

NORTHWESTERN UNIVERSITY

Bayesian Techniques for Image Recovery

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering and Computer Science

By

Sevket Derin Babacan

EVANSTON, ILLINOIS

December 2009

UMI Number: 3386901

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3386901

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

© Copyright by Sevket Derin Babacan 2009

All Rights Reserved

## ABSTRACT

Bayesian Techniques for Image Recovery

Sevket Derin Babacan

In applications where degraded images are acquired and where improving the quality of the imaging system is not an option, or reproducing the scene conditions in order to acquire another image is not possible, computational approaches provide a powerful means for restoring images. Image recovery is the process of estimating the information lost due to the acquisition or processing system and obtaining images with high quality and/or resolution from a set of degraded images. There is a growing need for more advanced image recovery algorithms due to the proliferation of new application areas such as medical and nano-imaging, compressive sensing, and computational photography. In addition, image recovery research is currently being utilized for designing new imaging hardware. Finally, as image recovery is a particular case of inverse problems, most of its results can be applied to a wide range of problems. Therefore, research in image recovery will remain very active in the future and will be of high interest in both industry and academia.

This dissertation considers Bayesian image recovery methods. We have developed novel Bayesian formulations for image recovery which provide consistent and systematic modeling

of the image, the imaging process and other unknowns. The main contributions include 1) providing fully automated methods to use complex priors within a Bayesian framework to capture the statistics of the unknown image and yield high quality restorations, and 2) employing variational distribution approximations to account for the errors and uncertainties during the estimation process. A powerful aspect of our approach is that all algorithm parameters are embedded in our framework using a fully Bayesian formulation, and are estimated simultaneously with the unknown image.

Using this framework, we developed novel algorithms for a number of image recovery problems, i.e., image restoration, (single and multi-frame) blind deconvolution, super resolution, compressive sensing, and light-field imaging. We show that some of the current approaches in Bayesian image recovery are special cases of our framework and that our formulation can be applied to a general class of recovery and parameter estimation problems. Experimental results demonstrate that the proposed approaches provide competitive and in many cases superior performance compared to the state-of-the-art methods without utilizing any prior knowledge and user supervision.

## Acknowledgements

This work would not have been possible without the assistance and encouragement of a number of people.

First, I would like to thank my advisor, Professor Aggelos K Katsaggelos, for his continuous support and guidance that helped to overcome obstacles that seemed insurmountable, and for always stimulating new intellectual areas to explore. I consider myself very lucky to have him as my advisor and friend. I am also greatly indebted to Professor Rafael Molina from University of Granada, who stands as a symbol of hard work and modesty for me. I truly enjoyed the opportunity to work together with Professors Katsaggelos and Molina who created a friendly environment which proved to be very productive. This work equally belongs to them.

I would like to thank the other members of my committee, Professor Ying Wu, and Professor Alan Sahakian, for offering their valuable input and feedback in writing this dissertation.

I am delighted to have Professor Khalid Sayood from University of Nebraska as my life-long mentor. He provided constant guidance and encouragement during my graduate studies, and most importantly, he taught me to pursue the things I truly like in my professional life.

The five years I spent in Evanston were full of invaluable experiences for me, both professionally and personally. I would like to thank all my coworkers in the Image and Video Processing Laboratory at Northwestern University for creating a stimulating and friendly environment. I also had an amazing time with my friends at Northwestern University. The friendship of many people completed my life during my graduate studies and left everlasting memories.

Khalid, Fusun, Sena and Sinan deserve a very special thanks for their support and love, and for becoming my family here in the USA.

My deepest gratitude goes to my family, Funda, Muhamrem and Oytun Babacan, as this work would not have been possible without them. They were always there for me at every step I have taken. Every success in my life is dedicated to them.

*Anne ve babama*

## Table of Contents

ABSTRACT	3
Acknowledgements	5
List of Tables	12
List of Figures	15
Chapter 1. Introduction	27
1.1. Introduction	27
1.2. Scope of Thesis	28
1.3. Thesis Outline	30
Chapter 2. Background	31
2.1. Bayesian Framework for Image Restoration and Blind Deconvolution	34
2.2. Bayesian Modelling	35
2.3. Bayesian Inference Methods	45
2.4. Super Resolution	52
Chapter 3. Total Variation Image Restoration Using Variational Distribution Approximation	61
3.1. Introduction	61
3.2. Bayesian Modeling and Inference	64

3.3.	Hyperpriors, prior, and observation model used in TV image Deconvolution	67
3.4.	Bayesian Inference and Variational Approximation of the posterior distribution for TV image restoration	69
3.5.	Experimental Results	81
3.6.	Conclusions	94
 Chapter 4. Generalized Gaussian Markov Random Field Image Restoration Using Variational Distribution Approximation		
4.1.	Introduction	96
4.2.	Bayesian Modeling	97
4.3.	Inference and Variational Approximation	99
4.4.	Experimental Results	104
4.5.	Conclusions	107
 Chapter 5. Total Variation Blind Deconvolution Using A Variational Approach		109
5.1.	Introduction	109
5.2.	Hierarchical Bayesian Modeling	110
5.3.	Bayesian Inference and Variational Approximation of the posterior distributions	112
5.4.	Experimental Results	121
5.5.	Conclusions	138
 Chapter 6. Bayesian Blind Deconvolution from Differently Exposed Image Pairs		147
6.1.	Introduction	147
6.2.	Problem Formulation	150
6.3.	Hierarchical Bayesian Model	152

6.4. Variational Bayesian Inference	157
6.5. Calculation of Posterior Distribution Approximations	161
6.6. Experimental Results	169
6.7. Conclusions	175
Chapter 7. Variational Bayesian Super Resolution	180
7.1. Introduction	180
7.2. Problem Formulation	183
7.3. Hierarchical Bayesian Model	184
7.4. Variational Bayesian Inference	189
7.5. Experimental Results	203
7.6. Conclusions	217
Chapter 8. Bayesian Compressive Sensing using Laplace Priors	222
8.1. Introduction	222
8.2. Bayesian Modeling	226
8.3. Bayesian Inference	232
8.4. Experiments	244
8.5. Conclusions	254
Chapter 9. Bayesian Compressive Sensing Using Non-Convex Priors	257
9.1. Introduction	257
9.2. Bayesian modeling	258
9.3. Inference procedure	262
9.4. Experiments	269

9.5. Conclusions	278
Chapter 10. Compressive Light Field Imaging	281
10.1. Introduction	281
10.2. Prior Work	284
10.3. Compressive Sensing of Light-Fields	286
10.4. Hierarchical Bayesian Model For Reconstruction	292
10.5. Reconstruction Algorithm	299
10.6. Experimental Results	303
10.7. Conclusions	304
Chapter 11. Conclusions	306
References	309
Appendix A. Calculation of the image estimates in Algorithms 1 and 2	330
Appendix B. Calculation of required expected values in Algorithm 1	333
Appendix C. Calculation of required expected values in Algorithm 5	335

## List of Tables

3.1	Proposed Algorithm I	74
3.2	Proposed Algorithm II	78
3.3	ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 9.	85
3.4	ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a 9x9 uniform blur.	86
3.5	Posterior means of the distributions of the hyperparameters, ISNR, and number of iterations using <i>ALG1</i> for the Lena image with 40 dB and 20 dB BSNR using $\bar{\alpha}^o = 1/23.84$ , and $\bar{\beta}^o = 1/0.16$ and $\bar{\beta}^o = 1/16$ , respectively, for different values of $\gamma_\beta$ and $\gamma_\alpha$ .	94
4.1	Proposed Algorithm	101
5.1	Proposed Algorithm I	116
5.2	Proposed Algorithm II	122
5.3	ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 9.	124
5.4	ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 5.	126

5.5	Posterior means of the distributions of the hyperparameters, ISNR, and number of iterations using <i>TVI</i> for the Lena image with 40 dB and 20 dB BSNR using $\bar{\alpha}_{\text{im}}^o = 0.042$ , $\bar{\alpha}_{\text{bl}}^o = 4.6 \times 10^8$ , and $\bar{\beta}^o = 6.25$ , respectively, for different values of $\gamma_{\alpha_{\text{im}}}$ , $\gamma_{\alpha_{\text{im}}}$ and $\gamma_{\beta}$ .	142
6.1	Proposed algorithm	167
7.1	Proposed Algorithm I	200
7.2	Proposed Algorithm II	202
7.3	Mean PSNRs with standard deviations in 20 experiments provided by the SR algorithms in different SNR levels for the case where motion information is exact.	207
7.4	Mean PSNRs with standard deviations in 20 experiments provided by the SR algorithms in different SNR levels for the case motion information is inaccurate (see text).	211
7.5	Mean MSEs with standard deviations of the motion parameters in 20 experiments provided by the proposed methods in different SNR levels.	212
8.1	Proposed Algorithm	242
8.2	Average reconstruction errors, running times and number of nonzero components for multi-scale CS reconstruction of the <i>Mondrian</i> image.	254
9.1	Average reconstruction errors, running times and number of nonzero components for multi-scale CS reconstruction of the <i>Mondrian</i> image.	278

- 9.2 Average reconstruction errors when only the 849 largest signal coefficients from the estimates in Table 9.1 are kept. 278

## List of Figures

3.1	Graphical model showing relationships between variables.	70
3.2	(a) Lena image; degraded with a Gaussian shaped PSF with variance 9 and Gaussian noise of variance: (b) 0.16 (BSNR = 40 dB), (c) 1.6 (BSNR = 30 dB), (d) 16 (BSNR = 20 dB).	83
3.3	Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) <i>MOL</i> method (ISNR = 3.90 dB), (b) <i>BFO1</i> method (ISNR = 4.72 dB), (c) <i>BFO2</i> method (ISNR = 4.5 dB), (d) <i>ALG1</i> method (ISNR = 4.84 dB), and (e) <i>ALG2</i> method (ISNR = 4.64 dB).	88
3.4	Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) <i>MOL</i> method (ISNR = 2.45 dB), (b) <i>BFO1</i> method (ISNR = 3.02 dB), (c) <i>BFO2</i> method (ISNR = 2.47 dB), (d) <i>ALG1</i> method (ISNR = 3.06 dB), and (e) <i>ALG2</i> method (ISNR = 2.58 dB).	89
3.5	Evolution of ISNR using <i>ALG1</i> for different values of $\gamma_\alpha$ and $\gamma_\beta$ for the restoration of the Lena image blurred with a Gaussian with variance 9, and (a) BSNR = 40 dB; (b) BSNR = 20 dB.	93
3.6	Evolution of ISNR with varying $\gamma_\alpha$ and $\bar{\alpha}^o$ for Lena image degraded with Gaussian blur with variance 9 at (a) 40 dB BSNR and (b) 20 dB BSNR (Note that $\bar{\alpha}^o = d \cdot \hat{\alpha}$ ).	94

4.1	(a) Graphical model showing relationships between variables, (b) the directions for the first order differences around the pixel $i$ .	98
4.2	(a) Original Lena Image, (b) Image degraded by a Gaussian shaped PSF with variance 9 and Gaussian noise of variance 0.16 (BSNR=40dB), (c) Restored image using Algorithm 3 with $p = 1.8$ (ISNR = 4.15dB), (c) Restored image using Algorithm 4 with $p = 1.6$ (ISNR = 3.78dB).	106
4.3	ISNR values obtained by different $p$ values with Lena image degraded by (a) a Gaussian blur with variance 9 and (b) a 9x9 uniform blur with Gaussian noise (BSNR = 40dB and 20dB).	107
5.1	Graphical model showing relationships between variables.	113
5.2	(a) Shepp-Logan phantom image; degraded with a Gaussian shaped PSF with variance 9 and Gaussian noise of variance: (b) 0.18 (BSNR = 40 dB), (c) 18 (BSNR = 20 dB).	123
5.3	Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) <i>TV1</i> algorithm (ISNR = 2.53 dB), (b) <i>TV2</i> algorithm (ISNR = 2.95 dB), (c) <i>SAR1</i> algorithm (ISNR = 2.10 dB), (d) <i>SAR2</i> algorithm (ISNR = 2.42 dB), (e) <i>TV1-NB</i> algorithm (ISNR = 4.33 dB), and (f) <i>TV2-NB</i> algorithm (ISNR = 4.31 dB).	125
5.4	Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) <i>TV1</i> algorithm (ISNR = 2.62 dB), (b) <i>TV2</i> algorithm (ISNR = -2.54 dB), (c) <i>SAR1</i> algorithm (ISNR = 1.06 dB), (d)	

*SAR2* algorithm (ISNR = -0.29 dB), (e) *TV1-NB* algorithm (ISNR = 3.31 dB), and (f) *TV2-NB* algorithm (ISNR = 3.29 dB). 131

5.5 Restorations of the Lena image blurred with a Gaussian PSF with variance 5 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 3.19 dB), (b) *TV2* algorithm (ISNR = 3.29 dB), (c) *SAR1* algorithm (ISNR = 2.35 dB), (d) *SAR2* algorithm (ISNR = 2.57 dB), (e) *TV1-NB* algorithm (ISNR = 4.98 dB), and (f) *TV2-NB* algorithm (ISNR = 4.93 dB). 132

5.6 Restorations of the Lena image blurred with a Gaussian PSF with variance 5 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 1.39 dB), (b) *TV2* algorithm (ISNR = -4.39 dB), (c) *SAR1* algorithm (ISNR = 0.36 dB), (d) *SAR2* algorithm (ISNR = -0.23 dB), (e) *TV1-NB* algorithm (ISNR = 2.92 dB), and (f) *TV2-NB* algorithm (ISNR = 2.83 dB). 133

5.7 Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 3.07 dB), (b) *TV2* algorithm (ISNR = 3.36 dB), (c) *SAR1* algorithm (ISNR = 1.64 dB), (d) *SAR2* algorithm (ISNR = 1.81 dB), (e) *TV1-NB* algorithm (ISNR = 4.16 dB), and (f) *TV2-NB* algorithm (ISNR = 4.15 dB). 134

5.8 Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.47 dB), (b) *TV2* algorithm (ISNR = -1.41 dB), (c) *SAR1* algorithm (ISNR = 1.56 dB), (d) *SAR2* algorithm (ISNR = -0.15 dB), (e) *TV1-NB* algorithm (ISNR = 4.28 dB), and (f) *TV2-NB* algorithm (ISNR = 4.27 dB). 135

- 5.9 Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 5 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.05 dB), (b) *TV2* algorithm (ISNR = 3.79 dB), (c) *SAR1* algorithm (ISNR = 1.91 dB), (d) *SAR2* algorithm (ISNR = 1.30 dB), (e) *TV1-NB* algorithm (ISNR = 7.57 dB), and (f) *TV2-NB* algorithm (ISNR = 7.29 dB). 136
- 5.10 Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 5 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.09 dB), (b) *TV2* algorithm (ISNR = -2.89 dB), (c) *SAR1* algorithm (ISNR = 1.46 dB), (d) *SAR2* algorithm (ISNR = -0.17 dB), (e) *TV1-NB* algorithm (ISNR = 4.68 dB), and (f) *TV2-NB* algorithm (ISNR = 4.65 dB). 137
- 5.11 One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 9 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*. 138
- 5.12 One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 9 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*. 138
- 5.13 One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 5 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*. 139
- 5.14 One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 5 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*. 139

- 5.15 One-dimensional slice through the origin of the original and estimated PSFs  
in the restoration of the Shepp-Logan phantom degraded by a Gaussian  
with variance 9 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and  
*SAR2*. 140
- 5.16 One-dimensional slice through the origin of the original and estimated PSFs  
in the restoration of the Shepp-Logan phantom degraded by a Gaussian  
with variance 9 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and  
*SAR2*. 140
- 5.17 One-dimensional slice through the origin of the original and estimated PSFs  
in the restoration of the Shepp-Logan phantom degraded by a Gaussian  
with variance 5 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and  
*SAR2*. 141
- 5.18 One-dimensional slice through the origin of the original and estimated PSFs  
in the restoration of the Shepp-Logan phantom degraded by a Gaussian  
with variance 5 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and  
*SAR2*. 141
- 5.19 ISNR evolution for different values of confidence parameters for the  
Algorithm 1 (*TV1*) applied to “Lena” image degraded by a Gaussian with  
variance 9 and BSNR = 40dB. (a) For fixed  $\gamma_\beta = 0$ , (b) for fixed  $\gamma_{\alpha_{bl}} = 0$ ,  
(c) for fixed  $\gamma_{\alpha_{im}} = 0$ , and (d) for fixed  $\gamma_\beta = 1$ . 142
- 5.20 Some restorations of the Lena image blurred with a Gaussian PSF with  
variance 9 and 40 dB BSNR using the *TV1* algorithm utilizing prior

knowledge through confidence parameters and positivity and symmetry constraints on the estimated blur. (a)  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (b)  $\gamma_{\alpha_{im}} = 0$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , (c)  $\gamma_{\alpha_{im}} = 0.6$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , and (d)  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 1$ . 143

