CHAPTER

Image denoising, local factor analysis, Bayesian Ying-Yang harmony learning[1]

6

Guangyong Chen, Fengyuan Zhu, Pheng Ann Heng, Lei Xu

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

Advances in Independent Component Analysis and Learning Machines

Edited by

Ella Bingham

Samuel Kaski

Jorma Laaksonen

Jouko Lampinen

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London, EC2Y 5AS, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-802806-3

Publisher: Matthew Deans
Acquisition Editor: Tim Pitts
Editorial Project Manager: Charlie Kent
Production Project Manager: Melissa Read
Designer: Greg Harris

CHAPTER

Image denoising, local factor analysis, Bayesian Ying-Yang harmony learning[1]

6

Guangyong Chen, Fengyuan Zhu, Pheng Ann Heng and Lei Xu
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

6.1 A BRIEF OVERVIEW ON DENOISING STUDIES

Distortion is often introduced into images through capturing instruments, data transmission media, image quantization, and discrete source of radiation. There are two typical image distortions [1]. One is called blur, which is intrinsic to image acquisition systems. The second distortion comes from environmental disturbances, which is usually called noises. Typically, an observed image X is regarded as generated from an additive model $X = \hat{X} + E$ with $\hat{X}$ representing the clean image and E denoting the noise that is independent from $\hat{X}$, as shown in Figure 6.1. This chapter focuses on removing this additive noise E, which is a challenging ill-posed problem. In the past decades, many efforts have been made and a brief overview is provided as follows.

6.1.1 SPATIAL FILTERING

A spatial filter is an image operator where each pixel x_t is changed by a function of the intensities of pixels in a neighbor $\mathcal{N}(x_t) \in \mathbb{R}^{d\times d}$. Based on the assumption that statistical property of images keep piecewise constant, numerous methods have been proposed for different noise types, such as mean filter, Gaussian filter, Wiener filter, weighted mean filter, anisotropic filter, bilateral filter for Gaussian noise, and median filter for Laplace noise [2].

Under the assumption of piecewise flat, the mean filter estimate

$$\bar{x}_t = \frac{1}{\#\mathcal{N}(x_t)} \sum_{x_j \in \mathcal{N}(x_t)} x_j, \tag{6.1}$$

where $\#S$ denotes the cardinality of the set S. When a pixel value x in $\mathcal{N}(x_t)$ is affected by a noise e from $G(e|0, \sigma_e^2)$, it follows from $x = \hat{x}_t + e$ that we have $G(x|\hat{x}_t, \sigma_e^2)$,

[1] All related codes and data set are available on the website: http://appsrv.cse.cuhk.edu.hk/~gychen/.

E. Bingham, S. Kaski, J. Laaksonen and J. Lampinen (Eds): Advances in Independent Component Analysis and Learning Machines.
http://dx.doi.org/10.1016/B978-0-12-802806-3.00006-3

FIGURE 6.1

The additive noise model. (a) An observed image, (b) the clean image, and (c) the additive noise. An example of flat patch is marked in the deep-color box, while the light color box contains a sharp edge.

where $G(u|\mu, \sigma^2)$ denotes a Gaussian distribution with the mean μ and the variance σ^2. By Eq. (6.1), we obtain

$$\bar{x}_t \sim G\left(\bar{x}_t|\hat{x}_t, \frac{\sigma_e^2}{d^2}\right). \tag{6.2}$$

which indicates that the reconstructed pixel $\bar{x}_t$ is a rough approximation of the original mean $\hat{x}_t$, whose accuracy heavily depends on the size $d \times d$ of $\mathcal{N}(x_t)$. A big d seemly increases the accuracy, but requires the property of local flat in a large neighbor. Usually, it is difficult to choose a suitable d.

Assume that the noise variance σ_e^2 is known and also $\hat{x}_t$ comes from a Gaussian with the mean $\bar{x}_t$ and the variance σ_x^2, the Wiener filter further improves the mean filter by

$$\bar{x}_{t|\text{Wiener}} = \bar{x}_t + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}(x_t - \bar{x}_t). \tag{6.3}$$

It involves $\bar{x}_t$ and thus still faces the difficulty of choosing a suitable d. Also, it is not easy to estimate σ_x^2.

Moreover, the assumption of piecewise constant does not hold anywhere on the images, such as the sharp edges shown in the blue box in Figure 6.1. Consequently, removing noise may blur sharp edges and destroy other subtle image information. Actually, all the spatial filtering methods suffer this same difficulty and limitation.

6.1.2 TRANSFORM-DOMAIN FILTERING

Also, numerous algorithms have been developed to separate noises in the transform domains, such as low-pass filter, Fourier Wiener filter, windowed Fourier Wiener filter, wavelet transform-based filter (WT), discrete cosine transform-based filter (DCT), etc., usually with a lower time complexity compared with spatial filtering methods.

Assume that images are piece smooth, a low-pass filter treats the high-frequency components in Fourier spectral as noise. However, as demonstrated in Figure 6.1,

the sharp edges contained in the blue box represent the high-frequency component in the Fourier domain, and thus discarding high-frequency components will blur the edges. Moreover, a practical implementation of low-pass filtering methods introduces a ringing effect into the reconstructed images.

Considering that the property of image can be piecewisely described, DCT divides a noisy image X into a local $d \times d$ window x_t, transforms x_t into a linear combination of d^2 frequency squares, and cuts off those with the corresponding coefficients below a small threshold. Widely used in JPEG compression, DCT has achieved promising denoising performance. However, it is difficult to partition noisy images into local windows with suitable size, which usually causes artifacts. Similar to a low-pass filter, DCT also fails if the local window contains a high-frequency pattern.

Spatial filtering and transform-domain filtering are linked via rewriting the spatial filtering into the following convolution

$$\bar{x} = X * K, \tag{6.4}$$

where K denotes the denoised kernel, for example, $K = \mathbf{1} \in \mathbb{R}^{d \times d}$ for the mean filter. By the convolution theorem, the counterpart of Eq. (6.4) in the Fourier transform is given as follows

$$\mathcal{F}(\bar{x}) = \mathcal{F}(X) \times \mathcal{F}(K), \tag{6.5}$$

where $\mathcal{F}(X)$ and $\mathcal{F}(K)$ denote the Fourier transform of the noisy image X and kernel K, respectively. Both spatial filtering and transform-domain filtering assume images are piecewise constant and eliminate noise at the cost of blurring subtle feature structures in noisy images.

