도 $\ln L(\Theta)$ 를 (국소적으로) 최대화 하는 추정치 $\hat{\Theta}$ 를 얻게 된다.

이상의 결과로 볼 때 MRF 하에서 정규혼합모형에 기반한 결측치 대체를 위한 EM 알고리즘은 혼합비율 $\{\pi_{ih}\}$ 의 추정 과정을 제외하면, Hunt와 Jorgensen (2003)의 독립성 가정 하에서의 그것과 완전히 일치한다. 이는 기존 프로그램의 약간의 수정만으로 MRF 하에서의 대체기법을 그대로 사용할 수 있게 하는 매우 유용한 사실이라 하겠다.

EM 알고리즘으로부터 사후확률 추정치 $\{\hat{\tau}_{ih}\}$ 와 모수 추정치 $\hat{\Theta}$ 을 얻은 후 우리는 두 가지 방법 중 하나로 결측치의 대체값을 구할 수 있다. 즉,

$$\hat{\mathbf{y}}_i^{\text{mis}} = \sum_{h=1}^g \hat{\tau}_{ih} \hat{\mathbf{y}}_{ih}^{\text{mis}}, \quad i = 1, \dots, n \quad (3.10)$$

을 계산하거나 적합한 혼합모형

$$f(\mathbf{y}_i; \hat{\Theta}) = \sum_{h=1}^g \hat{\pi}_{ih} \phi(\mathbf{y}_i; \hat{\mu}_h, \hat{\Sigma}_h), \quad i = 1, \dots, n \quad (3.11)$$

으로부터 1개의 자료를 랜덤하게 추출하여 관측된 자료는 그대로 두고 결측된 위치에 대응하여 추출된 자료를 삽입하는 방법이다.

그러나 저자는 다음 절에서 소개할 색조영상 결측치 대체를 위해서는 전자의 방법을 추천한다. 그 이유는 영상 고유의 특성 때문이라 할 수 있는데, 전자의 방법은 항상 가중평균으로 계산된 대체 색상을 제공하여 시각적으로 두드러지는 경우가 별로 없지만, 후자의 경우는 시각적 위화감이 드는 화소 대체값을 발생할 가능성을 항상 가지고 있다. 영상의 민감한 부분에서 이러한 대체값을 가지는 화소가 몇 개만 있더라도 전체 영상의 느낌을 매우 다르게 할 수 있는 위험이 있다. 따라서 다음 절에서 색조영상 실험에서는 전자의 방법을 사용하였다.

4. 영상실험

4.1. 사전 정의들

이 절에서는 색조영상의 결측 화소값의 대체를 위해 앞 절에서 제공한 방법을 적용하고자 한다. 일반적으로 영상분석에서 가우시안 마코프 랜덤 필드 모형은 화소들 사이의 국소적 관계를 표현하는데 보편적으로 이용된다. 그리고 색조영상은 0-255 사이의 값을 가지는 R(홍), G(녹), B(적) 3 성분(변량)으로 이루어져 있고, 이 세 변량 값으로 각 화소의 색상을 결정한다. 그런데 현실의 대부분의 영상에서 이 세 변량은 국소적으로 매우 큰 상관구조를 가진다. 따라서 색조영상에 3-변량 가우시안 마코프 랜덤 필드 혼합모형을 적용하는 것은 매우 자연스러운 것으로 판단된다.

다만, 영상의 일부에서는 화소 주변값들이 서로 유사하다기 보다는 점진적으로 부드럽게 변화하는 경향을 보이는 영역(gradation area)이 있다. 이런 영역에서는 주어진 작은 수의 색상 군집으로 명확히 구분하기 어려워 이런 부분에서의 대체값들은 마치 잡음처럼 시각적 위화감이 들게 할 수도 있다. 이러한 점을 보완하기 위해 본 연구에서는 결측치의 예측값인 식 (3.5) 대신에 결측화소를 i 를 중심으로 인근 영역($\mathcal{B}_i$)에 속한 화소값(결측치인 경우는 예측값)들의 가중평균을 사용할 것이다. 즉,

$$\mathbf{y}_{ih}^{\text{mis}(k)} = \sum_{j \in \mathcal{B}_i} b_j \left\{ \mu_{jh}^{\text{mis}(k)} + \Sigma_{jh}^{\text{mo}(k)} \Sigma_{jh}^{\text{oo}(k)-1} (\mathbf{y}_j^{\text{obs}} - \mu_{jh}^{\text{obs}(k)}) \right\} \quad (4.1)$$