5.21 One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Lena image degraded by a Gaussian with variance 8 and BSNR = 40dB with algorithm *TVI*. (a) True PSF, Estimated PSF with (b)  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (c)  $\gamma_{\alpha_{im}} = 0$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , (d)  $\gamma_{\alpha_{im}} = 0.6$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , and (e)  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 1$ . 144

5.22 (a) Observed Saturn image. (b) Non-blind Restoration with *TV2-NB*, (c) Restoration with *TV2* with  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (d) Restoration with *TVI* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ; (e) Restoration with *TV2* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ; (f) Restoration with *SAR1* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ; 145

5.23 One-dimensional slice through the origin of the theoretical and estimated PSFs in the restoration of the Saturn image. (a) Theoretical PSF, estimated PSF (b) using *TV2* with  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , and with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$  using (c) *SAR1*, (d) *TVI* and (e) *TV2* 146

6.1 Graphical model showing relationships between variables. 157

6.2 (a) Original Ephesus image, (b) observed noisy image simulating a short-exposure acquisition. 171

6.3	Blurred images simulating long-exposure photographs. The point spread function (PSF) used to generate each image is shown below the corresponding image. All PSFs have support $21 \times 21$ pixels and the images are of size $430 \times 270$ pixels. The values of the PSFs are linearly mapped to the $[0,255]$ range for visualization purposes.	171
6.4	Restoration results using the proposed algorithm. The restored images are shown in the top row, and the corresponding recovered PSFs are shown below the images. The values of the PSFs are linearly mapped to the $[0,255]$ range for visualization purposes.	172
6.5	Restoration results using the <i>deconvblind</i> routine in MATLAB.	172
6.6	An outdoor image pair. (a) Short exposure image (brightness level is corrected), (b) long exposure image, (c) denoised short exposure image, (d) restored image using the proposed algorithm, and (e) recovered PSF (support: $51 \times 51$ ).	176
6.7	Center regions of the images shown in Figure 6.6. (a) Short exposure image, (b) long exposure image, (c) denoised short exposure image, (d) result of the proposed algorithm.	177
6.8	Real image example (courtesy of [224]). (a) Short exposure image, (b) long exposure image, (c) denoised short exposure image, (d) restored image using [224], (e) recovered PSF using the proposed algorithm (support : $41 \times 41$ ), and (f) restored image using the proposed algorithm.	178
7.1	( <i>Left</i> ) Original HR image, ( <i>right</i> ) Five synthetically generated LR images.	205

7.2	Mean PSNR values of SR algorithms for different input SNR levels when (a) exact motion information is available, and (b) motion information is inaccurate.	208
7.3	Example estimated HR images from different SR methods in the case when SNR=25dB and motion information is exact. Results of (a) Bicubic interpolation (PSNR = 17.14dB), (b) <i>ZMT</i> (PSNR = 20.55dB), (c) <i>RSR</i> (PSNR = 26.41dB), and the proposed methods (d) <i>ALG1</i> (PSNR = 28.75dB), and (e) <i>ALG2</i> (PSNR = 28.58dB).	209
7.4	Example estimated HR images from different SR methods in the case when SNR=45dB and motion information is exact. Results of (a) Bicubic interpolation (PSNR = 17.16dB), (b) <i>ZMT</i> (PSNR = 20.53dB), (c) <i>RSR</i> (PSNR = 33.56dB), and the proposed methods (d) <i>ALG1</i> (PSNR = 36.81dB), and (e) <i>ALG2</i> (PSNR = 35.85dB).	210
7.5	Example estimated HR images from different SR methods in the case when SNR=25dB and motion information is inaccurate (see text). Results of (a) Bicubic interpolation (PSNR = 17.14dB), (b) <i>ZMT</i> (PSNR = 17.47dB), (c) <i>RSR</i> (PSNR = 17.41dB), and the proposed methods (d) <i>ALG1</i> (PSNR = 28.47dB), and (e) <i>ALG2</i> (PSNR = 28.34dB).	212
7.6	Example estimated HR images from different SR methods in the case when SNR=45 dB and motion information is inaccurate (see text). Results of (a) Bicubic interpolation (PSNR = 17.16dB), (b) <i>ZMT</i> (PSNR = 17.49dB), (c) <i>RSR</i> (PSNR = 17.44dB), and the proposed methods (d) <i>ALG1</i> (PSNR = 35.11dB), and (e) <i>ALG2</i> (PSNR = 33.40dB).	213

7.7	Estimated motion parameters by the algorithms <i>ALG1</i> and <i>ALG2</i> when SNR = 25dB. (a) Estimated translation parameters, (b) Estimated rotation angles. The resulting MSEs of the estimated parameters are 0.0029 for <i>ALG1</i> and $7.91 \times 10^{-4}$ for <i>ALG2</i> .	214
7.8	Estimated motion parameters by the algorithms <i>ALG1</i> and <i>ALG2</i> when SNR = 45dB. (a) Estimated translation parameters, (b) Estimated rotation angles. The resulting MSEs of the estimated parameters are $2.48 \times 10^{-5}$ for <i>ALG1</i> and 0.0017 for <i>ALG2</i> .	215
7.9	Super resolution results (4x resolution increase) by (a) bicubic interpolation, (b) <i>EF</i> , (c) <i>ZMT</i> , (d) <i>RSR</i> , (e) <i>ALG1</i> and (f) <i>ALG2</i> .	218
7.10	Super resolution results (3x resolution increase) by (a) bicubic interpolation, (b) <i>EF</i> , (c) <i>ZMT</i> , (d) <i>RSR</i> , (e) <i>ALG1</i> and (f) <i>ALG2</i> .	219
7.11	Super resolution results (5x resolution increase) by (a) bicubic interpolation, (b) <i>EF</i> , (c) <i>ZMT</i> , (d) <i>RSR</i> , (e) <i>ALG1</i> and (f) <i>ALG2</i> .	220
8.1	Directed acyclic graph representing the Bayesian model.	232
8.2	Number of measurements $M$ vs reconstruction error for the noise-free case resulting from different values of $\lambda$ . (a) Uniform spikes $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to $M > 90$ are not shown as the error rates are negligible.	247

- 8.3 Number of measurements  $M$  vs reconstruction error with noisy observations with different values of  $\lambda$ . (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown as the error rates converged. 248
- 8.4 Number of measurements  $M$  vs reconstruction error for the noise-free case for different algorithms. (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown as the error rates converged. 251
- 8.5 Reconstruction of uniform spikes signal with  $N=512$ ,  $M=120$ , and  $T=20$ .  
 (a) Original Signal, (b) Observation, Reconstructions with (c) Laplace, (d) BP, (e) BCS, (f) OMP, (g) FAR, and (h) GPSR. All reconstructions have negligible errors except GPSR with reconstruction error = 0.2186.  
 The error bars in (c) and (e) correspond to the estimated variances of the coefficients. 252
- 8.6 Number of measurements  $M$  vs reconstruction error with noisy observations for different algorithms. (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown for clarity as the error rates converged. 253

8.7	Examples of reconstructed Mondrian images using a multi-scale compressive sensing scheme by (a) linear reconstruction (error = 0.13325), (b) BP (error = 0.13874, time = 76.555 s, no. of nonzero components = 4096), (c) StOMP with FDR thresholding (error = 0.1747, time = 6.48 s, no. of nonzero components = 2032), (d) StOMP with FAR thresholding (error = 0.14673, time = 19.759 s, no. of nonzero components = 1196), (e) BCS (error = 0.14233, time = 16.086 s, no. of nonzero components = 1145) and (f) Laplace (error = 0.14234, time = 15.982, no. of nonzero components = 1125).	255
9.1	Graphical models representing the dependencies within the proposed Bayesian model.	262
9.2	Reconstruction errors obtained by the proposed method for different values of $p$ while varying the number of measurements $M$ ; (a) when no observation noise is present, and (b) when Gaussian noise with standard deviation 0.03 is added to the measurements.	271
9.3	Reconstruction errors obtained by the proposed method with varying the number of measurements $M$ for $p$ -values 0.01, 0.1, 0.3, 0.05, 0.7 and 1. The measurements are noiseless in (a) and Gaussian noise with standard deviation 0.03 is added to the measurements.	274
9.4	Comparison between the proposed method BCS-lp and IRLS algorithm for different $p$ values when the observations are (a) noiseless and (b) degraded with Gaussian noise with standard deviation 0.03.	275

9.5	Comparison between a number of CS reconstruction algorithms with varying number of measurements $M$ with (a) noiseless and (b) noisy measurements.	276
9.6	Examples of reconstructed Mondrian images using a multi-scale compressive sensing scheme. ( <i>Top-left</i> ) Original image; Reconstructed images by ( <i>Top-right</i> ) GPSR; ( <i>Middle-left</i> ) BP; ( <i>Middle-right</i> ) BCS; ( <i>Bottom-left</i> ) BCS-Laplace; ( <i>Bottom-right</i> ) BCS-lp.	280
10.1	The basic principle of utilizing a coded aperture to obtain light field images. The angular images are shown in (a), (b) and (c) when only corner blocks of the aperture are left open. Both horizontal and vertical parallax can be observed between these images (Horizontal and vertical dashed lines are shown to clearly denote the vertical and horizontal parallax, respectively). Figure (d) shows a captured image with the randomly coded aperture used in the proposed compressive sensing light field camera. All images are from a synthetic light field image (see Sec. 10.6).	288
10.2	Number of measurements $M$ vs relative reconstruction error (average over 50 runs).	304
10.3	Reconstruction examples. (a) Original angular image, reconstructed images from (b) 11 measurements (relative reconstruction error = $3.4 \times 10^{-4}$ ) and (c) 21 measurements (relative reconstruction error = $1.4 \times 10^{-5}$ ).	305

## CHAPTER 1

### Introduction

#### 1.1. Introduction

Image acquisition systems introduce degradation and noise to the image in many practical applications. The degradation might be, for instance, due to the finite resolution of the acquisition instrument, the relative motion between the imaging system and the scene, or atmospheric turbulence, while the noise can originate from the image formation process and transmission medium. The goal of image recovery is to restore the original scene from the degraded observations. Image recovery mainly consists of three problems: Image restoration, where the degradation introduced by the imaging system is assumed known or obtained experimentally; blind deconvolution, where only the observation is known; and super resolution, where the goal is to recover a higher resolution image from low-resolution observations. In addition, new recovery problems are being introduced at a face pace with the proliferation of new application areas such as medical and nano-imaging, compressive sensing, and computational photography.

All of these problems have been extensively studied and widely applied in many areas in science and engineering. These areas include astronomical imaging [131] [208], remote sensing [39], microscopy [96], medical imaging [161, 205], optics [201, 184], photography [250, 218], super-resolution applications [212], and motion tracking applications [66], among others.

An important question is the practical benefits of image recovery. Can one simply use a better image acquisition system to obtain higher-quality observations? Although this will

increase the quality of the images, due to physical limits and the practical cost of the imaging systems, and the suboptimality of the imaging conditions, the desired level can mostly not be reached. Moreover, existing images that can not be acquired again can have significant distortions. Signal processing based approaches are of high practical value in all of these cases.

Bayesian techniques are of the most commonly used methods in image recovery procedures. These approaches are oriented towards stochastic modeling of the degradations, observations and the original images and applying an inverse procedure to obtain approximations of the original images. Bayesian methods provide a systematic and consistent way to incorporate prior knowledge and desired constraints in the restoration. Additionally, most of the approaches in the recovery literature can conveniently be reproduced in a Bayesian framework [33].

## 1.2. Scope of Thesis

In this work Bayesian techniques for image recovery are studied in several novel forms. We will formulate the image recovery problems from a probabilistic perspective, and introduce novel frameworks and algorithms. The main novelty introduced in this work consists of introducing new priors for the image and the blur, full Bayesian analysis of the recovery problem, i.e., including parameter estimation additional to the estimation of the unknown original image and the degradation, and novel means for the inverse procedures.

This thesis covers a number of problems in image recovery, namely, image restoration, blind deconvolution, super resolution, and compressive sensing. We will follow a similar formulation for all problems. All unknowns and observed variables are treated as stochastic unknown variables in a Bayesian framework, and the joint probability distribution is defined incorporating these variables. By defining prior distributions and likelihood functions, these variables

can be inferred using the posterior distribution. In our work, we utilize variational distribution approximations which became very popular recently for reasons that will be discussed in this thesis. Our formulation of the recovery problems will provide approximations to the distributions of the unknown rather than point estimates. This new approach is useful in determining the uncertainty of the estimates, and in drawing new estimates from the approximated distribution. The proposed approaches are novel examples of spatially adaptive image recovery, which outperform the traditional spatially-nonadaptive approaches.

Moreover, utilizing the close connection between traditional image recovery methods and compressive sensing, we developed two novel methods for compressive sensing reconstruction problem. We have utilized non-Gaussian prior models, Laplace and non-convex priors based on  $l_p$ -norms, to impose sparsity to high degree on the solutions. The proposed algorithm based on Laplace priors provides competitive reconstruction performance to the state-of-the-art while having a low computation complexity. On the other hand, the proposed method based on non-convex signal priors achieve state-of-the-art reconstruction performance at the expense of increased computation complexity. We show that both methods include some of the existing methods in the literature as special cases, and therefore lead to useful insights for further improvement.

Finally, we developed a novel imaging application using compressive sensing ideas. Specifically, by utilizing a randomly coded non-refractive mask in front of the aperture, incoherent measurements of the light passing through different regions are encoded in the captured images. A novel reconstruction algorithm is proposed to recover the original light field image from these acquisitions. Using the principles of compressive sensing, we demonstrate that light field images with high angular dimension can be captured with only a few acquisitions.

### 1.3. Thesis Outline

This thesis is organized as follows. In Chapter 2, we introduce the problems image restoration, blind deconvolution and super resolution, and review the existing approaches in the literature. In Chapter 3 we develop a novel image restoration method based on TV-priors using a Bayesian formulation. An approach based on a generalized version of this prior is presented in Chapter 4. In Chapter 5 we present two novel algorithms for blind deconvolution based on similar formulations to image restoration. Chapter 6 presents a multi-frame blind deconvolution method specifically designed for image recovery from differently exposed image pairs. A variational Bayesian super-resolution method with simultaneous motion estimation is presented in Chapter 7. In Chapter 8 we formulate the compressive sensing reconstruction problem from a Bayesian perspective, and proposed the use of Laplace priors to impose sparsity to a higher extent than existing methods. Chapter 9 presents another Bayesian compressive sensing recovery method using non-convex  $l_p$ -norm based priors. In Chapter 10 we propose an application of compressive sensing for light-field image acquisition. Finally, chapter 11 contains concluding remarks.

## CHAPTER 2

### Background

In this chapter we will present a review of the image restoration, blind deconvolution and super resolution areas. Our approach in this review is to formulate the methods within the Bayesian framework. While many methods were developed by other formulations, they can also be derived using Bayesian formulations, and reviewing them in this fashion provides a consistent means to compare them.

We will start with the review of image restoration and blind deconvolution approaches since the formulations of these two problems are very close to each other. Generally speaking, blind deconvolution is a generalization of the image restoration problem where the blur is also unknown. Our review, however, will not be mutually exclusive from super resolution. We will also give examples from the super resolution while presenting specific elements in Bayesian analysis. We will conclude this chapter with Section 2.4 which is entirely devoted to super resolution. A brief review of compressive sensing is provided in Chapter 8.

In most applications the acquired images represent a degraded version of the original scene. These applications include astronomical imaging (e.g., using ground-based imaging systems or extraterrestrial observations of the earth and the planets), commercial photography, medical imaging (e.g., x-rays, digital angiograms, autoradiographs), and molecular and cellular bioimaging [120] [19] [205] [33]. The degradation can be due to the atmospheric turbulence, the relative motion between the camera and the scene, and the finite resolution of the acquisition instrument.