6.1.3 NONLOCAL MEANS METHODS

Instead of assuming images to be piecewise constant, nonlocal means methods observe that similar feature information is acquired repeatedly in images, for example, the sharp edge contained in the deep color box in Figure 6.2 appears elsewhere in this image, as marked in the light color box in Figure 6.2. Assume that a noisy patch x_t, as marked in the green box in Figure 6.2, is generated as follows

$$x_t = \hat{x}_t + e_t, \tag{6.6}$$

where $\hat{x}_t \in \mathbb{R}^{d \times d}$ denotes the clean patch and $e_t \in \mathbb{R}^{d \times d}$ denotes random noise with each element being independent of each other and also with $\hat{x}_t$. Given enough patches that contain the same feature information, a simple average of the searched patches will remove Gaussian noise and give a plausible estimation $\bar{x}_t$ of the clean patch $\hat{x}_t$. As shown in Figure 6.2(d), nonlocal means methods can preserve the sharp edge perfectly, compared with traditional spatial filtering and transform-domain filtering. Following this idea, several algorithms [3–7] have been proposed in the last 10 years, with promising performances.

One representative method is known as K-SVD [3,4]. By its nature, it could be implemented by iterating the step of updating the dictionary of representative patches

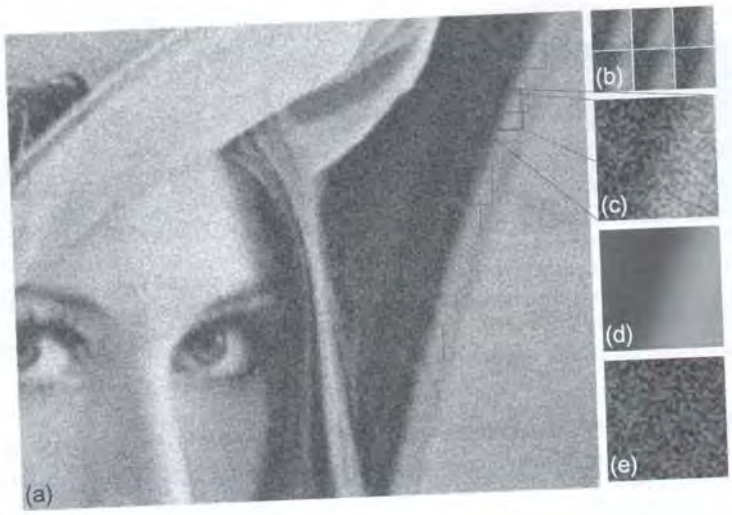

FIGURE 6.2

The motivation of nonlocal means methods. (a) A noisy image, where the predenoised patch is marked in a deep color box, (b) six representative patches similar to the deep color one, whose positions are marked in the light color boxes in (a), while (c), (d), and (e) denote the noisy patch, and the estimated clean patch, and the estimated noisy patch, respectively. As demonstrated in (d), the nonlocal means method recovers the sharp edge contained in the deep color box.

or features and the step of denoising a corrupted image based on the dictionary, such that a subset of patches is picked from the dictionary to form a sparsely weighted linear sum as a reconstruction of a corrupted patch subject to a given noise variance. Practically, to avoid high computing cost and also to improve denoising performance, K-SVD gets the dictionary obtained previously from a corpus of high-quality image database and usually with human interactive help. Then, denoising of the present corrupted image is based on this dictionary.

One other representative method is known as BM3D [5]. Those patches that match well with a reference patch under consideration are picked out and stacked to form a 3D array that is mapped into a transform domain where a collaborative filtering of the group is made by shrinkage. Then, the filtered 3D array is inversely transformed back to the estimates of those 2D patches. After processing all reference patches, the obtained estimates can overlap and then aggregate to form a final estimate of the true image.

Both K-SVD and BM3D need to know the noise variance in advance, which takes a critical role in not only getting the aggregating weights but also denoising either directly for K-SVD or via grouping and collaborative filtering for BM3D. Moreover, there are also some parameters that are heuristically chosen in advance

for both K-SVD and BM3D. Furthermore, K-SVD utilizes a pretrained dictionary while BM3D uses a general DCT basis, without considering the issue of controlling the dictionary complexity. Both the methods are suitable only to a limited range of images. Tacking these limitations, we propose a new denoising method called LFA-BYY.

6.2 LFA-BYY DENOISING METHOD

Instead of considering each item in the dictionary as an image patch or feature template as K-SVD and BM3D, LFA-BYY treats it as a parametric model that represents a family of image patches or a set of features. The dictionary consists of a set of parametric models that can be provided in one of three ways:

1. Learned from the present image under processing.
2. Obtained previously by the same method and other methods as well, including human help.
3. Obtained by updating the available dictionary on the present image.

In this chapter, we focus on the first way, though the study can be further generalized to the other ways.

To facilitate mathematical formulation, we stack each $d \times d$ image patch into a d^2 dimension vector. That is, we consider Eq. (6.6) in $\mathbb{R}^{d^2}$ instead of $\mathbb{R}^{d\times d}$. Each parametric model uses a factor analysis (FA) model featured by a $d^2 \times m_i$ matrix A_i with its columns spanning a subspace that locates at $\mu_i \in \mathbb{R}^{d^2}$. Each sample vector x_t comes from a mixture of FA models at a number of locations, also called local factor analysis (LFA) [8]. Specifically, we consider that x_t comes from

$$q(x_t|\theta) = \sum_i \alpha_i G(x_t|\mu_i, \Sigma_i),\ \theta = \{\alpha_i, \mu_i, A_i, \sigma_i^2, \Lambda_i\}_{i=1}^k,$$
$$G(x_t|\mu_i, \Sigma_i) = \int G(x_t|A_i y + \mu_i, \sigma_i^2 I) G(y|0, \Lambda_i)\mathrm{d}y,$$
$$\text{subject to } A_i^T A_i = I, \tag{6.7}$$

where $G(u|\mu, \Sigma)$ denotes a Gaussian with the mean vector μ and the covariance matrix Σ. Equivalently, we extend Eq. (6.6) into

$$x_t \text{ comes from } x_{i,t} \text{ with a probability } \alpha_i \text{ for } i = 1, \ldots, k,$$
$$x_{i,t} = \underbrace{A_i y_{i,t} + \mu_i}_{\hat{x}_{i,t}} + e_{i,t},\ \text{subject to } A_i^T A_i = I,$$
$$y_{i,t} \sim G(y_{i,t}|0, \Lambda_i),\ e_{i,t} \sim G(x_t|Ay + \mu_i, \sigma_i^2 I). \tag{6.8}$$

Turning the image under processing into a set $\mathcal{X} = \{x_t\}$ of samples, we estimate the parameter set θ and determine a set of integers $\mathcal{K} = \{k, \{m_i\}_{i=1}^k\}$ by the Bayesian Ying-Yang (BYY) harmony learning, which was first proposed in 1995 [9] and systematically developed over the past two decades [10,11].