와 같이 정의한다. 단, 여기서 b_j 는 결측화소 i 의 인근영역 $\mathcal{B}_i$ 에 속한 j 번째 화소에 대한 가중치이며, 화소 i 에 대해서는 가장 큰 가중치를 그리고 멀어질수록 작은 가중치를 주었다. 본 실험에서 사용한 가중치(윈도우)는

$$\mathbf{b} = \begin{pmatrix} 0.0947 & 0.1183 & 0.0947 \\ 0.1183 & 0.1478 & 0.1183 \\ 0.0947 & 0.1183 & 0.0947 \end{pmatrix} \quad (4.2)$$

과 같다. 이것은 화소 i 의 인근영역 $\mathcal{B}_i$ 를 화소거리 1 이내로 정의한 것이지만 영상의 상태에 따라 더 넓게 정의할 수도 있으며 가중치도 다르게 정의할 수도 있을 것이다.

한편, 본 절의 실험에서는 혼합성분의 개수 $g = 5$ 로 정하였으며, 깃스상수 $\beta = 1.0$ 으로 고정하여 사용하였다. 여기서도 색조영상의 색상의 변이 상태 정도가 클수록 g 를 크게 그리고 결측화소의 개수가 많을수록 β 를 보다 크게 정의할 수도 있을 것이다.

그리고 본 연구에서 모수들의 초기 추정치는, 우선 결측 화소값을 가장 가까이에 있는 관측 화소값으로 결측치를 대체한 후 k -means 군집기법으로 영상을 분할하였다. 그리고 (사후확률 1로서) 각 분할된 영역의 평균 색상과 공분산 행렬을 초기추정치 $\mu_h^{(0)} (h = 1, \dots, 5)$ 및 $\Sigma_h^{(0)} (h = 1, \dots, 5)$ 의 초기 추정치로 삼았으며, 각 분할 영역의 화소개수(n_h) 비율을 $\pi_{1h}^{(0)} = \pi_{2h}^{(0)} = \dots = \pi_{1n_h}^{(0)} (h = 1, \dots, 5)$ 로 하였다.

본 실험에서는 잘 알려진 Lena 영상 중 일부분인 147×113 크기의 실험 영상을 준비하였다 (그림 1 참조). 첫 실험은 랜덤 결측치를 R, G, B 각 변량에 10%, 30%, 50% 비율로 결측 화소를 생성하였는데

그림 1: 147 × 113 크기의 원 색조영상, 그리고 각각 R-변량, G-변량, B-변량 영상

표 1: 예측오차와 CPU time

각 변량의 결측비율	10%	20%	30%	40%	50%
예측오차 $\sqrt{\text{MSE}}$	15.58	15.72	17.00	17.91	19.45
CPU 시간(초)	70	158	292	553	981

각 비율에 대응한 영상을 Lena10, Lena30, Lenna50으로 부르겠다 (그림 2 참조). 그런데 이 경우 세 변량으로 이루어진 색조영상에서는 1개 이상의 원소에서 결측치를 가지는 화소의 개수는 훨씬 크게 증가하게 된다. 예를 들어, Lena50의 경우 약 88% 가량의 화소가 1개 변량 이상에서 결측치를 가지게 되어 그 화소들은 색상으로 표현되지 못한다. 한편 3개 변량 모두 결측된 경우는 약 14%였다. 그러나 그림 1에서 1개의 변량값 결측이면 노랑, 자주, 하늘색으로, 2개의 변량값 결측이면 적, 녹, 청색으로 그리고 3개 변량값 모두 결측되면 검정색으로 표현하였다. 그리고 R-, G-, B-변량 영상에서 결측치 화소는 검정색으로 나타내었다.

4.2. 실험 결과 및 해석

그림 2에 결측영상을 대체한 결과를 제공하였다. 결측비율 증가함에 따라 대체된 영상의 품질이 미세한 부분(예, 입 주변)에서 다소 떨어지기는 하나 모든 경우 원영상이 주는 주요 영상정보는 거의 유지하고 있음을 보여주고 있다. 특히 대체된 색조영상과 R-, G-, B-변량 영상을 볼 때 색상의 상태를 원영상의 그 것을 거의 유지하고 있다. 이는 혼합성분에 의한 ($g = 5$ 개의) 마코프 랜덤 필드의 국소적 특징을 살린 분할된 영상의 (가중치화 된) 색상들이 잘 결합된 결과로 보인다. 아울러 3-변량 사이의 상관구조가 색상을 유지하는데 큰 도움을 주는 것으로 판단된다. 그리고 뺄이나 어깨 부분에서 부드러운 색상변화가 비교적 잘 유지되고 있는 것은 식 (4.1)에 의한 평활의 효과로 나타난 것으로 보인다. 실제 식 (4.1) 대신 식 (3.5)을 직접 사용하여 평활을 생략할 경우 (결과를 제시하지는 않았지만) 그 부분들에서 잡음과 같이 거친 대체색상을 나타내었다.