A general degradation model is [120]

$$(2.1) \quad y(i, j) = S \left\{ \sum_m \sum_n h(i, j, m, n) x(m, n) \right\} \odot n(i, j),$$

where  $S(\cdot)$  represents a nonlinear function,  $h(i, j, m, n)$  is the impulse response of the blurring system at the location  $(i, j)$ ,  $n(i, j)$  is additive noise, and  $\odot$  represents pointwise operation. This is the most general form of a degradation model, since it includes pointwise nonlinearities such as saturation, nonstationary aberrations, spatial distortions, inter-channel effects, multiplicative noise, noise from several sources etc. However, most of the work in the literature does not take nonlinearity and signal-dependent noise into account (see, however, [223, 132, 189]), such that (2.1) reduces to

$$(2.2) \quad y(i, j) = \sum_m \sum_n h(i, j, m, n) x(m, n) + n(i, j),$$

The continuous version of this equation is called Fredholm integral equation of the first kind. This degradation model has found limited use in the literature, primarily due to the difficulty of estimating the spatially varying blur  $h(i, j, m, n)$ . In most practical applications, the blur can be assumed as space-invariant so that the following degradation model can be used

$$(2.3) \quad y(i, j) = \sum_m \sum_n h(i - m, j - n) x(m, n) + n(i, j),$$

We note that (2.2) and (2.3) can be written in matrix-vector form, by lexicographically ordering arrays  $y(i, j)$ ,  $x(i, j)$  and  $n(i, j)$ . Assuming that the images are of size  $m \times n = N$ , the degradation model can be written as

$$(2.4) \quad \mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n},$$

where  $\mathbf{y}$ ,  $\mathbf{x}$  and  $\mathbf{n}$  are  $N \times 1$  vectors obtained by some ordering of the corresponding images, and  $\mathbf{H}$  is a  $N \times N$  matrix. Note that (2.4) can also be written as

$$(2.5) \quad \mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{n},$$

The ordering to obtain the matrix-vector forms is typically lexicographic, although other forms such as raster-scan or interlaced are also possible. In the case of space-invariant degradation systems, the matrix  $\mathbf{H}$  in (2.4) is block-Toeplitz with Toeplitz blocks (BTTB), and can be approximated by a block circulant matrix with circulant blocks (BCCB). BCCB matrices have the very useful property that their eigenvalues are the 2D discrete Fourier coefficients of their defining sequences, and their eigenvectors are defined using their Fourier kernels. Using this property, Equation (2.4) can be written as follows

$$(2.6) \quad \mathbf{Y}(k, l) = \mathbf{H}(k, l)\mathbf{X}(k, l) + \mathbf{N}(k, l),$$

where  $\mathbf{Y}(k, l)$ ,  $\mathbf{H}(k, l)$ ,  $\mathbf{X}(k, l)$  and  $\mathbf{N}(k, l)$  are 2D Fourier transforms of the sequences  $y(i, j)$ ,  $h(i, j)$ ,  $x(i, j)$ , and  $n(i, j)$ , respectively. Note that this equation is valid under the assumption of

circular convolution, although linear convolution can always be achieved by appropriate zero padding of the 2D arrays.

The image restoration problem calls for finding an estimate of  $\mathbf{x}$  given  $\mathbf{y}$ ,  $\mathbf{H}$ , and knowledge about  $\mathbf{n}$  and possibly  $\mathbf{x}$  [120]. The literature on image restoration is rich (reviews and classifications of the major approaches can be found in [120], [19], [54], and references therein). Methods based on Bayesian formulations are of the most commonly used methods in the image restoration literature. Since this work is also based on a Bayesian formulation, we will focus on the literature of Bayesian methods. However, it should be noted that most of the other methods can also be derived using a Bayesian framework. Examples of such methods and their formulation under a Bayesian framework will also be given in what follows.

## 2.1. Bayesian Framework for Image Restoration and Blind Deconvolution

The fundamental principle of the Bayesian formulation is to treat all parameters and observable variables as unknown stochastic quantities, and form probabilistic distributions for these unknowns. Therefore, the original image  $\mathbf{x}$ , the blur  $\mathbf{h}$ , and the noise  $\mathbf{n}$  in (2.4) are treated as samples of random fields, with corresponding *prior* probability density functions (PDFs) that model our knowledge about the original image and the degradation process. Additionally, the PDFs of these unknowns depend on parameters  $\Omega$ , which are termed *hyperparameters*. The goal in Bayesian formulation is to form a joint global distribution using all unknowns, and perform inference using this distribution.

The hyperparameters  $\Omega$  can be assumed known, estimated separately from  $\mathbf{x}$  and  $\mathbf{h}$ , or estimated simultaneously by adopting a *hierarchical* Bayesian framework where they are also

assumed unknown and their PDFs are formed. The PDFs of the hyperparameters are called *hyperprior* distributions. The hierarchical model allows us to write the joint global distribution

$$(2.7) \quad p(\mathbf{x}, \mathbf{h}, \mathbf{y}, \Omega) = p(\Omega)p(\mathbf{x}, \mathbf{h}|\Omega)p(\mathbf{y}|\Omega, \mathbf{x}, \mathbf{h}),$$

Typically, we assume that  $\mathbf{x}$  and  $\mathbf{h}$  are *a priori* conditionally independent, given  $\Omega$ , i.e.,  $p(\mathbf{x}, \mathbf{h}|\Omega) = p(\mathbf{x}|\Omega)p(\mathbf{h}|\Omega)$ . The inference is performed using the posterior

$$(2.8) \quad p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \Omega)p(\mathbf{x}|\Omega)p(\mathbf{h}|\Omega)p(\Omega)}{p(\mathbf{y})}$$

In the following sections we study first various prior models for the image, blur, and hyperparameters that appeared in the literature. We will then proceed to analyze inference models for their estimation.

## 2.2. Bayesian Modelling

### 2.2.1. Observation Model

Due to the model in (2.4), the observation model is related to the PDF of the noise  $\mathbf{n}$ . A typically used model is stationary zero mean independent white Gaussian noise with distribution  $\mathcal{N}(\mathbf{n}|0, \beta^{-1})$ , that is,

$$(2.9) \quad p(\mathbf{n}) = p(\mathbf{y}|\mathbf{x}, h, \beta) = \left( \frac{\beta}{2\pi} \right)^{N/2} \exp \left[ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2 \right],$$

where  $\beta^{-1}$  denotes the variance.

Alternative noise models, such as Poisson noise arising in low-intensity imaging, or nonstationary noise are also assumed in certain models.

### 2.2.2. Parametric Prior Blur Models

Prior blur models can be chosen to be parametric, in which case  $p(\mathbf{h})$  is usually a uniform distribution. The unknown parameters of this parametric form may be experimentally computed, or be estimated using, for example, *Maximum Likelihood* methods (see [136, 127]). In the following subsections we will focus on the most popular parametric blur models in the literature.

Note that any blur model satisfies three constraints:

- Positivity:  $h(i, j) \geq 0$
- The blur PSF is real valued when the images are real.
- Energy conservation:  $\sum_i \sum_j h(i, j) = 0$ .

**Linear Motion Blur.** In general, relative motion of the camera and scene results in a temporal integration. If the camera movement or object motion is fast relative to the exposure period, it can be approximated as a linear motion blur, which is a 1D averaging filter.

An example of horizontal motion blur model is given by ( $L$  an even integer)

$$(2.10) \quad h(i, j) = \begin{cases} \frac{1}{L+1}, & -\frac{L}{2} \leq i \leq \frac{L}{2}, \\ & j = 0 \\ 0, & \text{otherwise.} \end{cases}$$

**Atmospheric Turbulence Blur.** This type of blur is common in remote sensing and aerial imaging applications. For long term exposure through the atmosphere a Gaussian PSF model is

a reasonably well approximation:

$$(2.11) \quad h(i, j) = K e^{-\frac{i^2 + j^2}{2\sigma^2}},$$

where  $K$  is the normalizing constant and  $\sigma^2$  is the variance that determines the severity of the blur. Alternative blur atmospheric blur models have been suggested in [164, 171]. In these works the PSF is approximated by the function

$$(2.12) \quad h(i, j) \propto (1 + \frac{i^2 + j^2}{R^2})^{-\delta},$$

where  $\delta$  and  $R$  are unknown parameters.

**Out-of-Focus Blur.** Photographical defocusing is a very common type of blurring, and it is caused by primarily by the finite aperture of the camera. Although a complete model of defocus blur depends on many parameters such as focal length, aperture number of the lens and the distance between the objects and camera, a uniform circular PSF model is generally used as an approximation, that is,

$$(2.13) \quad h(i, j) = \begin{cases} \frac{1}{\pi r^2}, & \sqrt{i^2 + j^2} \leq r \\ 0, & \text{otherwise.} \end{cases}$$

The uniform 2D blur is sometimes used as a cruder approximation to the out-of-focus blur; and it is also used as a model for sensor pixel integration in super-resolution restoration. This model is defined (with  $L$  an even integer) as

$$(2.14) \quad h(i, j) = \begin{cases} \frac{1}{(L+1)^2}, & -\frac{L}{2} \leq (i, j) \leq \frac{L}{2} \\ 0, & \text{otherwise.} \end{cases}$$

### 2.2.3. Prior Image and Blur Models

The prior distributions  $p(\mathbf{x}|\Omega)$  and  $p(\mathbf{h}|\Omega)$  should reflect our beliefs about the structure of  $\mathbf{x}$  and  $\mathbf{h}$ , and also constrain the space of possible solutions for them. This is necessary due to the ill-posed nature of the image restoration and blind deconvolution problems, and can also be interpreted under regularization. Several constraints on the image and the blur can be made, such as smooth, piecewise-smooth or textured. These descriptions can be modeled in a stochastic sense by forming prior distributions. A general exponential model is given by

$$(2.15a) \quad p(\mathbf{x}|\Omega) = \frac{1}{Z_x(\Omega)} \exp[-U_x(\mathbf{x}, \Omega)]$$

$$(2.15b) \quad p(\mathbf{h}|\Omega) = \frac{1}{Z_h(\Omega)} \exp[-U_h(\mathbf{h}, \Omega)]$$

where  $U(\cdot)$  are called the energy functions, and  $Z_x$  and  $Z_h$  are the normalizing terms. They may be assumed constant if the hyperparameters are known, or they must be calculated from  $\int \exp[-U_x(\mathbf{x}, \Omega)] d\mathbf{x}$  and  $\int \exp[-U_h(\mathbf{h}, \Omega)] d\mathbf{h}$ , respectively. Many different image and blur models in the literature can be put in the form of these exponential models. In the following subsections we will give details of some particular cases.

**2.2.3.1. Stationary Gaussian Models.** The most common model is the class of Gaussian models provided by  $U_x = \frac{1}{2}\alpha \|\mathbf{Lx}\|^2$ . Then, if  $\det|\mathbf{L}| \neq 0$ , the term  $Z_x$  in (2.15) becomes simply

$(2\pi)^{\frac{N}{2}} \alpha^{-\frac{N}{2}} \det|\mathbf{L}|^{-1}$ , which if we use a fixed stationary form for  $\mathbf{L}$  is simple to calculate. These models are often termed Simultaneous Autoregression (SAR) or Conditional Autoregression (CAR) models [200].

In the most basic case, where  $\mathbf{L} = \mathbf{I}$ , constraints are imposed on the magnitude of the intensity distribution of  $\mathbf{x}$ . A more common choice is  $\mathbf{L} = \mathbf{C}$ , where  $\mathbf{C}$  is the discrete Laplacian operator. Note that this selection for  $\mathbf{L}$  imposes constraints on the derivatives of the image. For instance, Molina *et al.* [169] used this model for both image and blur, giving

$$(2.16a) \quad p(\mathbf{x}|\alpha_{im}) \propto \alpha_{im}^{N/2} \exp\left[-\frac{1}{2}\alpha_{im} \|\mathbf{Cx}\|^2\right]$$

$$(2.16b) \quad p(\mathbf{h}|\alpha_{bl}) \propto \alpha_{bl}^{M/2} \exp\left[-\frac{1}{2}\alpha_{bl} \|\mathbf{Ch}\|^2\right].$$

The SAR model is suitable for  $\mathbf{x}$  and  $\mathbf{h}$  if it is assumed that the luminosity distribution is smooth on the image domain, and that the blur is a partially smooth function.

**2.2.3.2. Autoregressive Models.** A class of algorithms (see e.g. [136, 127]) model the observation  $\mathbf{y}$  as an Autoregressive Moving Average (ARMA) process. The observation equation (2.2) forms the Moving Average (MA) part of the model, whereas the original image is modeled as a 2-D Autoregressive (AR) process:

$$(2.17) \quad \mathbf{x} = \mathbf{Ax} + \mathbf{v},$$

where  $\mathbf{A}$  has a BTTB form, and  $\mathbf{v}$  is the excitation noise signal, driving the AR process. Assuming that  $\mathbf{v}$  is independent of the image,  $p(\mathbf{x})$  will be in the form of (2.15), with  $U_x = \frac{1}{2} \|\mathbf{(I-A)x}\|_{\Lambda_v}^2$

and  $Z_x = (2\pi)^{\frac{N}{2}} \det |\Lambda_v|^{\frac{1}{2}} \det |\mathbf{I} - \mathbf{A}|^{-1}$ , where  $\Lambda_v$  is the covariance matrix of  $\mathbf{v}$ . Note that unlike the SAR model, the AR coefficients also have to be estimated.

A related formulation to the stationary ARMA model is also considered by Katsaggelos and Lay in [127, 139, 126]. In these works, the AR model parameters are not estimated directly, but rather the defining sequence of the matrix  $\Lambda_x$  is found in the discrete frequency domain, along with the other parameters, under the assumption that the image model is stationary.

**2.2.3.3. Markov Random Field Models.** A class of models encountered extensively in image segmentation [70], classical image restoration [93], and also in super-resolution restoration [207] and BD [255, 65] are the Markov Random Field (MRF) models [248]. They are usually derived using local spatial dependencies.

We define the Gibbs distribution by setting  $U = \sum_{c \in \mathcal{C}} V_c(\mathbf{x})$  in (2.15), where  $V_c(\mathbf{x})$  is a *potential function* defined over *cliques*  $c$  in the image [248], and  $Z$  is termed the partition function. If quadratic potential functions are used, i.e.,  $V_c(\mathbf{x}) = (\mathbf{d}_c^T \mathbf{f})^2$ , we obtain the Gaussian Markov Random Field (GMRF) [37] or CAR [200] model, and the Gibbs distribution becomes a Gaussian:

$$(2.18) \quad p(\mathbf{x}) = \frac{1}{Z} \exp \left[ -\mathbf{x}^T \mathbf{B} \mathbf{x} \right] = \frac{1}{Z} \exp \left[ - \sum_{c \in \mathcal{C}} \mathbf{x}^T \mathbf{B}_c \mathbf{x} \right],$$

where  $\mathbf{B}_c$  is obtained from  $\mathbf{d}_c$  and satisfies  $[\mathbf{B}_c]_{\mathbf{s}_1, \mathbf{s}_2}$  are only non-zero when pixels  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are neighbors. Typically the vectors  $\mathbf{d}_c$  represent finite difference operators. The partition function is now equal to  $(2\pi)^{\frac{N}{2}} \det |\mathbf{B}|^{-\frac{1}{2}}$ .

Generalised Gaussian MRFs (GGMRFs) can also be obtained from this formulation with arbitrary non-quadratic potentials of a similar functional form:  $V_c(\mathbf{x}) = \rho(\mathbf{d}_c^T \mathbf{f})$ , where  $\rho$  is some

(usually convex) function, such as the *Huber function* [37] or p-norm (with  $p \geq 1$ ) based function,  $\rho(u) = |u|^p$ . This is similar to the use of potential functions used in anisotropic diffusion methods, the motivation being edge-preservation in the reconstructed image. Other extensions to the model consider hierarchical, or Compound GMRFs (CGMRFs), also with the goal of avoiding over-smoothing of edges [114, 93].

**2.2.3.4. Anisotropic Diffusion and Total Variation Type Models.** This class of priors incorporate non-quadratic functions on the image, with the aim of preserving edges by not over-penalizing discontinuities, i.e. outliers in the image gradient distribution, see [104, 54] for a unifying view of the probabilistic and variational approaches. Methods using this type of priors usually begin with a regularization formulation in the continuous image domain resulting in a Partial Differential Equations (PDEs) to be solved. However, these approaches can also be represented in Bayesian formulation and also reformulated in discrete domain.

The generalized regularization approach using anisotropic diffusion has been proposed by You and Kaveh [252]. In this formulation, convex functions  $\kappa(\cdot)$  and  $v(\cdot)$  of the image gradient  $|\nabla x(s)|$  and the PSF gradient  $|\nabla h(s)|$  respectively are used in defining regularization functionals:

$$(2.19a) \quad \mathcal{E}(x) = \int_{S_x} \kappa(|\nabla x(s)|) \, ds$$

$$(2.19b) \quad \mathcal{E}(h) = \int_{S_h} v(|\nabla h(s)|) \, ds.$$

This is in analogy with standard regularization procedures. Variational calculus is used to minimize (2.19a)-(2.19b), which results in a PDE for each variable. For instance, the solution for  $x$

in (2.19a) is given by:

$$(2.20) \quad \nabla_x \mathcal{E}(x) = \nabla \cdot \left( \frac{\kappa'(|\nabla x|)}{|\nabla x|} \nabla x \right) = 0.$$

Using a time evolution steepest descent method, we obtain the following PDE

$$(2.21) \quad \frac{\partial \hat{x}}{\partial t} = -\nabla_f \mathcal{E}(\hat{x}),$$

which clearly represents an anisotropic diffusion process. Thus, as time  $t$  progresses, directional smoothing occurs depending on the local image gradient. The strength and type of smoothing depends on the *diffusion coefficient* or *flux variable*,  $c$ , which is given by

$$(2.22) \quad c(|\nabla x|) = \frac{\kappa'(|\nabla x|)}{|\nabla x|}$$

Appropriate choice of  $c$  (or  $\kappa$ ) results in various types of restorations. For instance,  $\kappa(x) = \frac{1}{2}x^2$  and hence  $c(|\nabla x|) = 1$ , and  $\nabla_x \mathcal{E}(x) = \nabla^2(x)$ , i.e., a Laplacian operator [251], results in standard spatially-invariant isotropic regularization, or a CAR model. Another choice is  $\kappa(x) = x$  and hence  $c(|\nabla x|) = \frac{1}{|\nabla x|}$ , which results in *Total Variation* (TV) norm [58]. In this case, smoothing is only performed in the direction parallel to the edges, and smoothing orthogonal to the edges is completely suppressed.