Based on the obtained θ, each x_t is transformed into its cleaned counterpart $\hat{x}_t$ as follows

$$\hat{x}_t = A_j y_{j,t} + \mu_j, \quad j = \arg\max_i p(i|x_t),$$

$$y_{j,t} = \arg\max_y [G(x_t|A_j y + \mu_j, \sigma_j^2 I) G(y|0, \Lambda_j)], \tag{6.9}$$

where $p(i|x_t)$ is already computed during learning, representing the following posteriori probability

$$p(i|x_t) = \frac{\alpha_i G(x_t|\mu_i, \Sigma_i)}{\sum_j \alpha_j G(x_t|\mu_j, \Sigma_j)}. \tag{6.10}$$

Quoting the roles taken by the LFA model and the BYY harmony learning, this denoising method is thus known by the abbreviation LFA-BYY.

Equation (6.9) implements a concept that relates to the grouping and collaborative filtering in BM3D [5]. However, the method for implementing the concept is different. First, one group of patches is actually one class of the patches represented by its corresponding FA by Eq. (6.8). All the patches are classified into k-classes based on $p(i|x_t)$ that describes the fitness of the ith FA model to the patch x_t, while BM3D searches the similar patches by Euclidean distance and groups them based on a prespecified threshold. Second, making collaborative filtering by BM3D involves heuristic settings with human help, which is no longer necessary in Eq. (6.9). The component μ_j is actually a counterpart of the result by a low-pass filtering along the stacking direction by BM3D. Since the patches in a group or class may come from different locations of the image, it is reasonable to regard them as a sequence of identically and independently distributed (i.i.d.) elements and thus only consider their mean value with other cross element relations ignored. Put together, the linear transform $x_t \to y_{j,t} \to \hat{x}_t$ is a counterpart of the collaborative filtering (i.e., 2D patches $\to$ 3D array $\to$ filtered 3D array $\to$ 2D patches). Lastly, all the patches are classified exclusively into k-classes without overlap, thus there is no need of an aggregation like that used by BM3D to smooth out inconsistency, which may cause artifacts as well.

Also, the above LFA extends the K-SVD [3,4]. In the degenerated case that there is only one FA $x_t = Ay_t + e_t$, the matrix A is the counterpart of the dictionary used in the K-SVD, while y_t is the counterpart of the sparse weights by the K-SVD. Being different, the LFA model corresponds to an organized hierarchical dictionary, with not only subsets $\{A_j\}_{j=1}^k$ in a lower layer but also $\{\mu_j\}_{j=1}^k$ in an upper layer, such that the above addressed grouping and collaborative filtering can be also performed. Moreover, the K-SVD considers weights as sparse coding that is approximately obtained by some pursuit algorithm, while the LFA considers the weights separately in each FA by $y_{j,t}$ coded by independent Gaussian distributions. Furthermore, K-SVD also uses a dictionary that is collected from a large number of clean images, while the LFA model utilizes an adaptive dictionary learned from the present noisy image.

Even critically, both K-SVD and BM3D require that the noise variance is preestimated because they have not taken the issue of controlling dictionary complexity into consideration, while estimating the noise variance is closely associated with appropriately controlling the complexity of the dictionary. A high complexity leads to an over-fitting problem and formulates noise as feature information, thus creating artifacts. In contrast, a lower complexity introduces an under-fitting problem and treats the feature information as noise, thus smoothing out the subtle feature information. The dictionary complexity is featured by $\mathcal{K}$ for the local FA model by Eqs. (6.7) and (6.8). With k and each m_i appropriately determined, we can get the noise variance σ_i^2 estimated appropriately. Therefore, we may break the limitation K-SVD and BM3D, that is, we not only need to know the noise variance but also its applicability to the heterogeneous noises, namely, different groups of image patches may be affected by noises with different variances.

Therefore, how to determine one appropriate $\mathcal{K}$ effectively is a key problem for learning LFA. In the next section, we will see that the BYY harmony learning provides a favorable solution to this problem, and also the LFA model is actually different from those existing models called mixture of factor analyzers (MFA) [12,13], because each FA model is different from the traditional FA for facilitating to determine $\mathcal{K}$ effectively.

6.3 BYY HARMONY LEARNING ALGORITHM FOR LFA

The task of determining one appropriate $\mathcal{K}$ is called model selection. A conventional model selection approach is featured by a two-stage implementation. The first stage is called parameter learning for estimating all the unknown parameters θ for every value of $\mathcal{K}$ in a prespecified set that enumerates all the candidate models under consideration. The second stage selects the best candidate by a model selection criterion, for example, Akaike's Information Criterion [14], Bayesian Information Criterion [15], Minimum Message Length [16], etc. However, a larger k and m_i imply more unknown parameters, which makes parameter estimation become less reliable such that the criterion evaluation reduces its accuracy, see Section 2.1 in [11] for a detailed discussion. Moreover, such two-stage procedure suffers a huge-computing cost.

Instead of two-stage implementation, automatic model selection is featured by determining one appropriate $\mathcal{K}$ during parameter learning on θ, with $\mathcal{K}$ starting at a value large enough and then gradually shrinking. An early effort is rival penalized competitive learning (RPCL) [17,18], featured by the cluster number automatically determined during learning. Two types of Bayesian-related approaches can perform automatic model selection. One is the BYY harmony learning and the other is variational Bayesian (VB) [19,20]. The VB introduces a function to approximate the marginal likelihood by Jensen's inequality and employs an EM-like algorithm to optimize the lower bound. The model selection of VB is realized by incorporating

appropriate prior distributions of unknown parameters. As empirically demonstrated by [21,22], BYY is capable of selecting suitable model complexity automatically even without imposing any priors on the parameters, and outperforms VB with Dirichlet-Normal-Gamma prior distribution [23].