표 1에 각 변량의 결측비율을 10%씩 증가시키면서 결측화소값 당 예측오차($\sqrt{\text{MSE}}$)를 수록하였다. 단,

$$\text{MSE} = \frac{\sum_{i \in M} (y_i - \hat{y}_i^{\text{mis}})^2}{N - 1}$$

이며, M 은 모든 결측원소의 집합이며, N 은 M 의 원소 개수를 나타낸다. 결측률이 증가하면서 예측오차 역시 증가하기는 하지만, 255의 변량값으로 범위로 볼 때 그 값들이 대체적으로 작을 뿐 아니라 결측률에 비해 그 증가 정도도 작다. 다시 말해서 결측화소의 대체색상이 원영상의 화소색상 사이로부터 크게 벗어나지 않으며, 10% 씩 결측률이 증가할 때 약 1 정도의 대체변량값의 변화 정도만 있음을 보여주고 있다.

그림 2: 3개 쌍들의 첫번째 행은 각각 *Lena10*, *Lena30*, *Lena50* 결측영상과 대응하는 *R*-, *G*-, *B*-영상이고 두번째 행은 예측된 대체영상이다. 1개의 변량값 결측이면 노랑, 자주, 하늘색으로, 2개의 변량값 결측이면 적, 녹, 청색으로 그리고 3개 변량값 모두 결측되면 검정색으로 표현하였다. *R*-, *G*-, *B*-변량 영상에서 결측치 화소는 검정색으로 나타내었다.

그림 3: 3개 쌍들의 첫번째 행은 각각 *LenaSQ5*, *LenaSQ10*, *LennaSQ15* 결측영상과 대응하는 *R*-, *G*-, *B*-영상이고 두번째 행은 예측된 대체영상이다. 사각형 결측의 첫 행은 *R*-, *G*-, *B*-변량값이 모두 결측된 경우이며, 두 번째 행은 *R*-, *G*-변량값이 결측된 경우이고, 세 번째와 네 번째 행은 각각 *G*-변량값과 *R*-변량값이 결측된 경우이다.

한편 알고리즘 모든 경우 70회 반복전에 충분히 수렴하였는데, 측정된 CPU 시간(하드웨어: Pentium(R) Dual-Core E5200@2.0GHz, 프로그래밍 언어: Matlab R2009a)은 결측률의 증가에 따라 점차 증가하되 증가율 역시 더욱 커지는 것으로 나타났다 (표 1 참조). 이것은 관측 자료(즉, 불완비 자료)가 적을수록 수렴속도가 느려지는 EM 알고리즘 고유의 특징 때문이라 할 수 있다.

두번째 실험은 5×5 , 10×10 및 15×15 크기의 정사각형 결측을 가진 영상 LenaSQ5, LenaSQ10 및 LenaSQ15에 대한 제안된 방법의 대체능력을 실험하고자 한다. 각 결측영상에서 사각형 결측의 첫 행은 R-, G-, B-변량값이 모두 결측된 경우이며, 두 번째 행은 R-, G-변량값이 결측된 경우이고, 세 번째와 네 번째 행은 각각 G-변량값과 R-변량값이 결측된 경우이다. 이와 같은 결측은 랜덤결측이 아니라 결측이 집단으로 모여 있기 때문에 대체의 어려움이 더 클 것으로 예상된다.

그림 3에 대체 결과를 수록하였다. 대체적으로 만족스러운 결과를 보이고 있으나, 결측 블록의 크기가 클수록 그리고 결측 변량이 많을수록 대체 능력이 떨어지는 것으로 나타났다. 특히 모든 영상의 첫 행의 결측 블록들의 경우 (즉, 3-변량이 모두 결측된 경우), 결측화소들이 주변의 평균 색상으로 강력히 대체되려는 경향을 보이고 있다. 그러나 이러한 특성은 랜덤 결측인 경우는 장점일 수 있지만 LenaSQ15의 첫 행의 오른쪽 마지막 결측 블록의 경우에서 처럼 대체 블록이 영상의 객체의 윤곽 특성을 잘 살리지 못하는 문제점이 있다. 그렇지만 대단위 블록 결측을 제외한 나머지 결측 블록에 대해서는 모두 양호한 대체능력을 보여주고 있다.