For the two cases represented above, the corresponding image priors can be written as

$$(2.23) \quad p(\mathbf{x}) \propto \exp \left[ -\alpha_{im} \sum_i ((\Delta^h \mathbf{x})_i^2 + (\Delta^v \mathbf{x})_i^2) \right]$$

for the Laplacian; and

$$(2.24) \quad p(\mathbf{x}) \propto \exp \left[ -\alpha_{\text{im}} \sum_i \sqrt{(\Delta^h \mathbf{x})_i^2 + (\Delta^v \mathbf{x})_i^2} \right]$$

for the TV norm, where  $\Delta_i^h$  and  $\Delta_i^v$  are linear operators corresponding to horizontal and vertical first order differences, at pixel  $i$ , respectively.

Many other diffusion coefficients are proposed in the literature, including very complex structural operators (see [242] for a review). An interesting method is [252], where a combination of the Laplacian and TV is used. The smoothing strength is increased using the Laplacian in smooth areas with low gradient magnitude, and decreased using the TV norm in areas where large intensity transitions occur in order to preserve edges while still removing noise.

Šroubek and Flusser [218] use a similar scheme to those already mentioned, but they write the anisotropic diffusion model in the form of (2.15) using the following discretization of (2.19a)

$$(2.25) \quad p(\mathbf{x}, c(\mathbf{x})) = \frac{1}{Z_x} \exp \left[ -\frac{1}{2} \mathbf{x}^T \mathbf{B}(c) \mathbf{x} \right]$$

The diffusion is set equal to the edge strengths between two pixels in a hidden line process, such that a spatially-varying weights matrix  $\mathbf{B}$  can be formed from local image gradients. Similar formulations are also proposed in [250, 135, 124]. Note that a very generalized formulation is also proposed in the regularization context in [125].

#### 2.2.4. Hyperprior Models

The estimation of the hyperparameters  $\Omega$  is an important problem since they determine the performance of the algorithms and therefore play an important role in Bayesian image restoration,

blind deconvolution and super resolution. This estimation problem is introduced in the hierarchical Bayesian paradigm as a second stage, where , as explained before, the first stage consists of the formulation of  $p(\mathbf{x}|\Omega)$ ,  $p(\mathbf{h}|\Omega)$ , and  $p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \Omega)$ .

A large part of the Bayesian literature is devoted to finding hyperprior distributions  $p(\Omega)$  for which  $p(\Omega, \mathbf{x}, \mathbf{h}|\mathbf{y})$  can be calculated in a straightforward way or be approximated. These are the so called conjugate priors [23], which were developed extensively in Raiffa and Schlaifer [194].

Besides providing for easy calculation or approximations of  $p(\Omega, \mathbf{x}, \mathbf{h}|\mathbf{y})$ , conjugate priors have, as we will see later, the intuitive feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with the parameters updated by the sample information.

The *a priori* models for the hyperparameters depend on the type of the unknown parameters, and different models are proposed in the literature. For parameters corresponding to inverses of variances the gamma distribution is used, given by

$$(2.26) \quad p(\omega) = \Gamma(\omega|a_\omega^o, b_\omega^o) = \frac{(b_\omega^o)^{a_\omega^o}}{\Gamma(a_\omega^o)} \omega^{a_\omega^o-1} \exp[-b_\omega^o \omega],$$

where  $\omega > 0$  denotes a hyperparameter,  $b_\omega^o > 0$  is the scale parameter, and  $a_\omega^o > 0$  is the shape parameter. These parameters are assumed known. The gamma distribution has the following mean, variance and mode:

$$(2.27) \quad E[\omega] = \frac{a_\omega^o}{b_\omega^o}, \quad Var[\omega] = \frac{a_\omega^o}{(b_\omega^o)^2}, \quad Mode[\omega] = \frac{a_\omega^o - 1}{b_\omega^o}.$$

Note that the mode does not exist when  $a_\omega^o \leq 1$  and that mean and mode do not coincide.

There are several important reasons for selecting Gamma distributions for the hyperpriors. First, gamma distribution is conjugate for the variance of the Gaussian, and therefore the posteriors will also have Gamma distributions in the Bayesian formulation. Second, as will be shown later, their update equations will exhibit interesting similarities to some previously derived results in the literature.

For components of mean vectors the corresponding conjugate prior is a normal distribution. Additionally, for covariance matrices the hyperprior is given by an inverse Wishart distribution (see [90]).

We observe, however, that in general most of the methods proposed in the literature use the *uninformative* prior model

$$(2.28) \quad p(\Omega) = \text{constant}.$$

### 2.3. Bayesian Inference Methods

There are a number of different ways to estimate the image and blur using (2.8). Many methods in the literature attempt to provide point estimates to the parameters  $\mathbf{x}$  and  $\mathbf{h}$ , which reduces the problem to the one of optimization. However, different methodologies estimate the distributions of these parameters [90, 175, 117], which have some advantages over other methods. The distributions can either be approximated or simulated. In this section we will present different inference methods.

### 2.3.1. Maximum a Posteriori and Maximum Likelihood

The *Maximum A Posteriori* (MAP) solution is obtained by maximizing the posterior probability density, that is,

$$(2.29) \quad \{\hat{\mathbf{x}}, \hat{\mathbf{h}}, \hat{\Omega}\}_{\text{MAP}} = \underset{\mathbf{x}, \mathbf{h}, \Omega}{\operatorname{argmax}} p(\mathbf{y} | \mathbf{x}, \mathbf{h}, \Omega) p(\mathbf{x} | \Omega) p(\mathbf{h} | \Omega) p(\Omega).$$

On the other hand, the *Maximum Likelihood* (ML) solution attempts to maximize the likelihood  $p(\mathbf{y} | \mathbf{x}, \mathbf{h}, \Omega)$  with respect to the parameters:

$$(2.30) \quad \{\hat{\mathbf{x}}, \hat{\mathbf{h}}, \hat{\Omega}\}_{\text{ML}} = \underset{\mathbf{x}, \mathbf{h}, \Omega}{\operatorname{argmax}} p(\mathbf{y} | \mathbf{x}, \mathbf{h}, \Omega).$$

Note that in this case the parameters present only in  $p(\mathbf{x}, \mathbf{h} | \Omega)$  cannot be estimated. The ML method is widely used, and can also be seen as a non-Bayesian method. It should be noted that ML solution is identical to the MAP solution when uninformative (flat) priors for  $\mathbf{x}$  and  $\mathbf{h}$  are used in (2.29). Some approaches utilize flat priors for some parameters but not for others. Assuming known values for the hyperparameters is equivalent to forming degenerate distributions (impulse functions) for the hyperpriors. For instance, by assuming

$$(2.31) \quad p(\Omega) = \delta(\Omega, \Omega_0) = \begin{cases} 1, & \text{if } \Omega = \Omega_0 \\ 0, & \text{otherwise} \end{cases}$$

the MAP and ML solutions become respectively

$$(2.32) \quad \{\hat{\mathbf{x}}, \hat{\mathbf{h}}\}_{\text{MAP}} = \underset{\mathbf{x}, \mathbf{h}}{\operatorname{argmax}} p(\mathbf{y} | \mathbf{x}, \mathbf{h}, \Omega_0) p(\mathbf{x} | \Omega_0) p(\mathbf{h} | \Omega_0)$$

$$(2.33) \quad \{\hat{\mathbf{x}}, \hat{\mathbf{h}}\}_{\text{ML}} = \underset{\mathbf{x}, \mathbf{h}}{\operatorname{argmax}} p(\mathbf{y} | \mathbf{x}, \mathbf{h}, \Omega_0).$$

Many deconvolution methods can fit into these formulations by using different forms of likelihood functions, image, blur and hyperparameter priors, and the optimization methods. It should be noted that many regularization-based approaches can also be derived using this formulation. In regularization approaches, the inverse problem is formulated as a constrained optimization problem, where the cost function is  $\|\mathbf{y} - \mathbf{Hx}\|_W^2$ . This term ensures fidelity to data. Additionally, constraints on the solutions are imposed by using regularization terms. Generally, these constraints ensure smoothness of the image and the blur, that is, the high frequency energy of the image and the blur is minimized. The effect of the regularization terms is controlled by the regularization parameters, which basically represent the trade off between fidelity to the data and desirable properties (smoothness) of the solutions.

For example, the classical regularized image restoration formulation used in [122, 124, 135] can be derived using (2.32) and (2.33). This formulation is extended to blind deconvolution problem in [250], which can be given in relaxed minimization form as follows:

$$(2.34) \quad \hat{\mathbf{x}}, \hat{\mathbf{h}} = \underset{\mathbf{f}, \mathbf{h}}{\operatorname{argmin}} [\|\mathbf{y} - \mathbf{Hx}\|_W^2 + \lambda_1 \|\mathbf{L}_x \mathbf{x}\|^2 + \lambda_2 \|\mathbf{L}_h \mathbf{h}\|^2],$$

where  $\lambda_1$  and  $\lambda_2$  are the Lagrange multipliers for each constraint, and  $\mathbf{L}_x$  and  $\mathbf{L}_h$  are the regularization operators. The regularization operators are chosen to be Laplacians multiplied by a spatially-varying weights term, calculated as in [122, 124, 75, 125] from the local image variance in order to provide some spatial adaptivity to avoid oversmoothing edges.

**2.3.1.1. Iterated Conditional Modes.** A major problem in the solution of (2.32) is the simultaneous estimation of the variables  $\mathbf{x}$  and  $\mathbf{h}$ . A widely used approach is *Alternating Minimization*

(AM), which basically is minimization with respect to one unknown while holding the others constant. The main advantage of this algorithm is its simplicity due to the linearization of the objective function (see (2.34)). This optimization procedure corresponds to the Iterated Conditional Modes (ICM) proposed by Besag [25]. AM has been applied to standard regularization approaches [250, 63], and to the anisotropic diffusion and TV type models.

There are various numerical methods to solve the associated PDEs resulting from AM. These include the classical Euler, Newton or Runge-Kutta methods; or recently developed approaches, such as time-marching [203], primal-dual methods [55], lagged diffusivity fixed point schemes [238], and half-quadratic regularization [50] (similar to the discrete schemes in [92, 91]). All of these methods employ techniques to discretize and linearize the PDEs to approximate the solution.

### 2.3.2. Minimum Mean-Squared Error

The MAP estimate does not take into account the whole posterior PDF. If the posterior is sharply peaked about the maximum MAP estimate gives the best possible solution. However, there are cases where the posterior can be broad (heavy-tailed) or even multimodal. As mentioned in [168], for a Gaussian in high dimensions most of the probability *mass* is concentrated away from the probability *density* peak.

The Minimum Mean-Squared Error (MMSE) estimate instead minimizes the expected mean square error between the estimates and the true values, and therefore calculates the mean value of  $p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})$ . In practice finding MMSE estimates analytically is generally difficult, though it is possible with sampling based methods (§2.3.5) and can be approximated using variational Bayesian methods (§2.3.4).

### 2.3.3. Marginalizing Hidden Variables

Another method of approaching the problem is to marginalize out some unknowns and perform inference on the others. For instance, in the Evidence analysis [168, 166], where we first calculate

$$(2.35) \quad \hat{\mathbf{h}}, \hat{\Omega} = \operatorname{argmax}_{\mathbf{h}, \Omega} \int_{\mathbf{x}} p(\Omega) p(\mathbf{x}, \mathbf{h} | \Omega) p(\mathbf{y} | \Omega, \mathbf{x}, \mathbf{h}) d\mathbf{x}$$

and then finding the restoration as

$$(2.36) \quad \hat{\mathbf{x}}|_{\hat{\mathbf{h}}, \hat{\Omega}} = \operatorname{argmax}_{\mathbf{x}} p(\mathbf{x} | \hat{\Omega}) p(\mathbf{y} | \hat{\Omega}, \mathbf{x}, \hat{\mathbf{h}}).$$

Another way is to marginalize  $\mathbf{h}$  and  $\Omega$  to directly obtain

$$(2.37) \quad \hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} \int_{\mathbf{h}, \Omega} p(\Omega) p(\mathbf{x}, \mathbf{h} | \Omega) p(\mathbf{y} | \Omega, \mathbf{x}, \mathbf{h}) d\mathbf{h} \cdot d\Omega,$$

which is called the Empirical analysis [166]. Note that the marginalized variables are also called hidden variables.

The expectation-maximization (EM) algorithm, first described in [69], is a very popular techniques in signal processing for iteratively solving ML and MAP problems. Its convergence to a *local* maximum of the likelihood or the posterior distribution is guaranteed. It is also particularly well-suited to providing solutions to inverse problems in image restoration, blind deconvolution, and super resolution, since the parameters to be estimated can be regarded as *hidden data*.

### 2.3.4. Variational Bayesian Approach

Variational methods are generalizations of the EM algorithm to find ML and MAP estimates. Although the EM algorithm is very useful in a wide range of application, its application is not possible in many problems since the posterior distribution cannot be calculated. The key to the variational methods is to approximate the posterior distribution  $p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})$  by a simpler distribution  $q(\mathbf{x}, \mathbf{h}, \Omega)$ . This approximation is obtained by minimizing the Kullback-Leibler (KL) divergence between these two distributions. The estimates to the unknowns can be computed based on this approximation to the posterior distribution. Additionally, studying the variational approximation  $q(\mathbf{x}, \mathbf{h}, \Omega)$  allows the examination of the quality of the estimates, such as measuring their uncertainty.

In BD the variational approximation provides a tractable distribution  $q(\mathbf{x}, \mathbf{h}, \Omega)$  to the intractable distribution  $p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})$ . For an arbitrary PDF  $q(\mathbf{x}, \mathbf{h}, \Omega)$ , the goal is to minimize the KL divergence, given by

$$\begin{aligned} KL(q(\mathbf{x}, \mathbf{h}, \Omega) \| p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})) &= \int q(\mathbf{x}, \mathbf{h}, \Omega) \log \left( \frac{q(\mathbf{x}, \mathbf{h}, \Omega)}{p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})} \right) d\mathbf{x} \cdot d\mathbf{h} \cdot d\Omega \\ &= \int q(\mathbf{x}, \mathbf{h}, \Omega) \log \left( \frac{q(\mathbf{x}, \mathbf{h}, \Omega)}{p(\mathbf{x}, \mathbf{h}, \Omega, \mathbf{y})} \right) d\mathbf{f} \cdot d\mathbf{h} \cdot d\Omega + \text{const}, \end{aligned}$$

(2.38)

which is always nonnegative and equal to zero only when  $q(\mathbf{x}, \mathbf{h}, \Omega) = p(\mathbf{x}, \mathbf{h}, \Omega | \mathbf{y})$ , which corresponds to the EM result.

To reduce the computational complexity and to find analytical forms for the parameter distributions, the mean field approximation is used to factorize the PDF  $q(\mathbf{x}, \mathbf{h}, \Omega)$  such that

$$(2.39) \quad q(\mathbf{x}, \mathbf{h}, \Omega) = q(\mathbf{x})q(\mathbf{h})q(\Omega).$$

Based on this approximation, an iterative procedure to estimate the distributions of each parameter can be developed, which is a form of AM. Note that some parameters can be assumed to have degenerate distributions, so that point estimates for these parameters are calculated. More details on the variational methods will be given later when we present our algorithms.

### 2.3.5. Sampling Methods

The most general approach to perform inference is to simulate the posterior distributions numerically. This allows to perform inference on arbitrary models in high dimensional spaces where analytic solutions cannot be found. Methods such as Markov Chain Monte Carlo (MCMC) (see, e.g., [175, 7, 202]) attempt to approximate the posterior distribution by the statistics of samples generated from a Markov Chain.

The simplest example of MCMC is the Gibbs sampler which has been used in classical image restoration in conjunction with MRF image models [93]. Gibbs sampler draws samples from conditional distributions of each parameter given the others, and the conditioning is on the most recently sample values of the parameters. Note that this is similar to the variational Bayesian approach where instead of drawing samples expectations of the distributions are used. Similarly, with use of *Simulated Annealing* [93], the ICM formulation can be considered a deterministic approximation of the sampling, where the conditional distributions are replaced by degenerate distributions at their modes (termed “instantaneous freezing” in [25]).

In theory, sampling methods can be shown to be optimal in performing inference on the posterior, and their convergence is guaranteed. However, in practice, they are computationally very intensive, and it is hard to tell when the convergence has occurred. Despite these, sampling methods are utilized in a wide range of applications. For instance, they are used in [167] where both direct inference and variational methods are difficult to perform because of the form of the blur prior.