The LFA model by Eqs. (6.7) and (6.8) has already taken the issue of facilitating automatic model selection into consideration, where the FA model is actually different from the traditional FA. To avoid confusion, the traditional one is named FA-a, while each FA in Eqs. (6.7) and (6.8) is named FA-b. The orthonormal constraint of $A_i^T A_i = I$ is removed by FA-a that imposes $\Lambda_i = I$. In the studies of BYY harmony learning, FA-b was preferred as discussed in Item 9.4 in [24] and Section 3 in [25]. Two FA types are equivalent in terms of maximizing the likelihood, but behave very differently when selecting model complexity is taken into consideration, with further details referred to Section 3.2.1 in [26] for a recent overview. Extensive empirical experiments in [27] have shown that the BYY harmony learning and VB perform reliably and robustly better on FA-b than on FA-a, while BYY outperforms VB considerably, especially on FA-b. Moreover, a mixture of FA-a models is called MFA [12,13], while a mixture of FA-b models, namely the one given by Eqs. (6.7) and (6.8), is called LFA [8], also see Figures 3 and 9 in [11]. Again, empirical experiments in [28] have confirmed that learning LFA by BYY outperforms learning MFA by BYY, learning LFA by VB, and learning MFA by VB [13].

The BYY harmony learning was first proposed in [9] and systematically developed in the past two decades. BYY harmony learning on typical structures leads to new model selection criteria, new techniques for implementing learning regularization, and a class of algorithms that implement automatic model selection during parameter learning. Also, BYY harmony learning offers a theoretical explanation of RPCL. Further details and the latest systematical introduction about BYY harmony learning can be found in [10,11].

Here, we directly adopt Algorithm 5: "BYY learning for local factor analysis" given in [10], which is rewritten as Algorithm 1 in this chapter. It implements the maximization of the harmony measure in a BYY system. On the LFA model by Eqs. (6.8) and (6.7), the harmony measure is formulated as follows

$$\begin{aligned} &\mathrm{H}(p\|q) = \sum_{t=1}^{N}\sum_{i=1}^{k}\int p(i|x_t)p(y|i,x_t)\ln[\alpha_i G(x_t|A_i y+\mu_i,\sigma_i^2 I)G(y|0,\Lambda_i)]\mathrm{d}y, \\ &\text{s.t.:} A_i^T A_i = I, i = 1,\ldots,k, \\ &\text{s.t.:}\mathrm{KL}\left(p(i|x_t)p(y|i,x_t)\middle\|\frac{\alpha_i G(x_t|A_i y+\mu_i,\sigma_i^2 I)G(y|0,\Lambda_i)}{q(x_t|\theta)}\right) = 0, \end{aligned} \tag{6.11}$$

which is maximized to determine not only $\theta, p(i|x_t)$ and $p(y|i,x_t)$, but also $\mathcal{K}$. With the help of the Lagrange method, an iterative procedure is used to implement this maximization, from which we are led to Algorithm 1.

ALGORITHM 1 BYY LEARNING FOR LOCAL FACTOR ANALYSIS

Initialize $\theta = \{\alpha_i, A_i, \Lambda_i, \sigma_i^2, \mu_i\}_{i=1}^k$, and η is controlled as described in Section 2.3 in [10].
Repeat the following two steps **until** converged.

***Yang* Step**: We get:

$p_{i,t} = p(i|x_t, \theta^{\text{old}})$ by Eq. (6.13);

$\Gamma_{i,y|x}^{\text{new}} = \frac{\eta}{1+\eta}\sigma_i^{2\ \text{old}}(\mathbf{I} + \sigma_i^{2\ \text{old}}\Lambda_i^{\text{old}-1})^{-1}$;

$W_i^{\text{new}} = \Gamma_{i,y|x}^{\text{old}} A_i^{\text{old}\,T}\sigma_i^{\text{old}\,-2}$.

***Ying* Step:** *get* $\alpha_i^{\text{new}}, \mu_i^{\text{new}}, A_i^{\text{new}}, \sigma_i^{2\ \text{new}}, \nu_i^{\text{new}}, \Lambda_i^{\text{new}}$ by:

$\alpha_i^{\text{new}} = \frac{1}{N}\sum_{t=1}^N p_{i,t}$, $\mu_i^{\text{new}} = \frac{1}{N\alpha_i^{\text{new}}}\sum_{t=1}^N p_{i,t}x_t$;

$y_{t,i} = W_i^{\text{new}}(x_t - \mu_i)$, $e_{t,i} = x_t - \mu_i^{\text{new}} - A_i^{\text{old}}y_{t,i}$;

$\sigma_i^{2\ \text{new}} = \frac{1}{d^2}Tr\left[A_i^{\text{old}}\Gamma_{i,y|x}^{\text{new}}A_i^{\text{old;T}} + \frac{1}{N\alpha_i^{\text{new}}}\sum_{t=1}^N p_{i,t}e_{t,i}e_{t,i}^T\right]$;

$\Lambda_i^{\text{new}} = \text{diag}\left(\Gamma_{i,y|x}^{\text{new}} + \frac{1}{N\alpha_i^{\text{new}}}\sum_{t=1}^N p_{i,t}y_{t,i}y_{t,i}^T\right)$;

$A_i^{\text{new}} = G_S\left[\sum_{t=1}^N p_{i,t}(x_t - \mu_i^{\text{new}}y_{t,i}^T)\right]\left[\sum_{t=1}^N p_{i,t}(y_{t,i}y_{t,i}^T + \Gamma_{i,y|x}^{\text{new}})\right]^{-1}$.

where $G_S[\phi]$ denotes a Gram-Schmidt operator that orthogonalizes ϕ, i.e., $\phi\phi^T = I$.
TRIMMING:

if one $\lambda_{i,l}^{\text{new}}$ of $\Lambda_i = \text{diag}[\lambda_{i,1}, \ldots, \lambda_{i,m_i}]$ tends to 0, discard the lth column of A_i, let $m_i = m_i - 1$; if $\alpha_i^{\text{new}} \to 0$ or $\alpha_i^{\text{new}}\sigma_i^{2\ \text{new}} \to 0$, discard $G(x|A_i y + \mu_i, \sigma_i^2 I)$ and $G(y|0, \Lambda_i)$, let $k = k - 1$.

The above $\text{H}(p\|q)$ is a specific derivation of the following general form

$$H(p||q) = \int p(R|X)p(X)\ln[q(X|R)q(R)]\text{d}X\text{d}R, \tag{6.12}$$

where R consists of not only θ, y, i explicitly but also $\mathcal{K}$ implicitly. Maximizing $H(p||q)$ forces $q(X|R)q(R)$ (called the Ying structure) to match $p(R|X)p(X)$ (called the Yang structure). There are always certain structural constraints imposed on the Ying-Yang structures with an additional one coming from $p(X)$ to accommodate a finite size of samples, because a perfect equality $q(X|R)q(R) = p(R|X)p(X)$ may not be really reached but still be approached as close as possible. At this equality, $H(p||q)$ becomes the negative entropy that describes the complexity of the BYY system. Further maximizing it will decrease the system complexity and thus provides an ability for determining an appropriate $\mathcal{K}$.