5. 결론 및 토의

본 논문에서는 MRF 하에서 다중 결측치 대체를 위한 정규혼합모형의 사용을 제안하였다. 그 이유로서 MRF 하에서 다중 결측치 대체기법은 자체로 정규혼합모형의 적합과 동일하는 점을 밝혔다. 결측치 대체를 위한 EM 알고리즘을 제공하였으며, 이를 바탕으로 색조영상 결측화소 대체문제에 적용하여 실험을 통해 그 실효성을 보였다. 특히 EM 알고리즘의 E-step에서 결측화소 주변의 평활을 제안하였는데, 이것은 색조영상 대체에 잘 적응하였다.

다만, 제안된 방법은 (랜덤) 결측률이 증가할수록 수렴속도는 느려지고 처리시간도 급격히 증가한다. 이것은 관측 자료(즉, 불완비 자료)가 적을수록 수렴속도가 느려지는 EM 알고리즘 고유의 특징 때문이라 할 수 있다. 따라서 수렴속도를 향상시킬 수 있는 개선법이 요구된다. 또한 대단위 완전 블록결측의 경우 대체값이 객체의 윤곽특성과 세부적 특성을 잘 살리지 못하는 점을 개선하기 위해서는 보다 현실적이고 다양한 제약을 두어야 할 것으로 보인다.

참고 문헌

- Besag, J. (1975). Statistical analysis of non-lattice data, *The Statistician*, **24**, 179–195.
- Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society B*, **48**, 259–302.
- Blanchet, J. and Vignes, M. (2009). A model-based approach to gene clustering with missing observation reconstruction in a Markov random field framework, *Journal of Computational Biology*, **16**, 475–486.
- Dass, S. C. and Nair, V. N. (2003). Edge detection, spatial smoothing, and image reconstruction with partially observed multivariate data, *Journal of the American Statistical Association*, **98**, 77–89.
- Hunt, L. and Jorgensen, M. (2003). Mixture model clustering mixed data with missing information, *Computation Statistics & Data Analysis*, **41**, 429–440.
- Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology*, **12**, 1–16.
- Kim, D., Lee, Y. and Oh, H. S. (2006). Hierarchical likelihood-based wavelet method for denoising signals with missing data, *IEEE Signal Processing Letters*.

- Little, J. and Rubin, D. (2002). *Statistical Analysis of Missing Data*, Wiley, New York.
- Marker, D. A., Judkins, D. R. and Winglee, M. (2002). *Large-Scale Imputation for Complex Surveys*, In: Groves, R. M. and Eltinge, E. L.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.
- Ogawa, T., Haseyama, M. and Kitajima, H. (2006). Restoration of Missing Intensity of Still Images by Using Optical Flows, *System and Computers in Japan*, **37**, 1786–1795.
- Owen, A. (1986). Contribution to the discussion of paper by B.D.Ripley, *Canadian Journal of Statistics*, **14**, 106–110.
- Qian and Titterington. (1992). Stochastic relaxations and EM algorithms for Markov random fields, *Journal of Statistical Computation and Simulation*, **40**, 55–69.
- Zio, M. D., Guarnera, U. and Luzi, O. (2007). Imputation through finite Gaussian mixture models, *Computation Statistics & Data Analysis*, **51**, 5305–5316.

2009년 10월 접수; 2009년 10월 채택

Imputation of Multiple Missing Values by Normal Mixture Model under Markov Random Field: Application to Imputation of Pixel Values of Color Image

Seung-Gu Kim^{1,a}

^aDepartment of Data & Information, Sangji University

Abstract

There very many approaches to impute missing values in the *iid.* case. However, it is hardly found the imputation techniques in the Markov random field(MRF) case. In this paper, we show that the imputation under MRF is just to impute by fitting the normal mixture model(NMM) under several practical assumptions. Our multivariate normal mixture model based approaches under MRF is applied to impute the missing pixel values of 3-variate (R, G, B) color image, providing a technique to smooth the imputed values.

Keywords: Multiple missing values, imputation, Markov random field, EM algorithm, color image.

¹ Professor, Department of Data & Information, Sangji University, 660 Woosan-Dong, Wonju, Kangwon 220-702, Korea.
E-mail: sgukim@sangji.ac.kr