## 2.4. Super Resolution

Super resolution (SR) describes the process of obtaining an high resolution (HR) image or a sequence of HR images from a set of low resolution (LR) observations. Super resolution is also referred to as resolution enhancement (RE) in the literature. The scope of this proposal only includes motion-based spatial and temporal RE, although there are other forms too (see, for example, motion-free RE [61] and hyperspectral image enhancement [170, 173]).

In SR, the LR observations are under-sampled, blurred and warped versions of the HR image(s). They are generally acquired by multiple sensors imaging the same scene, or by a single sensor acquiring images over a period of time. For static scenes the LR images only exhibit global displacements, whereas for dynamic scenes they can also have local displacements due to object motion.

A simple solution to obtain higher-resolution observations is to increase the resolution of the imaging device. However, this may not be feasible due to the increased cost, the increased data transfer rate, and also because increasing the number of pixels in a sensor increases the shot noise [188]. Signal processing techniques offer significant advantages to improve the resolution where obtaining HR observations are not feasible.

### 2.4.1. Bayesian Modeling

SR problem can also be cast within the Bayesian framework. Let  $\mathbf{x}_k$  and  $\mathbf{y}_k$  denote the  $k^{th}$  HR and LR frame, respectively. We assume that the LR frames are of size  $N \times M$ , and the HR frames are of size  $PN \times PM$ , so that the magnification factor  $P$  both horizontally and vertically. The set of frames can be ordered to form new vectors  $\mathbf{x}$ , and  $\mathbf{y}$ . We denote by the vector  $\mathbf{d}_{l,k}$  the warping model parameters in mapping frame  $\mathbf{x}_k$  to  $\mathbf{x}_l$ . The vector  $\mathbf{d}_{l,k}$  contains the motion information for compensating frame  $\mathbf{x}_k$  to  $\mathbf{x}_l$ , that is, except the occluded areas, each pixel in  $\mathbf{x}_l$  can be predicted from  $\mathbf{x}_k$  using  $\mathbf{d}_{l,k}$ . The motion information is subpixel accuracy. The set of vectors  $\mathbf{d}_{l,k}$  can be ordered similarly to form the vector  $\mathbf{d}$ . In the case of still image SR there is only one HR image  $\mathbf{x}$ , but the same notation can be applied. The SR problem is to find estimates to  $\mathbf{x}$  and  $\mathbf{d}$  based on the observations  $\mathbf{y}$ , which will be generically denoted by  $\mathbf{o}$ .

By treating all parameters and observable variables as unknown stochastic quantities, we again form prior distributions, likelihood functions and finally the joint probability distribution as follows:

$$(2.40) \quad p(\Omega, \mathbf{x}_k, \mathbf{d}, \mathbf{o}) = p(\Omega)p(\mathbf{x}_k, \mathbf{d} | \Omega)p(\mathbf{o} | \Omega, \mathbf{x}_k, \mathbf{d}),$$

where the unknown variables are  $\mathbf{x}_k$ ,  $\mathbf{d}$  and the hyperparameters  $\Omega$ , the observation is treated as a sample of a random field. Note that the hyperparameters may also be assumed known. Also,  $p(\mathbf{o} | \mathbf{x}_k, \mathbf{d})$  is termed the *likelihood* of the observations. The Bayesian inference is performed on the *posterior*

$$(2.41) \quad p(\Omega, \mathbf{x}_k, \mathbf{d} | \mathbf{o}) = \frac{p(\Omega)p(\mathbf{x}_k, \mathbf{d} | \Omega)p(\mathbf{o} | \Omega, \mathbf{x}_k, \mathbf{d})}{p(\mathbf{o})}.$$

As in image restoration and blind deconvolution, a MAP solution can be found using (2.41). However, as before, a major problem is to find simultaneous estimates of the variables  $\mathbf{x}_k$  and  $\mathbf{d}$ . The SR literature concentrated mainly on two AM methodologies which provide approximations to the MAP solution. The first one, referred to as *alternate*, uses a cyclic coordinate descent procedure [102], where the HR images and the displacements are found at different steps by holding the other one constant. The second one, referred to as *sequential*, assumes the displacements are known or estimated separately, and provides estimates to the HR images using these displacements in (2.41). Observe that assuming that a parameter is known is equivalent to using a degenerate distribution for it. A degenerate distribution on a variable  $\omega$  is defined as:

$$(2.42) \quad p(\omega) = \delta(\omega, \omega_0) = \begin{cases} 1, & \text{if } \omega = \omega_0 \\ 0, & \text{otherwise} \end{cases}$$

#### 2.4.2. Low-Resolution Image Formation Models

First we define the relation between two HR frames. Assuming constant illumination conditions in the scene, we can write this relation as follows:

$$(2.43) \quad \mathbf{x}_l = \mathbf{C}(\mathbf{d}_{l,k})\mathbf{x}_k,$$

where  $\mathbf{C}(\mathbf{d}_{l,k})\mathbf{x}_k$  is a  $(PN \times PM) \times (PN \times PM)$  matrix that maps frame  $\mathbf{x}_l$  to frame  $\mathbf{x}_k$ .

The next important step is to establish the model of the acquisition of the LR frames. There are two main models, namely the warp-blur model and the blur-warp model. In the warp-blur model, the warping of the image is applied before it is blurred. Then, the LR image  $\mathbf{y}_l$  is related to the HR image  $\mathbf{x}_l$  by

$$(2.44) \quad \mathbf{y}_l = \mathbf{A}_l \mathbf{H}_l \mathbf{x}_l + \eta_l, \quad l = 1, 2, \dots, L,$$

where the matrix  $\mathbf{H}_l$  of size  $(PN \times PM) \times (PN \times PM)$  describes the filtering of the HR image,  $\mathbf{A}_l$  is the downsampling matrix of size  $NM \times (PN \times PM)$ , and  $\eta_l$  denotes the observation noise. The matrices  $\mathbf{A}_l$  and  $\mathbf{H}_l$  are generally assumed to be known (see, however, [7]). Combining (2.43) and (2.44) we obtain the following equation describing the acquisition of an LR image  $\mathbf{y}_l$  from the unknown HR image  $\mathbf{x}_k$

$$(2.45) \quad \mathbf{y}_l = \mathbf{A}_l \mathbf{H}_l \mathbf{C}(\mathbf{d}_{l,k}) \mathbf{x}_k + \eta_l + \mu_{l,k} = \mathbf{A}_l \mathbf{H}_l \mathbf{C}(\mathbf{d}_{l,k}) \mathbf{x}_k + \mathbf{e}_{l,k},$$

where  $\mu_{l,k}$  represents the registration noise and  $\mathbf{e}_{l,k}$  represents the combined acquisition and registration noise. Note from (2.45) that the warp  $\mathbf{C}(\mathbf{d}_{l,k})$  is applied before the blur  $\mathbf{H}_{l,k}$  on  $\mathbf{x}_k$ .

In the blur-warp model, the HR image is first blurred and then warping and downsampling is applied. In this case, the observation model becomes

$$(2.46) \quad \mathbf{y}_l = \mathbf{A}_l \mathbf{M}(\mathbf{m}_{l,k}) \mathbf{B}_l \mathbf{x}_k + \mathbf{w}_{l,k},$$

where  $\mathbf{w}_{l,k}$  denotes the acquisition and registration noise,  $\mathbf{B}_l$  is the blurring matrix,  $\mathbf{M}(\mathbf{m}_{l,k})$  is the motion compensation operator with the use of the motion vector  $\mathbf{m}_{l,k}$ , and  $\mathbf{A}_l$  is again the downsampling matrix.

#### 2.4.3. High-Resolution Image Models in Super Resolution

We proceed by giving examples of image prior models used in super resolution. Some methods use a noninformative flat prior for the HR images  $\mathbf{x}_k$  [219, 222, 111], so that

$$(2.47) \quad p(\mathbf{x}_k) \propto \text{constant}$$

Another prior model is the SAR model mentioned before in Section 2.2.3.1, which is given in super resolution context as

$$(2.48) \quad p(\mathbf{x}_k) \propto \exp\left[-\frac{\lambda}{2} \|\mathbf{C}\mathbf{x}_k\|^2\right]$$

Two other prior models are based on the Huber functional and were proposed by Schultz and Stevenson [206] and Hardie *et al.* [106] (see Park *et al.* [188] for additional references on prior models). More recently, total variation [46], anisotropic diffusion [130], and compound models [195] have all been applied to the SR problem.

Bilateral total variation prior model is proposed in [79], which is given by

$$(2.49) \quad p(\mathbf{x}_k) \propto \exp \left[ -\lambda \sum_{l=-P}^P \sum_{\substack{m=0, \\ l+m \geq 0}}^P \alpha^{|l|+|m|} \| \mathbf{x}_k - S_x^l S_y^m \mathbf{x}_k \|_1 \right],$$

where  $\| \theta \|_1 = \sum |\theta_i|$  denotes the  $l1$ -norm,  $\lambda$  is a scale parameter, the operators  $S_x^l$  and  $S_y^m$  are shift operators ( $l$  and  $m$  pixels in the horizontal and vertical directions, respectively) and  $0 < \alpha < 1$  is applied to provide a spatially decaying effect to the summation of the  $l1$ -norm terms.

Chan *et al.* [51] propose the use of biorthogonal wavelets, where the achr image distribution is given by

$$(2.50) \quad p(\mathbf{x}_k) \propto \exp \left[ -\frac{\lambda}{2} \| \mathbf{x}_k \|^2 \right].$$

On the other hand, Willet *et al.* [244] formulate the LR observation model in terms of the discrete wavelet transform (DWT) coefficients  $\theta$  of the HR image  $\mathbf{x}_k$  [153] and use

$$(2.51) \quad p(\theta) \propto \exp[-\tau \| \theta \|_1]$$

as the image prior model, where  $\tau$  is a scale parameter.

The final class of SR methods utilize learning approaches for the modeling of the HR images. Principal component analysis (PCA) is used in [47] to learn the prior distributions from a dataset of HR images. In this case, the HR image distribution can be given as

$$(2.52) \quad p(\mathbf{x}_k) \propto \exp \left[ -\frac{\lambda}{2} \| (I - VV^t)(\mathbf{x}_k - \mu) \|^2 \right],$$

where  $\Sigma$  is the diagonal matrix of component variances obtained from the PCA,  $\mu$  is the average of the training images, and  $\lambda$  is a scale parameter.. The PCA methodology is also used in [102]. Learning approaches with the use of sampling exemplars are also used in [18, 44, 85, 190, 240].

#### 2.4.4. Bayesian Inference Models in Super Resolution

As shown before, the joint probability distribution can be written as in (2.40). The work in SR literature primarily dealt with two approximations, namely, *alternate* and *sequential* to provide approximations to the solution

$$(2.53) \quad \hat{\mathbf{x}}_k, \hat{\mathbf{d}}, \hat{\Omega} = \underset{\mathbf{x}_k, \mathbf{d}, \Omega}{\operatorname{argmax}} p(\mathbf{x}_k, \mathbf{d}, \Omega | \mathbf{o}) = \underset{\mathbf{x}_k, \mathbf{d}, \Omega}{\operatorname{argmax}} p(\Omega) p(\mathbf{x}_k | \Omega) p(\mathbf{d} | \Omega) p(\mathbf{o} | \Omega, \mathbf{x}_k, \mathbf{d})$$

Most models assume that the image intensities and the motion vectors are independent given  $\Omega$ , i.e.,  $p(\mathbf{d}, \mathbf{x}_k | \Omega) = p(\mathbf{d} | \Omega) p(\mathbf{x}_k | \Omega)$ . Utilizing this assumption, the posterior on which the inference will be performed can be written as

$$(2.54) \quad p(\mathbf{x}_k, \mathbf{d}, \Omega | \mathbf{o}) = \frac{p(\Omega) p(\mathbf{x}_k | \Omega) p(\mathbf{d} | \Omega) p(\mathbf{o} | \Omega, \mathbf{x}_k, \mathbf{d})}{p(\mathbf{o})}.$$

The term  $p(\mathbf{o}|\Omega, \mathbf{x}_k, \mathbf{d})$  is called the *likelihood* of the observations and it is generally assumed to be a Gaussian distribution which corresponds to the Gaussian noise assumption in the LR image formation process. We studied different image priors in Section 2.4.3. Additionally, there have been a number of prior models for the hyperparameters  $\Omega$  proposed in the literature. They can be assumed known, which corresponds to the degenerate distribution

$$(2.55) \quad p(\Omega) = \delta(\Omega, \Omega_0) = \begin{cases} 1, & \text{if } \Omega = \Omega_0 \\ 0, & \text{otherwise} \end{cases}.$$

When  $\Omega$  is unknown most of the methods proposed in the literature use the model

$$(2.56) \quad p(\Omega) = \text{constant}.$$

However, as in image restoration and blind deconvolution, conjugate priors can be used for hyperpriors, see Section 2.2.4 for details. To our knowledge, these hyperprior models have only been used by Humblot and Ali Mohammad-Djafari [110] in SR problems.

There are a number of methods in estimating the unknown variables using (2.54). Considering the similarity between (2.54) and (2.8), most of these methods are identical to the methods used in image restoration and super resolution. We will therefore will not give explicit details of each method but state the differences when necessary.

Simultaneous estimation of the HR image and the motion parameters using MAP is utilized in [105, 159, 211]. These methods are examples of AM methods and therefore are ICM procedures. Tipping and Bishop [226] pointed out that marginalization, and therefore *evidence*

and *empirical* based analysis can be used in HR image estimation by integrating out one of the unknowns.

Tom and Katsaggelos [229, 230, 231] have used the EM algorithm in SR image reconstruction assuming an AR prior HR image model with unknown parameters and unknown global translations between HR images. The prior model used on such translation was an uniform distribution (see also [249]).

Evidence analysis is applied in [172] to first estimate the variances of the HR image prior distribution and the image formation model and then the HR image. The same formulation is also achieved from a regularization point of view in [36, 180, 181, 182, 183]. Other methods utilizing evidence-based analysis are [226, 110].

Variational Bayesian approaches utilizing distribution approximations are used in SR problems in [220]. Finally, Gibbs sampler has been used in [110] within a MAP formulation.

## CHAPTER 3

# Total Variation Image Restoration Using Variational Distribution Approximation

### 3.1. Introduction

In this chapter we develop a novel image restoration method based on total variation image priors using a Bayesian formulation. We will utilize the standard formulation of the image degradation model, introduced in Chapter 2, given in matrix-vector form by

$$(3.1) \quad \mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n},$$

where the  $N \times 1$  vectors  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{n}$  again represent respectively the original image, the available noisy and blurred image, and the noise with independent elements of variance  $\sigma_n^2 = \beta^{-1}$ , and  $\mathbf{H}$  represents the known blurring matrix. The images are assumed to be of size  $m \times n = N$ , and they are lexicographically ordered into  $N \times 1$  vectors. The restoration problem calls for finding an estimate of  $\mathbf{x}$  given  $\mathbf{y}$ ,  $\mathbf{H}$ , and knowledge about  $\mathbf{n}$  and possibly  $\mathbf{x}$  [120].

A number of approaches have been developed in providing solutions to the restoration problem (see, for example, [120], [19], [54], and references therein). A straightforward approach to the restoration problem is to use least squares estimation and select  $\bar{\mathbf{x}}$ , an estimate of the original

---

<sup>0</sup>This work has appeared in [15, 13]

image, as

$$(3.2) \quad \bar{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} \frac{1}{Z_{noise}(\beta)} \exp \left[ -\frac{1}{2} \beta \| \mathbf{y} - \mathbf{Hx} \|^2 \right],$$

where  $Z_{noise}(\beta) = (2\pi/\beta)^{N/2}$ . However, as is well known, this approach does not lead to useful restorations in most cases. Use of prior knowledge about the original image can improve the restoration results. Within the Bayesian framework this knowledge is encapsulated as a prior distribution  $p(\mathbf{x})$ .

A general model for the prior distribution  $p(\mathbf{x})$  is a Markov Random Field (MRF) which is characterized by its Gibbs distribution given by

$$(3.3) \quad p(\mathbf{x}|\alpha) = \frac{1}{Z(\alpha)} \exp \{-\alpha F(\mathbf{x})\},$$

where  $Z(\alpha)$  is the partition function with a constant  $\alpha$  and  $F$  is the energy function of the form  $F(\mathbf{x}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{x})$ , where  $\mathcal{C}$  denotes a set of cliques (i.e., set of connected pixels) for the MRF, and  $V_c$  is a potential function defined on a clique.

A critical issue is the choice of the energy function. We use the Total Variation (TV) image prior [204] whose energy function is the discrete version of the total variation integral defined as

$$(3.4) \quad F_{TV}(\mathbf{x}) = \int |\nabla(\mathbf{x})| d\mathbf{x} .$$

We will explicitly write the form of the prior model in the next section.