Observing Algorithm 1, such a model selection ability is reflected mainly in its Ying step, namely

$$p(i|x_t, \theta) = \frac{[\alpha_i G(x_t|\mu_i, \Sigma_i)]^{\frac{1+\eta}{\eta}}}{\sum_{i=1}^k [\alpha_i G(x_t|\mu_i, \Sigma_i)]^{\frac{1+\eta}{\eta}}},$$

$$p(y|i, x_t) = G\left(y|y^*, \frac{\eta}{1+\eta}\Sigma_{p(y|i,x_t)}\right),$$

$$y^* = (\mathbf{I} + \sigma_i^2 \Lambda_i^{-1})^{-1} A_i^T (x_t - \mu_i),$$

$$\Sigma_{p(y|i,x_t)} = \sigma_i^2 (\mathbf{I} + \sigma_i^2 \Lambda_i^{-1})^{-1}, \tag{6.13}$$

where μ_i, Σ_i are the same as in Eq. (6.7). The Lagrange parameter η reflects an agreement of balance between the Ying and the Yang. Not only the difference between the above $p(i|x_t, \theta)$ and its posteriori probability counterpart by Eq. (6.10) but also the difference between the above $p(y|i, x_t)$ and its posteriori probability counterpart $G(y|y^*, \Sigma_{p(y|i,x_t)})$ are featured by η, which makes $p(i|x_t, \theta)$ and $p(y|i, x_t)$ become more selective for automatic model selection on $\mathcal{K}$. As a result, the sparsity for each observable sample x_t is realized by $\hat{x}_t$ via the linear transform $x_t \to y_{j,t} \to \hat{x}_t$ by Eq. (6.9).

When $\eta = \infty$, the Yang step will degenerate into the E-step of the classical expectation-maximization (EM) algorithm. With $p(i|x_t), p(y|i, x_t)$ replaced by their posteriori probability counterparts, or letting $\eta = \infty$, the Ying step of Algorithm 1 also becomes the same as the same M-step with the classical EM algorithm. Further discussions can be found in Section 2.3 in [10].

6.4 EXPERIMENTS AND DISCUSSION

6.4.1 COMPARATIVE RESULTS WITH COMPETING ALGORITHMS

To make a systematical assessment, we compare LFA-BYY with four state-of-the-art image denoising algorithms: BM3D [5], K-SVD for color image denoising [29], expected patch log likelihood (EPLL) [6], and multispectral image denoising (Msi) [30], on the benchmark Kodak image processing dataset that contains 24 natural images. Each image is polluted by Gaussian noise with 10 different intensities from $\sigma = 10$ to $\sigma = 100$.

We perform each of the competitive algorithms with the code provided with its published paper and websites as follows:

- BM3D (http://www.cs.tut.fi/~foi/GCF-BM3D/);
- K-SVD (http://www.ipol.im/pub/art/2012/llm-ksvd/);
- EPLL (http://people.csail.mit.edu/danielzoran/); and
- Msi (www.cs.cmu.edu/yiyang/Publications.html).

All the free parameters of these approaches are set as suggested in the original papers. We also provide the exact noise intensity of each polluted image as input to all the competing algorithms but not to LFA-BYY as it is able to estimate the noise intensity of a polluted image during the denoising procedure. The patch size of our algorithm is set to be 8×8.

The performance of each algorithm is evaluated by comparing the similarity between the original clean images and denoised ones with the measurements of peak

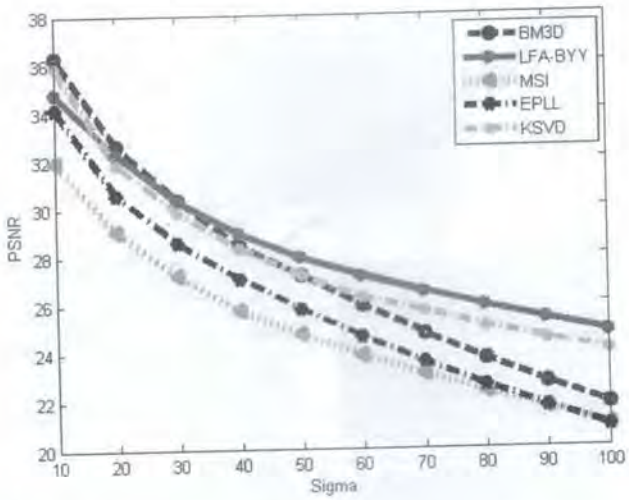

FIGURE 6.3

Average PSNR value of 24 natural images over different noise intensities.

signal-to-noise ratio (PSNR), structure similarity (SSIM) [31], and feature similarity (FSIM) [32]. The larger these measurements are, the more similar a denoised image is with the original.

Figures 6.3, 6.4, and 6.5 illustrate the experimental results. It can be observed that LFA-BYY consistently produces promising experimental results. Especially, on images with large noise intensity ($\sigma > 20$), LFA-BYY consistently produces the most robust and superior performances over the competing methods. Figure 6.6 demonstrates results on an image with noise intensity $\sigma = 100$, from which we notice that LFA-BYY can better preserve the detailed features of images polluted with large noise. The experiments echo that LFA-BYY can appropriately control the complexity of the LFA model (i.e., the dictionary) and learn the noise intensity per image under processing. In comparison, other methods either utilize a pre-trained dictionary (K-SVD, EPLL, Msi) or actually a general DCT basis (BM3D), without considering the issue of controlling the dictionary complexity. This is the reason why they all not only need to know the noise intensities in advance but also under-perform LFA-BYY, especially when the patches are polluted by large noises. In such cases, LFA-BYY will accordingly learn an LFA model with a reduced complexity to ignore unreliable details, while the competing methods still use the same dictionaries.