If the hyperparameters  $\alpha$  and  $\beta$  are known, following the Bayesian paradigm (see [104] for the unification of probabilistic and variational estimation) it is customary to select, as the

restoration of  $\mathbf{x}$ , the image  $\mathbf{x}_{(\alpha,\beta)}$  defined by

$$(3.5) \quad \mathbf{x}_{(\alpha,\beta)} = \operatorname{argmax}_{\mathbf{x}} p(\mathbf{x}|\alpha)p(\mathbf{y}|\mathbf{x},\beta) = \operatorname{argmin}_{\mathbf{x}} [\alpha F_{\text{TV}}(\mathbf{x}) + \frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2].$$

Not much work has been reported in the literature on the joint parameter and image estimation when the parameters  $\alpha$  and  $\beta$  are not known (see [54, 53] for recent developments in variational modeling and inference). Rudin *et al.* [204] consider the minimization of  $F_{\text{TV}}(\mathbf{x})$  constrained by  $\|\mathbf{y} - \mathbf{Hx}\|^2 = N\hat{\sigma}_{\mathbf{n}}^2$ , where  $\hat{\sigma}_{\mathbf{n}}^2$  represents an estimate of the noise variance, and then proceed to estimate both the image and the associated Lagrange multiplier to this constrained optimization problem. Bertalmio *et al.* [24] make the Lagrange multiplier region dependent. Bioucas-Dias *et al.* [26], using their majorization-minimization approach [27], propose a Bayesian method to estimate the original image and  $\alpha$  assuming that an estimate of the noise variance is available. To our knowledge no work has been reported on the simultaneous estimation of the parameters  $\alpha$  and  $\beta$  and the image and also on the estimation of the uncertainty of those estimates (only point estimates of the parameters and image have been provided).

In this chapter we use the Bayesian paradigm to jointly estimate the image and unknown hyperparameters ( $\alpha$  and  $\beta$ ) in image restoration when the TV image prior is used. The estimation procedure will not provide only point estimates of the image and the hyperparameters but also the probability distributions that approximate the posterior distribution of the hyperparameters and the original image given the observation.

The chapter is organized as follows: Section 3.2 presents a general description of the Bayesian modeling and inference of the TV restoration problem, which includes a brief discussion on estimation procedures (inference methods) that provide point or probability distribution estimates. The actual parameter hyperpriors, image prior, and observation models used

are then presented in Section 3.3. Section 3.4 describes the variational approach to distribution approximation for TV image restoration and how inference is performed. We propose different approximations of the posterior distribution of the image and the unknown hyperparameters, and compare them to other approaches reported in the literature. Finally, in Section 3.5 experimental results and comparisons with other methods are shown, and Section 3.6 concludes the chapter.

### 3.2. Bayesian Modeling and Inference

The Bayesian modeling of the TV restoration problem requires first the definition of a joint distribution  $p(\alpha, \beta, \mathbf{x}, \mathbf{y})$  of the observation,  $\mathbf{y}$ , the unknown image,  $\mathbf{x}$ , and the hyperparameters  $\alpha$  and  $\beta$ . To model the joint distribution, we utilize the hierarchical Bayesian paradigm (see, for example, [168, 160, 87, 169]). In the hierarchical approach to image restoration we have at least two stages. In the first stage, knowledge about the structural form of the observation noise and the structural behavior of the image is used in forming  $p(\mathbf{y}|\mathbf{x}, \beta)$  and  $p(\mathbf{x}|\alpha)$ , respectively. These noise and image models depend on the unknown hyperparameters  $\alpha$  and  $\beta$ . In the second stage a hyperprior on the hyperparameters is defined, thus allowing for the incorporation of information about these hyperparameters into the process.

For  $\alpha, \beta, \mathbf{x}, \mathbf{y}$  the following joint distribution is defined

$$(3.6) \quad p(\alpha, \beta, \mathbf{x}, \mathbf{y}) = p(\alpha)p(\beta)p(\mathbf{x}|\alpha)p(\mathbf{y}|\mathbf{x}, \beta),$$

and inference is based on  $p(\alpha, \beta, \mathbf{x}|\mathbf{y})$ .

Three crucial questions have to be addressed when modeling and performing inference for image restoration problems using the hierarchical Bayesian paradigm. The first one relates to

the definition of  $p(\alpha)$  and  $p(\beta)$ . We should be able to deal with the case of known hyperparameters which correspond to degenerate distributions for  $p(\alpha)$  and  $p(\beta)$ , but also with more realistic situations including the cases when some knowledge about these parameters is available or when only the observation  $\mathbf{y}$  is available to estimate them.

The second crucial problem to be considered is to decide how inference will be carried out. A commonly used approach in image restoration (called the Evidence analysis [168]) consists of estimating the hyperparameters  $\alpha, \beta$  by using

$$(3.7) \quad (\hat{\alpha}, \hat{\beta}) = \underset{\alpha, \beta}{\operatorname{argmax}} p(\alpha, \beta | \mathbf{y}) = \underset{\alpha, \beta}{\operatorname{argmax}} \int p(\alpha, \beta, \mathbf{x} | \mathbf{y}) d\mathbf{x} = \underset{\alpha, \beta}{\operatorname{argmax}} \int p(\alpha, \beta, \mathbf{x}, \mathbf{y}) d\mathbf{x}$$

and then estimating the image by solving

$$(3.8) \quad \hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} p(\mathbf{x} | \hat{\alpha}, \hat{\beta}, \mathbf{y}).$$

Another approach, also commonly used in image restoration, is the so called empirical analysis [166], which consists of calculating the restoration by solving

$$(3.9) \quad \bar{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \int \int p(\alpha, \beta, \mathbf{x}, \mathbf{y}) d\alpha d\beta.$$

These inference procedures aim at optimizing a given function and not at obtaining posterior distributions that can be analyzed or simulated to obtain additional information about the quality of the estimates. Instead of having a distribution over all possible values of the parameters and the image, the above inference procedures choose a specific set of values. This means that we have neglected many other interpretations of the data. If the posterior is sharply peaked, other values of the hyperparameters and the image will have a much lower posterior probability but,

if the posterior is broad, choosing a unique value will neglect many other choices of them with similar posterior probabilities.

The third crucial problem to be solved when using the Bayesian paradigm on TV image restoration is to decide how to calculate  $p(\alpha, \beta, \mathbf{x}|\mathbf{y})$ , which is in general a challenging task. An approach is provided by the variational distribution approximation. This approximation can be thought of as being between the Laplace approximation (see, for instance, [86, 87]) and sampling methods [7]. The basic underlying idea, as will be explained later, is to approximate  $p(\alpha, \beta, \mathbf{x}|\mathbf{y})$  with a simpler distribution. See the very interesting theses [162], [21], books [215], [28] and book chapter [117] for a comprehensive introduction to variational methods.

The last few years have seen a growing interest in the application of variational methods [117, 162] to inference problems. These methods attempt to approximate posterior distributions with the use of the Kullback-Leibler cross-entropy [133]. Application of variational methods to Bayesian inference problems include graphical models and neural networks [117], independent component analysis [162], mixtures of factor analyzers, linear dynamic systems, hidden Markov models [21], support vector machines [30] and blind deconvolution problems (see [163], [144] and [169]).

In this chapter we propose a method that uses a TV prior distribution for the image, and gamma distributions for the unknown parameter (hyperparameter) of the prior and the image formation noise. We apply variational methods to approximate the posterior probability of the unknown image and hyperparameters and propose two different approximations of the posterior distribution. We use the obtained posterior approximation to gain additional insight into the estimated hyperparameters and image.

### 3.3. Hyperpriors, prior, and observation model used in TV image Deconvolution

We first describe the TV prior model as well as the observation model we use in the first stage of the hierarchical Bayesian paradigm. Then, since the prior and observation models depend on unknown hyperparameters we proceed to explain the hyperprior distributions we utilize for these hyperparameters.

#### 3.3.1. First stage: prior models on images

As image model we use the TV prior, given by

$$(3.10) \quad p(\mathbf{x}|\alpha) \propto \frac{1}{Z_{\text{TV}}(\alpha)} \exp[-\alpha \text{TV}(\mathbf{x})],$$

where  $Z_{\text{TV}}(\alpha)$  is the partition function and

$$(3.11) \quad \text{TV}(\mathbf{x}) = \sum_i \sqrt{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2},$$

where the operators  $\Delta_i^h(\mathbf{x})$  and  $\Delta_i^v(\mathbf{x})$  correspond to, respectively, the horizontal and vertical first order differences, at pixel  $i$ , that is,  $\Delta_i^h(\mathbf{x}) = x_i - x_{l(i)}$  and  $\Delta_i^v(\mathbf{x}) = x_i - x_{a(i)}$ , with  $l(i)$  and  $a(i)$  denoting the nearest neighbors of  $i$ , to the left and above, respectively.

Unless we want to use very simple estimation procedures for the hyperparameter  $\alpha$  we need to calculate (approximate) the partition function  $Z_{\text{TV}}(\alpha)$ . Using

$$(3.12) \quad \int \int \exp \left[ -\alpha \sqrt{u^2 + v^2} \right] du dv = 2\pi/\alpha^2$$

we can utilize the following approximation of  $p(\mathbf{x}|\alpha)$  in (3.10) proposed in [27],

$$(3.13) \quad p(\mathbf{x}|\alpha) = c \alpha^{N/2} \exp[-\alpha TV(\mathbf{x})],$$

where again  $N$  is the size of the original image  $\mathbf{x}$ , and  $c$  is a constant. Note that the idea of approximating partition functions in image priors to be able to estimate distribution parameters has also been used in [144].

The probability distribution corresponding to the observation model in (3.1) is given by

$$(3.14) \quad p(\mathbf{y}|\mathbf{x}, \beta) \propto \beta^{N/2} \exp\left[-\frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2\right]$$

### 3.3.2. Second stage: hyperpriors on the hyperparameters

A large part of the Bayesian literature is devoted to finding hyperprior distributions  $p(\alpha, \beta)$  for which  $p(\alpha, \beta, \mathbf{x}|\mathbf{y})$  can be either calculated in a straightforward way or be closely approximated. These are the so called conjugate priors [23] which have the intuitive feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with the parameters updated by the sample information.

We will assume that each of the hyperparameters  $\omega \in \{\alpha, \beta\}$  has as hyperprior the gamma distribution,  $\Gamma(\omega|a_\omega^o, b_\omega^o)$ , defined by

$$(3.15) \quad p(\omega) = \Gamma(\omega|a_\omega^o, b_\omega^o) = \frac{(b_\omega^o)^{a_\omega^o}}{\Gamma(a_\omega^o)} \omega^{a_\omega^o-1} \exp[-\omega b_\omega^o],$$

where  $b_\omega^o > 0$  and  $a_\omega^o > 0$  are respectively the scale and shape parameters, which are assumed to be known. We will discuss their calculation in Section 3.5. The gamma distribution has the

following mean, variance, and mode

$$(3.16) \quad E[\omega] = a_\omega^o/b_\omega^o, \quad Var[\omega] = a_\omega^o/(b_\omega^o)^2, \quad \text{Mode}[\omega] = (a_\omega^o - 1)/b_\omega^o.$$

There are several important reasons for selecting Gamma distributions for the hyperpriors. First, the Gamma distribution is conjugate for the variance of the Gaussian, and therefore the posteriors will also have Gamma distributions in the Bayesian formulation. Second, as will be shown later, their update equations will exhibit interesting similarities to some previously derived results in the literature.

Finally, combining the first and second stages of the problem modeling we have the following global distribution

$$(3.17) \quad p(\alpha, \beta, \mathbf{x}, \mathbf{y}) = p(\alpha)p(\beta)p(\mathbf{x}|\alpha)p(\mathbf{y}|\mathbf{x}, \beta),$$

where  $p(\alpha)$ ,  $p(\beta)$ ,  $p(\mathbf{x}|\alpha)$ , and  $p(\mathbf{y}|\mathbf{x}, \beta)$  have been defined in (3.15), (3.13), and (3.14). The joint probability model is shown in graphical form in Fig. 3.1 using a directed acyclic graph.

### 3.4. Bayesian Inference and Variational Approximation of the posterior distribution for TV image restoration

The Bayesian paradigm dictates that inference on  $(\alpha, \beta, \mathbf{x})$  should be based on

$$(3.18) \quad p(\alpha, \beta, \mathbf{x} | \mathbf{y}) = \frac{p(\alpha, \beta, \mathbf{x}, \mathbf{y})}{p(\mathbf{y})},$$

where  $p(\alpha, \beta, \mathbf{x}, \mathbf{y})$  is given by (3.17).

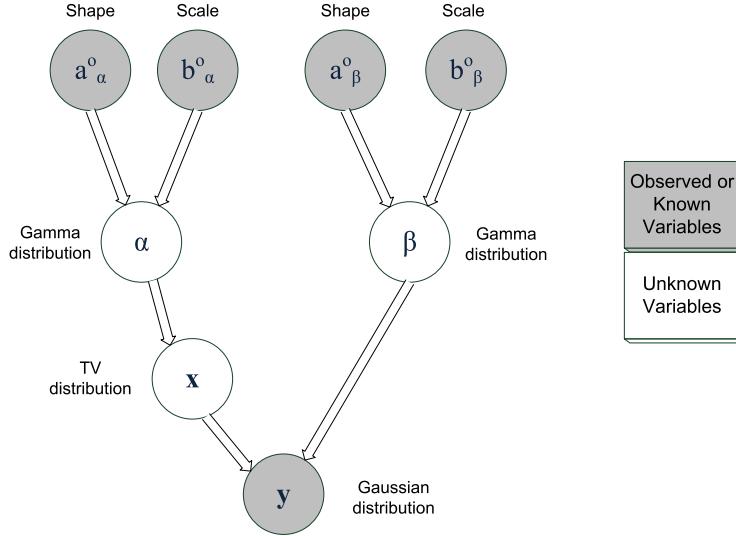


Figure 3.1. Graphical model showing relationships between variables.

Because  $p(\alpha, \beta, \mathbf{x} | \mathbf{y})$  can not be found in closed form, since

$$(3.19) \quad p(\mathbf{y}) = \int \int \int p(\alpha, \beta, \mathbf{x}, \mathbf{y}) d\mathbf{x} d\beta d\alpha$$

can not be calculated analytically, we apply variational methods to approximate this distribution by the distribution  $q(\alpha, \beta, \mathbf{x})$ . We utilize a mean field approximation [187] for the posterior distributions of  $\alpha$ ,  $\beta$ , and  $\mathbf{x}$  so that these posterior distributions are assumed to be independent given the observations. We will later show that particular selections of the distributions  $q(\alpha, \beta)$  and  $q(\mathbf{x})$  lead to the hyperparameters and image point estimates provided by the evidence and empirical analysis described in Section 3.2. Notice, however, that unless the distributions  $q(\alpha, \beta)$  and  $q(\mathbf{x})$  are degenerate, the variational approximation provides us with additional information that goes beyond simple point estimates.

The variational criterion used to find  $q(\alpha, \beta, \mathbf{x})$  is the minimization of the Kullback-Leibler divergence, given by

$$\begin{aligned}
C_{KL}(q(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x} | \mathbf{y})) &= \int \int \int q(\alpha, \beta, \mathbf{x}) \log \left( \frac{q(\alpha, \beta, \mathbf{x})}{p(\alpha, \beta, \mathbf{x} | \mathbf{y})} \right) d\alpha d\beta d\mathbf{x} \\
&= \int \int \int q(\alpha, \beta, \mathbf{x}) \log \left( \frac{q(\alpha, \beta, \mathbf{x})}{p(\alpha, \beta, \mathbf{x}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x} + \text{const}, \\
(3.20) \quad &= \mathcal{M}(q(\mathbf{x}, \alpha, \beta)) + \text{const},
\end{aligned}$$

which is always non-negative and equal to zero only when  $q(\alpha, \beta, \mathbf{x}) = p(\alpha, \beta, \mathbf{x} | \mathbf{y})$ .

Due to the form of the TV prior the above integral is difficult to evaluate (note that also for the same reason the evidence and empirical estimates described in Section 3.2 are difficult to calculate). We can however majorize the TV prior by a function which renders the integral easier to calculate. Let us consider the following inequality, also used in [27], which states that for any  $w \geq 0$  and  $z > 0$

$$(3.21) \quad \sqrt{wz} \leq \frac{w+z}{2} \Rightarrow \sqrt{w} \leq \frac{w+z}{2\sqrt{z}}.$$

Let us also define for  $\alpha, \mathbf{x}$ , and any  $N$ -dimensional vector  $\mathbf{u} \in (R^+)^N$ , with components  $u_i$ ,  $i = 1, \dots, N$ , the following functional

$$(3.22) \quad \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u}) = \alpha^{N/2} \exp \left[ -\frac{\alpha}{2} \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + u_i}{\sqrt{u_i}} \right].$$

Now, using inequality (3.21) with  $w = (\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2$  and  $z = u_i$  and comparing (3.22) with (3.13) we obtain

$$(3.23) \quad p(\mathbf{x} | \alpha) \geq c \cdot \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u}).$$

As will be shown later, vector  $\mathbf{u}$  is a quantity that needs to be computed and has an intuitive interpretation related to the unknown image  $\mathbf{x}$ . This leads to the following lower bound for the joint probability distribution

$$\begin{aligned} p(\alpha, \beta, \mathbf{x}, \mathbf{y}) &\geq c \cdot p(\alpha)p(\beta)\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u})p(\mathbf{y}|\mathbf{x}, \beta) \\ (3.24) \quad &= \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y}). \end{aligned}$$

By defining

$$(3.25) \quad \tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u}) = \int \int \int q(\alpha, \beta, \mathbf{x}) \log \left( \frac{q(\alpha, \beta, \mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x},$$

and utilizing inequality (3.24) we obtain

$$(3.26) \quad \mathcal{M}(q(\mathbf{x}, \alpha, \beta)) \leq \min_{\mathbf{u}} \tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u}).$$

Therefore, by finding a sequence of distributions  $\{q^k(\alpha, \beta, \mathbf{x})\}$  that monotonically decreases  $\tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u})$  for a fixed  $\mathbf{u}$  a sequence of an ever decreasing upper bound of  $C_{KL}(q^k(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y}))$  is also obtained due to (3.20). However, also minimizing  $\tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u})$  with respect to  $\mathbf{u}$  generates a sequence of vectors  $\{\mathbf{u}^k\}$  that tightens the upper-bound for each distribution  $q^k(\alpha, \beta, \mathbf{x})$ . Therefore, the two sequences  $\{q^k(\alpha, \beta, \mathbf{x})\}$  and  $\{\mathbf{u}^k\}$  are coupled. We develop below an iterative algorithm (Algorithm 1) to find such sequences.

Inequality (3.21) provides a local quadratic approximation to the TV prior. Had a fixed  $\mathbf{u}^o$  been used a global conditional auto-regression model approximating the TV prior would have

been obtained. Clearly, the procedure which updates  $\mathbf{u}$  will provide a tighter upper bound for  $\mathcal{M}(q(\mathbf{x}, \alpha, \beta))$ , since we are using  $\min_{\mathbf{u}} \tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u})$  instead of  $\tilde{\mathcal{M}}(q(\mathbf{x}, \alpha, \beta), \mathbf{u}^o)$ .

Finally we note that the process to find the best posterior distribution approximation of the image in combination with  $\mathbf{u}$  is a very natural extension of the Majorization-Minimization approach to function optimization (see [137]) and that local majorization has also been applied to variational logistic regression [113], as well as, to the inference of its parameters (see [29] and also [31]).

The following algorithm can therefore be used for calculating the approximating posteriors  $q(\alpha, \beta, \mathbf{x}) = q(\mathbf{x})q(\alpha, \beta)$ .

Let us now further develop each of the steps of the above algorithm. To calculate  $q^k(\mathbf{x})$  we observe that differentiating the integral on the right hand side of (3.27) with respect to  $q(\mathbf{x})$  and setting it equal to zero we obtain

$$(3.31) \quad q^k(\mathbf{x}) \propto \exp \left\{ E_{q^k(\alpha, \beta)} [\ln \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^k)] \right\},$$

which represents an  $N$ -dimensional Gaussian distribution with parameters

$$(3.32) \quad \text{cov}_{q^k(\mathbf{x})}[\mathbf{x}] = \left( E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{H} + E_{q^k(\alpha)}[\alpha] (\Delta^h)^t W(\mathbf{u}^k) (\Delta^h) + E_{q^k(\alpha)}[\alpha] (\Delta^v)^t W(\mathbf{u}^k) (\Delta^v) \right)^{-1},$$

and

$$(3.33) \quad E_{q^k(\mathbf{x})}[\mathbf{x}] = \text{cov}_{q^k(\mathbf{x})}[\mathbf{x}] E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{y},$$

Table 3.1. Proposed Algorithm I

**Algorithm 1.** Posterior parameter and image distributions estimation in TV restoration using  $q(\alpha, \beta, \mathbf{x}) = q(\alpha, \beta)q(\mathbf{x})$ .

Given  $\mathbf{u}^1 \in (R^+)^N$  and  $q^1(\alpha, \beta)$ , an initial estimate of the distribution  $q(\alpha, \beta)$ , for  $k = 1, 2, \dots$  until a stopping criterion is met:

(1) Find

$$(3.27) \quad q^k(\mathbf{x}) = \operatorname{argmin}_{q(\mathbf{x})} \int \int \int q^k(\alpha, \beta) q(\mathbf{x}) \log \left( \frac{q^k(\alpha, \beta) q(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^k, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

(2) Find

$$(3.28) \quad \mathbf{u}^{k+1} = \operatorname{argmin}_{\mathbf{u}} \int \int \int q^k(\alpha, \beta) q^k(\mathbf{x}) \log \left( \frac{q^k(\alpha, \beta) q^k(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

(3) Find

(3.29)

$$q^{k+1}(\alpha, \beta) = \operatorname{argmin}_{q(\alpha, \beta)} \int \int \int q(\alpha, \beta) q^k(\mathbf{x}) \log \left( \frac{q(\alpha, \beta) q^k(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^{k+1}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

Set

$$(3.30) \quad q(\alpha, \beta) = \lim_{k \rightarrow \infty} q^k(\alpha, \beta), \quad q(\mathbf{x}) = \lim_{k \rightarrow \infty} q^k(\mathbf{x}).$$

where  $W(\mathbf{u}^k)$  is an  $N \times N$  diagonal matrix of the form

$$(3.34) \quad W(\mathbf{u}^k) = \operatorname{diag} \left( \frac{1}{\sqrt{u_i^k}} \right), \quad i = 1, \dots, N,$$

and  $\Delta^h$  and  $\Delta^v$  represent the  $N \times N$  convolution matrices associated to the first order horizontal and vertical differences, respectively.

To calculate  $\mathbf{u}^{k+1}$  we have from (3.28) that

$$(3.35) \quad \mathbf{u}^{k+1} = \operatorname{argmin}_{\mathbf{u}} \sum_i \frac{\mathbf{E}_{q^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2] + u_i}{\sqrt{u_i}}$$

and consequently

$$(3.36) \quad u_i^{k+1} = E_{q^k(x)}[(\Delta_i^h(x))^2 + (\Delta_i^v(x))^2], \quad i = 1, \dots, N.$$

Notice that  $q^k(\alpha, \beta)$  is not required in calculating  $\mathbf{u}^{k+1}$ . It is clear from (3.36) that the vector  $\mathbf{u}^{k+1}$  is a function of the spatial first order differences of the unknown image  $\mathbf{x}$  under the distribution  $q^k(\mathbf{x})$  and represents the local spatial activity of  $\mathbf{x}$ . Therefore, matrix  $W(\mathbf{u}^k)$  in (3.34) can be interpreted as the *spatial adaptivity* matrix since it controls the amount of smoothing at each pixel location depending on the strength of the intensity variation at that pixel, as expressed by the horizontal and vertical intensity gradient. That is, for the pixels with high spatial activity the corresponding entries of  $W(\mathbf{u}^k)$  are very small or zero, which means that no smoothness is enforced, while for the pixels in a flat region the corresponding entries of  $W(\mathbf{u}^k)$  are very large, which means that smoothness is enforced. This matrix  $W(\mathbf{u}^k)$  has also been referred to as the visibility matrix [6] since it describes the masking property of the human visual system, according to which noise is not visible in high spatial activity regions (its high frequencies are masked by the edges), while it is visible in the low spatial frequency (flat) regions. The visibility matrix and its complementary matrix  $\mathbf{I} - W(\mathbf{u}^k)$  have been used in iterative image restoration in [125].

By differentiating the integral on the right hand side of (3.29) with respect to  $q(\alpha, \beta)$  and setting it equal to zero we obtain

$$(3.37) \quad q^{k+1}(\alpha, \beta) \propto \exp \left\{ E_{q^k(x)} [\ln F(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^{k+1})] \right\}$$

and thus

$$(3.38) \quad q^{k+1}(\alpha, \beta) = q^{k+1}(\alpha)q^{k+1}(\beta),$$

where  $q^{k+1}(\alpha)$  and  $q^{k+1}(\beta)$  are gamma distributions given respectively by

$$(3.39) \quad q^{k+1}(\alpha) \propto \alpha^{N/2+a_\alpha^o-1} \exp \left[ -\alpha \left( \sum_i \sqrt{u_i^{k+1}} + b_\alpha^o \right) \right],$$

$$(3.40) \quad q^{k+1}(\beta) \propto \beta^{N/2+a_\beta^o-1} \exp \left[ -\beta \left( \frac{E_{q^k(x)} \| \mathbf{y} - \mathbf{Hx} \|^2}{2} + b_\beta^o \right) \right].$$

The means of these gamma distributions are given by

$$(3.41) \quad E_{q^{k+1}(\alpha)}[\alpha] = \frac{N/2 + a_\alpha^o}{\sum_i \sqrt{u_i^{k+1}} + b_\alpha^o} \quad E_{q^{k+1}(\beta)}[\beta] = \frac{N/2 + a_\beta^o}{E_{q^k(x)} \| \mathbf{y} - \mathbf{Hx} \|^2 / 2 + b_\beta^o}$$

The calculation of  $E_{q^k(x)}[\mathbf{x}]$ ,  $\text{cov}_{q^k(x)}[\mathbf{x}]$ ,  $E_{q^k(x)}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2]$ , and  $E_{q^k(x)}[\| \mathbf{y} - \mathbf{Hx} \|^2]$  is carried out in Appendices A and B.

Note that we have

$$(3.42) \quad (E_{q^{k+1}(\alpha)}[\alpha])^{-1} = \frac{a_\alpha^o}{a_\alpha^o + \frac{N}{2}} \frac{b_\alpha^o}{a_\alpha^o} + \frac{N/2}{(a_\alpha^o + \frac{N}{2})} \frac{2 \sum_i \sqrt{u_i^{k+1}}}{N},$$

$$(3.43) \quad (E_{q^{k+1}(\beta)}[\beta])^{-1} = \frac{a_\beta^o}{a_\beta^o + \frac{N}{2}} \frac{b_\beta^o}{a_\beta^o} + \frac{N/2}{(a_\beta^o + \frac{N}{2})} \frac{E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2]}{N}.$$

and thus

$$(3.44) \quad (E_{q^{k+1}(\alpha)}[\alpha])^{-1} = \gamma_\alpha \frac{1}{\bar{\alpha}^o} + (1 - \gamma_\alpha) \frac{\sum_i \sqrt{u_i^{k+1}}}{N/2},$$

$$(3.45) \quad (E_{q^{k+1}(\beta)}[\beta])^{-1} = \gamma_\beta \frac{1}{\bar{\beta}^o} + (1 - \gamma_\beta) \frac{E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2]}{N},$$

where  $\bar{\alpha}^o = a_\alpha^o/b_\alpha^o$ ,  $\bar{\beta}^o = a_\beta^o/b_\beta^o$ , and

$$(3.46) \quad \gamma_\alpha = \frac{a_\alpha^o}{a_\alpha^o + \frac{N}{2}}, \quad \gamma_\beta = \frac{a_\beta^o}{a_\beta^o + \frac{N}{2}}.$$

Equation (3.46) indicates that  $\gamma_\alpha$  and  $\gamma_\beta$ , both taking values in the interval  $[0, 1)$ , can be understood as normalized confidence parameters. As can be seen from (3.44) and (3.45), the inverses of the means of the hyperpriors are represented as convex combinations of their initial values and their maximum likelihood (ML) estimates. These ML estimates have been derived before either empirically or by using regularization formulations [121] [125]. According to (3.44) and (3.45), when they are equal to zero, no confidence is placed on the initial values of the hyperparameters and ML estimates are used, while when they are asymptotically equal to one, the prior knowledge of the mean is fully enforced, i.e., no estimation of the hyperparameters is performed.

Case of particular interest is when

$$(3.47) \quad a_\alpha^o = a_\beta^o = 1 \text{ and } b_\alpha^o = b_\beta^o = \infty,$$

which corresponds to a flat hyperprior distribution. This type of hyperprior modeling makes the observation responsible for the whole estimation process.

In the proposed model for estimating the posterior distribution of the image and the unknown hyperparameters no assumptions were made about  $q(\mathbf{x})$  and  $q(\alpha, \beta)$ . We study now the case when  $q(\mathbf{x})$  is a degenerate distribution, that is, a distribution which takes one value with probability one and the rest with probability zero. In the iterative procedure we describe next

Table 3.2. Proposed Algorithm II

**Algorithm 2.** Posterior parameter and image distributions estimation in TV restoration using  $q(\alpha, \beta, \mathbf{x}) = q(\alpha, \beta)q(\mathbf{x})$  with  $q(\mathbf{x})$  a degenerate distribution.

Given  $q^1(\alpha, \beta)$ , an initial estimate of the distribution  $q(\alpha, \beta)$  and  $\mathbf{u}^1 \in (R^+)^N$ , for  $k = 1, 2, \dots$  until a stopping criterion is met:

(1) Calculate

$$(3.48) \quad \begin{aligned} \mathbf{x}^k &= \left( E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{H} + E_{q^k(\alpha)}[\alpha] (\Delta^h)^t W(\mathbf{u}^k) (\Delta^h) + E_{q^k(\alpha)}[\alpha] (\Delta^v)^t W(\mathbf{u}^k) (\Delta^v) \right)^{-1} \\ &\times E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{y} \end{aligned}$$

(2) Calculate

$$(3.49) \quad \mathbf{u}_i^{k+1} = (\Delta_i^h(\mathbf{x}^k))^2 + (\Delta_i^v(\mathbf{x}^k))^2, \quad i = 1, \dots, N.$$

(3) Calculate

$$(3.50) \quad q^{k+1}(\alpha, \beta) = q^{k+1}(\alpha)q^{k+1}(\beta)$$

where  $q^{k+1}(\alpha)$  and  $q^{k+1}(\beta)$  are gamma distributions given respectively by

$$(3.51) \quad q^{k+1}(\alpha) \propto \alpha^{N/2+a_\alpha^o-1} \exp \left[ -\alpha \left( \sum_i \sqrt{u_i^{k+1}} + b_\alpha^0 \right) \right],$$

$$(3.52) \quad q^{k+1}(\beta) \propto \beta^{N/2+a_\beta^o} \exp \left[ -\beta \left( \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}^k\|^2}{2} + b_\beta^o \right) \right].$$

Set

$$(3.53) \quad q(\alpha, \beta) = \lim_{k \rightarrow \infty} q^k(\alpha, \beta), \quad \hat{\mathbf{x}} = \lim_{k \rightarrow \infty} \mathbf{x}^k.$$

we use  $\mathbf{x}^k$  to denote the value  $q^k(\mathbf{x})$  takes with probability one. We then have the following procedure.

Two additional factorizations of the distribution  $q(\alpha, \beta, \mathbf{x})$  can be used. The first one corresponds to assuming that  $q(\alpha, \beta)$  is a degenerate distribution. In this case, selecting as image estimate the mean value of the limiting  $q(\mathbf{x})$  distribution in the corresponding algorithm is equivalent to performing the evidence analysis for the TV restoration problem. The second one

corresponds to assuming that both  $q(\alpha, \beta)$  and  $q(\mathbf{x})$  are degenerate distributions. The corresponding algorithm is equivalent to maximizing alternatively in the hyperparameters and image the lower bound of  $p(\alpha, \beta, \mathbf{x})$  given in (3.24). In other words, the estimation procedure is an iterated conditional mode (ICM) algorithm [25].

To end this section we comment on two particular hyperparameter distributions  $p(\alpha, \beta)$ . The first one is obtained when both  $\beta$  and  $\alpha$  are known quantities. Then Algorithm 2 with  $\gamma_\alpha = \gamma_b = 1$ ,  $\bar{\alpha}^o = \alpha$ , and  $\bar{\beta}^o = \beta$ , provides the same solution with

$$(3.54) \quad \begin{aligned} \underline{\mathbf{x}} &= \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \frac{\beta}{2} \| \mathbf{y} - \mathbf{Hx} \|^2 + \alpha \operatorname{TV}(\mathbf{x}) \right\} \\ &= \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \| \mathbf{y} - \mathbf{Hx} \|^2 + \frac{2\alpha}{\beta} \operatorname{TV}(\mathbf{x}) \right\}. \end{aligned}$$

If  $\alpha = k/2$  with  $k = 0.064$ , the estimate of (3.54) is the one used in [27], and referred to as algorithm *BFO1* in Section 3.5.