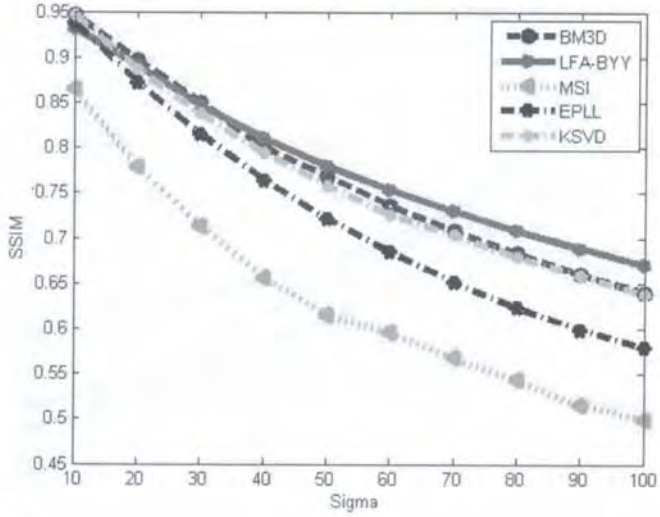

FIGURE 6.4

Average SSIM value of 24 natural images over different noise intensities.

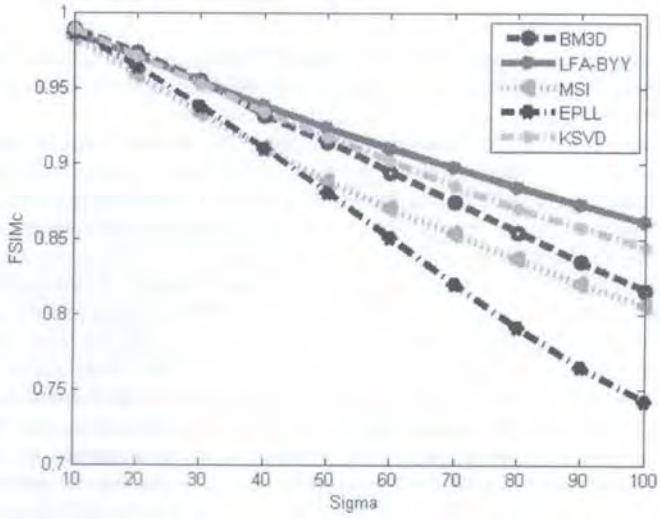

FIGURE 6.5

Average FSIM value of 24 natural images over different noise intensities.

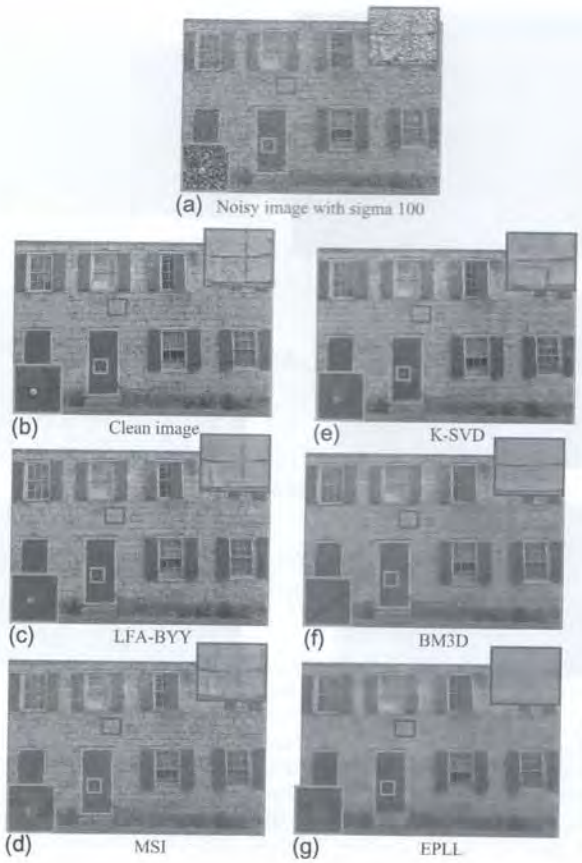

FIGURE 6.6

Results on an image with a real noise intensity $\sigma = 100$.

6.4.2 FAVORABLE FEATURES OF LFA-BYY

6.4.2.1 Robustness to different image datasets

LFA-BYY learns the dictionary of features from a noisy image adaptively per image under processing and thus the performances remain robust to different image datasets. Also, other algorithms use either a pretrained dictionary or a general basis, and thus are only suitable to a limited range of images. For example, as BM3D utilizes a DCT basis to describe features, it is more suitable for processing

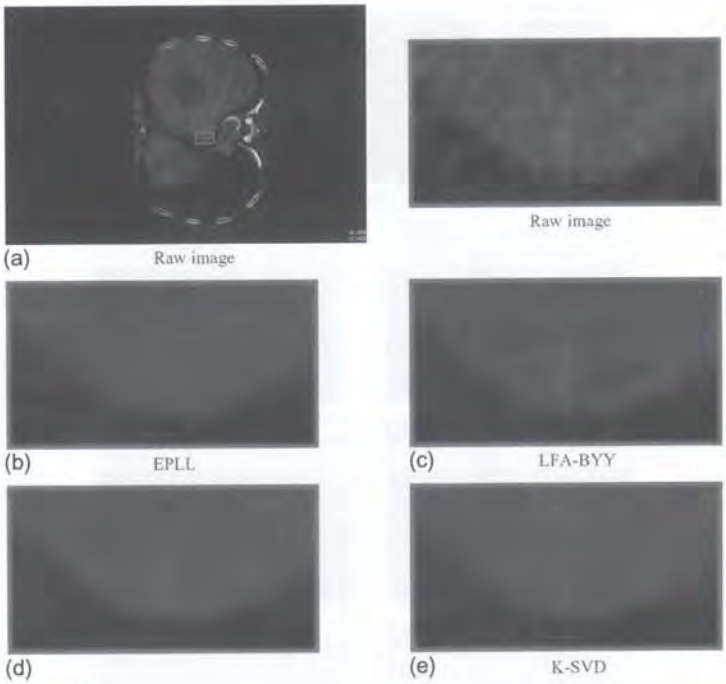

FIGURE 6.7

Comparative result on practical medical images. The structure shown in the box has some clinical meaning, but the noise around this structure may disturb the judgment of doctors. LFA-BYY enhances the clinical structure and removes the surrounding noise concurrently, while other algorithms remove the noise at the cost of smoothing detailed structures.

natural images. Figure 6.7 illustrates comparative results of denoising medical images. Significant detailed feature is smoothed by other algorithms while preserved by LFA-BYY.

6.4.2.2 Robustness to unknown noise intensity

LFA-BYY is not only able to learn the noise intensity per image under processing applicable to the heterogeneous noises on one image. In addition, another competing method needs a two-stage procedure, namely noise intensity estimation and image denoising with the estimated noise intensity. Not only can this procedure be highly

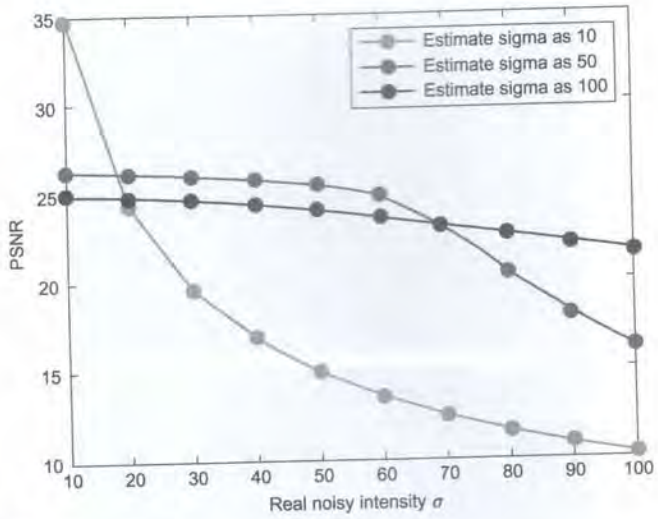

FIGURE 6.8

The performance of BM3D under different noise intensities where the estimated noise intensity is supposed to be 10, 20, and 100, respectively.

time-consuming but also noise intensity estimation is a difficult task while a poor estimation can seriously affect their performances.

Figure 6.8 demonstrates the sensitivity of BM3D to a noise intensity estimation. With a poor estimated noise intensity, the denoising result can be highly affected. When the estimated noise intensity is higher than the real one, the detailed features of an image will be considered as noise, resulting in an over-smoothed image as shown in Figures 6.9(c) and 6.10(c). On the other hand, when the estimated noise intensity is lower than the real one, the noise cannot be eliminated efficiently as shown in Figure 6.11. Moreover, the results shown in Figure 6.10 coincide with the results given in Figures 6.3, 6.4, and 6.5, which state that the LFA-BYY algorithm can preserve much more detail than BM3D when real σ is large.

6.4.2.3 No free parameters

There are usually many heuristically picked parameters in the existing denoising algorithms, and the performance of these algorithms can be highly affected by the inappropriate setting of these parameters. In contrast, LFA-BYY does not contain such parameters. All the knowledge comes from the image under processing, producing consistent performances on different images.

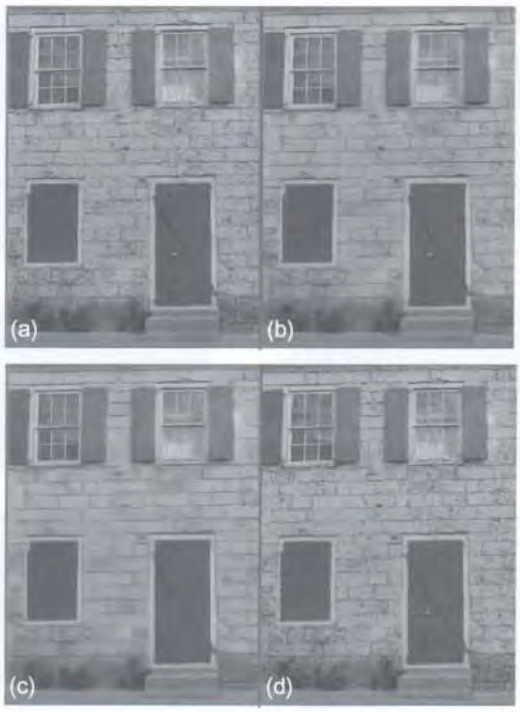

FIGURE 6.9

The denoising performances on one image with the true noise intensity $\sigma = 10$. (a) The denoised image by BM3D provided with the estimated intensity $\sigma = 10$; (b) the denoised image by BM3D provided with the estimated intensity $\sigma = 50$; (c) the denoised image by BM3D provided with the estimated intensity $\sigma = 100$; and (d) the denoised image by LFA-BYY with noise intensity determined automatically.

6.5 CONCLUDING REMARKS

The proposed novel image denoising method LFA-BYY learns the LFA model (i.e., the dictionary) per image whilst processing. With the help of the BYY harmony learning, LFA-BYY can appropriately control the dictionary complexity and learn the noise intensity from the present image under processing, while existing state-of-the-art methods have not considered the issues. In comparison with

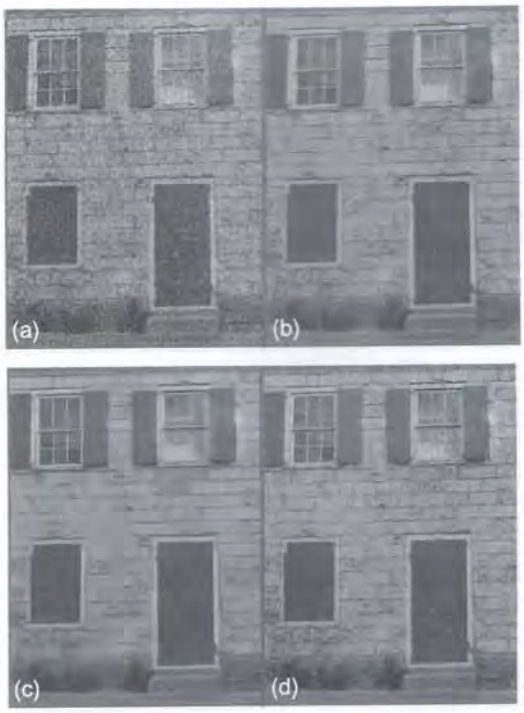

FIGURE 6.10

The denoising performances on one image with the true noise intensity $\sigma = 50$. (a) The denoised image by BM3D provided with the estimated intensity $\sigma = 10$; (b) the denoised image by BM3D provided with the estimated intensity $\sigma = 50$; (c) the denoised image by BM3D provided with the estimated intensity $\sigma = 100$; and (d) the denoised image by LFA-BYY with noise intensity determined automatically.

four methods, BM3D, K-SVD, EPLL, and Msi, on the benchmark Kodak image processing dataset that contains 24 natural images and also additional medical data, experiments have shown that LFA-BYY has not only obtained competitive results on images polluted by a small noise but also outperformed these competing methods when the noise intensity increases beyond a point, especially with significant improvements as the noise intensity becomes large.

The patch size $d \times d$ in this chapter is chosen following the previous nonlocal means methods. This size will influence the performance of image denoising

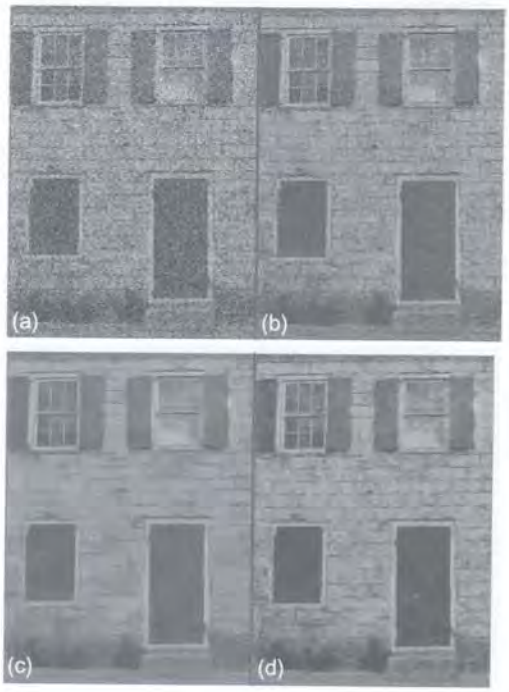

FIGURE 6.11

The denoising performances on one image with the true noise intensity $\sigma = 100$. (a) The denoised image by BM3D provided with the estimated intensity $\sigma = 10$; (b) the denoised image by BM3D provided with the estimated intensity $\sigma = 50$; (c) the denoised image by BM3D provided with the estimated intensity $\sigma = 100$; and (d) the denoised image by LFA-BYY with noise intensity determined automatically.

especially when images are polluted with strong noise. Further investigation for appropriate patch sizes and other possible improvements are left for further work.

REFERENCES

[1] A. Buades, B. Coll, J.-M. Morel, A review of image denoising algorithms, with a new one, Multiscale Model. Simul. 4 (2) (2005) 490-530.

[2] R.C. Gonzalez, R.E. Woods, Digital image processing (2002). Prentice Hall, Upper Saddle River, NJ.

[3] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311-4322.
[4] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (12) (2006) 3736-3745.
[5] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-D transform-domain collaborative filtering, IEEE Trans. Image Process. 16 (8) (2007) 2080-2095.
[6] D. Zoran, Y. Weiss, From learning models of natural image patches to whole image restoration, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 479-486.
[7] S. Gu, L. Zhang, W. Zuo, X. Feng, Weighted nuclear norm minimization with application to image denoising, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[8] L. Xu, BYY harmony neural networks, structural RPCL, and topological self-organizing on mixture models, Neural Netw. 15 (2002) 1125-1151.
[9] L. Xu, Bayesian-Kullback coupled Ying-Yang machines: unified learnings and new results on vector quantization, in: Proceedings of International Conference on Neural Information Processing, 1995, pp. 977-988.
[10] L. Xu, Further advances on Bayesian Ying-Yang harmony learning, Appl. Inform. (in press).
[11] L. Xu, Bayesian Ying-Yang system, best harmony learning, and five action circling, Front. Electr. Electron. Eng. China 5 (3) (2010) 281-328.
[12] G.E. Hinton, P. Dayan, M. Revow, Modeling the manifolds of images of handwritten digits, IEEE Trans. Neural Netw. 8 (1) (1997) 65-74.
[13] Z. Ghahramani, M.J. Beal, Variational inference for Bayesian mixtures of factor analysers, in: NIPS, 1999, pp. 449-455.
[14] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716-723.
[15] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461-464.
[16] C.S. Wallace, D.M. Boulton, An information measure for classification, Comput. J. 11 (2) (1968) 185-194.
[17] L. Xu, A. Krzyzak, E. Oja, Unsupervised and supervised classifications by rival penalized competitive learning, in: Proceedings. 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems, IEEE, 1992, pp. 496-499.
[18] L. Xu, A. Krzyzak, E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, IEEE Trans. Neural Netw. 4 (4) (1993) 636-649.
[19] R.M. Neal, G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, Springer, 1998, pp. 355-368.
[20] H. Attias, A variational Bayesian framework for graphical models, Adv. Neural Inform. Process. Syst. 12 (1-2) (2000) 209-215.
[21] L. Shi, S. Tu, L. Xu, Learning Gaussian mixture with automatic model selection: a comparative study on three Bayesian related approaches, Front. Electr. Electron. Eng. China 6 (2) (2011) 215-244.
[22] G. Chen, H. Pheng, H. Ann, L. Xu, Projection embedded BYY learning algorithm for Gaussian mixture based clustering, Appl. Inform. 1 (2014), 2.

[23] A. Corduneanu, C.M. Bishop, Variational Bayesian model selection for mixture distributions, in: Artificial Intelligence and Statistics, vol. 2001, Morgan Kaufmann, Waltham, MA, 2001, pp. 27-34.
[24] L. Xu, Bayesian Ying-Yang system and theory as a unified statistical learning approach: (i) unsupervised and semi-unsupervised learning, in: Brain-Like Computing and Intelligent Information Systems, Springer-Verlag, Heidelberg, 1997, 241-274.
[25] L. Xu, Bayesian Ying-Yang system and theory as a unified statistical learning approach: (iii) models and algorithms for dependence reduction, data dimension reduction, ICA and supervised learning, in: Lecture Notes in Computer Science: Proc. of International Workshop on Theoretical Aspects of Neural Computation, 1997, pp. 43-60.
[26] L. Xu, On essential topics of BYY harmony learning: current status, challenging issues, and gene analysis applications, Front. Electr. Electron. Eng. China 7 (2012) 147-196.
[27] S.K. Tu, L. Xu, Parameterizations make different model selections: empirical findings from factor analysis, Front. Electr. Electron. Eng. China 6 (2011) 256-274 (a special issue on Machine Learning and Intelligence Science: IScIDE2010 (B)).
[28] L. Shi, Z.-Y. Liu, S. Tu, L. Xu, Learning local factor analysis versus mixture of factor analyzers with automatic model selection, Neurocomputing 139 (2014) 3-14.
[29] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Image Process. 17 (1) (2008) 53-69.
[30] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, B. Zhang, Decomposable nonlocal tensor dictionary learning for multispectral image denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2949-2956.
[31] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600-612.
[32] L. Zhang, D. Zhang, X. Mou, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378-2386.