The second hyperparameter distribution  $p(\alpha, \beta)$  is obtained when only  $\beta$  is known, that is,  $\gamma_b = 1$ ,  $\bar{\beta}^o = \beta$ , and when  $p(\alpha) \propto \alpha^{-1}$  and  $\gamma_\alpha = 0$ . Then Algorithm 2 at convergence provides (see (3.44))

$$(3.55) \quad E_{q(\alpha)}[\alpha] = \frac{N/2}{\operatorname{TV}(\hat{\mathbf{x}})},$$

and the solution for the image in (3.48) satisfies

$$(3.56) \quad -\beta \mathbf{H}^t (\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}) + \frac{N/2}{\operatorname{TV}(\hat{\mathbf{x}})} [(\Delta^h)^t W(\hat{\mathbf{u}})(\Delta^h) + (\Delta^v)^t W(\hat{\mathbf{u}})(\Delta^v)] \hat{\mathbf{x}} = 0$$

with

$$(3.57) \quad \hat{u}_i = (\Delta_i^h(\hat{\mathbf{x}}))^2 + (\Delta_i^v(\hat{\mathbf{x}}))^2, \quad i = 1, \dots, N.$$

Now, regularizing  $W(\mathbf{u})$  by using  $\text{diag}(1/\sqrt{u_i + \varepsilon})$  where  $\varepsilon$  is a small positive constant to obtain a differentiable TV norm we have

$$(3.58) \quad \frac{\partial}{\partial \mathbf{x}} \text{TV}(\mathbf{x}) \Big|_{\mathbf{x}=\hat{\mathbf{x}}} = [(\Delta^h)^t W(\hat{\mathbf{u}})(\Delta^h) + (\Delta^v)^t W(\hat{\mathbf{u}})(\Delta^v)]\hat{\mathbf{x}}.$$

Therefore, (3.56) can be rewritten as

$$(3.59) \quad -\beta \mathbf{H}^t(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}) + \frac{N/2}{\text{TV}(\hat{\mathbf{x}})} \frac{\partial}{\partial \mathbf{x}} \text{TV}(\mathbf{x}) \Big|_{\mathbf{x}=\hat{\mathbf{x}}} = 0.$$

That is, for this particular selection of  $p(\alpha, \beta)$ , Algorithm 2 provides the solution of

$$(3.60) \quad \hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \left\{ \frac{\beta}{2} \| \mathbf{y} - \mathbf{Hx} \|^2 + \frac{N}{2} \log \text{TV}(\mathbf{x}) \right\}.$$

Interestingly, this image estimate coincides with the image estimate proposed in [26], and referred to as algorithm *BFO2* in Section 3.5, which is obtained as

$$(3.61) \quad \begin{aligned} \bar{\mathbf{x}} &= \operatorname{argmax}_{\mathbf{x}} \left\{ \log \left( \exp \left[ -\frac{\beta}{2} \| \mathbf{y} - \mathbf{Hx} \|^2 \right] \int \alpha^{N/2-1} \exp[-\alpha \text{TV}(\mathbf{x})] d\alpha \right) \right\} \\ &= \operatorname{argmin}_{\mathbf{x}} \left\{ \frac{\beta}{2} \| \mathbf{y} - \mathbf{Hx} \|^2 + \frac{N}{2} \log \text{TV}(\mathbf{x}) \right\}, \end{aligned}$$

Clearly, Algorithm 2 is a generalization of the algorithms presented in [27] and [26].

### 3.5. Experimental Results

We performed a number of experiments to evaluate the performance of the proposed algorithms and also to compare them with other image restoration methods in the literature. We present results with Algorithm 1 (denoted by *ALG1*), Algorithm 2 (denoted by *ALG2*) and the TV-based approaches in [27] and [26], denoted (see the end of the previous section) by *BFO1* and *BFO2*, respectively. As already shown algorithms *BFO1* and *BFO2* are special cases of *ALG2*. We will elaborate on the differences and similarities of the methods in conjunction with the results. As in [27] and [26] we use a conjugate gradient algorithm (CG) to find the *BFO1* and *BFO2* image estimates.

We also included results obtained with the use of the algorithm in [166] which models the image distribution by a simultaneous autoregression (SAR) model [200] instead of a TV model and simultaneously estimates the prior and image hyperparameters. This algorithm will be denoted by *MOL* in the results. Comparing TV-based algorithms with this method provided useful insights about the proposed approaches.

In evaluating the upper bound of the performance of the proposed algorithms we also provide results obtained by the algorithms denoted by *ALG1-TrueU*, *ALG2-TrueU*, *ALG1-True*, and *ALG2-True*. For the *ALG1-TrueU* and *ALG2-TrueU* algorithms the noise variance  $1/\beta$  is known (since we are dealing with synthetic experiments), and  $\alpha$  and  $\mathbf{u}$  are calculated using the original image ( $\alpha$  from the equation  $\alpha = \frac{N/2}{\sum_i \sqrt{u_i}}$  and  $\mathbf{u}$  from (3.36) and (3.49)).

All three parameters are computed once and thus they are not updated during the iterations. For the *ALG1-True* and *ALG2-True* algorithms  $\alpha$  and  $\beta$  are treated as in *ALG1-TrueU* and *ALG2-TrueU*, but  $\mathbf{u}$  is evaluated iteratively.

In our results we provided the improvement in signal-to-noise ratio (ISNR) as an objective measure of the quality. The ISNR is defined as  $10\log_{10}(\|x - y\|^2 / \|x - \hat{x}\|^2)$ , where  $x$ ,  $y$  and  $\hat{x}$  are the original, observed, and estimated images, respectively. In the tables we present in this section, we report the ISNR values, number of iterations, and estimated noise variances using a conjugate gradient (CG) approach (values in parentheses are obtained using a gradient descent (GD) approach to solve (3.33) and (3.48), as further discussed in Appendix A). Note that since the parameter  $1/\beta$  is not estimated by the algorithms *BFO1* and *BFO2*, but it is assumed known, the corresponding entries are denoted by “-”. For all experiments,  $\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 < 10^{-4}$  (or  $E_{q^k(\mathbf{x})}[\mathbf{x}]$  instead of  $\mathbf{x}^k$ ) is used to terminate the algorithms, and a threshold of  $10^{-6}$  is used to terminate the CG and GD iterations.

For the first set of experiments, we synthetically degraded the "Lena" and "Cameraman" images and the "Shepp-Logan" phantom with a Gaussian blur with variance 9 and additive Gaussian noise. We experimented with three noise levels, corresponding to blurred signal-to-noise ratios (BSNR) of 40, 30, and 20 dB. The original Lena image is shown in Fig. 3.2(a) and the degraded versions with the three noise levels in Figs. 3.2(b)-(d) (the corresponding noise variances are equal to 0.16, 1.6, and 16).

Flat hyperpriors on the hyperparameters are used as initial conditions, i.e.,  $a_\alpha^o = a_\beta^o = 0$  and  $b_\alpha^o = b_\beta^o = 0$ . The initially selected values for  $E_{q^1(\alpha)}[\alpha]$  and  $E_{q^1(\beta)}[\beta]$  for both *ALG1* and *ALG2* methods were equal to

$$(3.62) \quad E_{q^1(\alpha)}[\alpha] = \frac{N/2}{\sum_i \sqrt{u_i}}, \quad E_{q^1(\beta)}[\beta] = \frac{N/2}{\|\mathbf{y} - \mathbf{H}\mathbf{y}\|^2 / 2},$$



Figure 3.2. (a) Lena image; degraded with a Gaussian shaped PSF with variance 9 and Gaussian noise of variance: (b) 0.16 (BSNR = 40 dB), (c) 1.6 (BSNR = 30 dB), (d) 16 (BSNR = 20 dB).

that is, we used the observations to initialize the hyperprior means. The observed image is used as the initial value of  $\mathbf{x}$ , and the initial value of  $\mathbf{u}$  is calculated from this observed image. Note that the algorithms are initialized automatically without any manual input.

The ISNR values, the number of iterations, and the estimates of the noise variance  $1/\beta$  are shown in Table 3.3 (it is noted that the true value of the noise variance is reported for the algorithms with the "True" suffix). In the second set of experiments, the same images are degraded by a 9x9 uniform blur and additive Gaussian noise. The corresponding results are shown in Table 3.4.

It is clear that knowledge of the noise and image parameters provides an advantage for *BFO1*; this method outperforms other methods in nearly all noise levels. However, both *ALG1* and *ALG2* result in comparable, in some cases even higher ISNR values, despite the fact that no prior information is assumed about the degradation process. We will later show that with the use of hyperpriors on the unknown hyperparameters higher ISNR values to the ones obtained by *BFO1* can be achieved by the *ALG1* and *ALG2* algorithms.

Table 3.3. ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 9.

BSNR	Method	Lena			Cameraman			Shepp-Logan		
		ISNR (dB)	iterations	1/ $\beta$	ISNR (dB)	iterations	1/ $\beta$	ISNR (dB)	iterations	1/ $\beta$
40dB	<i>MOL</i>	3.90	26	0.16	2.73	35	0.30	3.67	40	0.45
	<i>BFO1</i>	4.72	20	-	3.51	19	-	7.07	19	-
	<i>BFO2</i>	4.50	19	-	3.27	16	-	5.88	17	-
	<i>ALG1</i>	4.78(4.84)	15(10)	0.16 (0.16)	3.39(3.38)	21(16)	0.30(0.30)	6.69(6.46)	36(35)	0.45(0.45)
	<i>ALG2</i>	4.49(4.64)	17(16)	0.16 (0.16)	3.26(3.28)	18(17)	0.30(0.30)	5.63(5.56)	20(21)	0.45(0.45)
	<i>ALG1-TrueU</i>	6.95	3	0.16	5.35	13	0.30	59.33	7	0.45
	<i>ALG2-TrueU</i>	6.95	4	0.16	5.33	18	0.30	58.91	7	0.45
	<i>ALG1-True</i>	4.84(4.89)	12(8)	0.16	3.49(3.44)	16(12)	0.30	6.60(6.42)	36(37)	0.45
	<i>ALG2-True</i>	4.65(4.86)	21(13)	0.16	3.49(3.49)	18(14)	0.30	5.95(6.09)	16(19)	0.45
30dB	<i>MOL</i>	3.13	27	1.62	2.14	35	3.05	2.91	49	4.53
	<i>BFO1</i>	3.87	24	-	2.89	18	-	5.15	18	-
	<i>BFO2</i>	3.56	21	-	2.47	16	-	3.94	16	-
	<i>ALG1</i>	3.87(4.03)	24(16)	1.60(1.60)	2.63(2.68)	26(23)	4.60(4.60)	4.31(4.30)	43(44)	4.60(4.60)
	<i>ALG2</i>	3.55(3.67)	20(21)	1.60(1.60)	2.41(2.49)	19(23)	4.78(4.70)	3.72(3.78)	16(39)	4.78(4.70)
	<i>ALG1-TrueU</i>	6.01	2	1.60	4.55	5	3.00	43.23	10	4.50
	<i>ALG2-TrueU</i>	6.01	2	1.60	4.55	6	3.00	36.38	9	4.50
	<i>ALG1-True</i>	3.94(4.12)	22(14)	1.60	2.86(2.91)	20(16)	3.00	4.31(4.31)	43(44)	4.50
	<i>ALG2-True</i>	3.81(3.95)	23(20)	1.60	2.87(2.95)	19(19)	3.00	4.13(4.21)	16(53)	4.50
20dB	<i>MOL</i>	2.45	4	16.05	1.64	3	30.23	2.24	3	45.11
	<i>BFO1</i>	3.02	20	-	2.13	16	-	3.56	17	-
	<i>BFO2</i>	2.47	23	-	2.23	29	-	2.20	13	-
	<i>ALG1</i>	2.87(3.06)	30(23)	16.22(16.20)	1.72(1.78)	38(36)	31.69(31.57)	1.85(1.86)	47(48)	48.66(48.61)
	<i>ALG2</i>	2.42(2.58)	23(34)	16.74(16.48)	1.42(1.58)	16(30)	33.52(32.62)	2.05(1.72)	12(53)	51.11(49.74)
	<i>ALG1-TrueU</i>	5.07	2	16.00	3.71	5	30.00	24.05	91	45.00
	<i>ALG2-TrueU</i>	5.07	2	16.00	3.70	6	30.00	21.56	72	45.00
	<i>ALG1-True</i>	3.05(3.25)	25(21)	16.00	2.13(2.20)	27(27)	30.00	2.20(2.21)	48(48)	45.00
	<i>ALG2-True</i>	2.92(3.04)	20(28)	16.00	2.10(2.20)	15(29)	30.00	2.49(2.28)	14(51)	45.00

Table 3.4. ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a 9x9 uniform blur.

BSNR	Method	Lena			Cameraman			Shepp-Logan		
		ISNR (dB)	iterations	1/ $\beta$	ISNR (dB)	iterations	1/ $\beta$	ISNR (dB)	iterations	1/ $\beta$
40dB	<i>MOL</i>	6.79	21	0.17	6.16	25	0.31	5.82	27	0.48
	<i>BFO1</i>	8.34	8	-	8.55	8	-	14.22	12	-
	<i>BFO2</i>	8.35	9	-	8.25	12	-	12.01	12	-
	<i>ALGI</i>	8.42(8.24)	21(19)	0.12(0.12)	8.57(8.34)	28(24)	0.25(0.25)	13.69(13.32)	19(20)	0.40(0.40)
	<i>ALG2</i>	8.37(8.36)	16(15)	0.12(0.13)	8.46(8.29)	18(21)	0.27(0.27)	13.26(12.9)	15(17)	0.44(0.44)
	<i>ALG1-TrueU</i>	10.93	3	0.17	11.34	9	0.31	58.62	7	0.46
	<i>ALG2-TrueU</i>	10.93	3	0.17	11.33	13	0.31	58.41	7	0.46
	<i>ALG1-True</i>	8.48(8.43)	7(7)	0.17	8.58(8.19)	9(8)	0.31	13.53(13.52)	21(17)	0.46
30dB	<i>ALG2-True</i>	8.46(8.40)	7(9)	0.17	8.53(8.51)	10(11)	0.31	12.14(12.68)	12(16)	0.46
	<i>MOL</i>	4.62	30	1.71	3.98	39	3.22	4.10	41	4.69
	<i>BFO1</i>	6.08	11	-	5.68	10	-	8.88	12	-
	<i>BFO2</i>	5.64	12	-	4.65	14	-	6.91	13	-
	<i>ALGI</i>	5.89(5.89)	13(14)	1.49(1.49)	5.41(5.41)	20(19)	3.04(3.04)	7.77(7.75)	31(29)	4.60(4.60)
	<i>ALG2</i>	5.58(5.61)	14(14)	1.64(1.64)	4.38(4.39)	15(17)	3.60(3.60)	6.50(6.85)	15(29)	4.84(4.77)
	<i>ALG1-TrueU</i>	8.23	2	1.66	8.26	5	3.10	44.63	8	4.60
	<i>ALG2-TrueU</i>	8.23	2	1.66	8.26	6	3.10	38.01	7	4.60
20dB	<i>ALG1-True</i>	5.96(5.97)	9(8)	1.66	5.70(5.72)	11(11)	3.10	7.72(7.70)	30(29)	4.60
	<i>ALG2-True</i>	6.02(6.04)	10(11)	1.66	5.61(5.64)	12(13)	3.10	7.14(7.44)	13(24)	4.60
	<i>MOL</i>	2.94	18	17.07	2.26	4	32.42	2.66	5	45.68
	<i>BFO1</i>	4.09	14	-	3.31	14	-	5.57	16	-
	<i>BFO2</i>	4.14	16	-	2.12	20	-	2.95	14	-
	<i>ALGI</i>	3.72(3.83)	13(13)	17.00(16.91)	2.42(2.46)	23(22)	34.05(34.00)	3.01(3.03)	35(34)	49.97(49.88)
	<i>ALG2</i>	3.15(3.28)	12(22)	17.93(17.59)	1.94(2.12)	16(28)	36.52(35.56)	2.64(2.54)	12(40)	53.67(51.92)
	<i>ALG1-TrueU</i>	8.23	2	16.67	5.33	5	31.00	25.10	85	46.00
10dB	<i>ALG2-TrueU</i>	8.23	2	16.67	5.33	6	31.00	22.97	68	46.00
	<i>ALG1-True</i>	5.95(5.96)	9(8)	16.67	3.33(3.38)	14(14)	31.00	3.51(3.51)	33(33)	46.00
	<i>ALG2-True</i>	6.02(6.04)	10(11)	16.67	3.25(3.32)	17(18)	31.00	3.34(3.37)	15(34)	46.00

The important point to note here is that *ALG1* and *ALG2* generally perform better than *BFO2* and *MOL*. The proposed methods generally result in higher ISNR values than *BFO2*, although the noise variance is assumed to be known in *BFO2*. The *MOL* algorithm is outperformed by other methods in all experiments, although the noise variance is very accurately estimated. This comparison clearly shows that the spatially adaptive deconvolution and noise removal achieved by TV-based restoration methods provides a significant improvement over methods like *MOL* which do not incorporate spatial adaptivity in the estimation procedure.

We also note that the proposed methods are robust to the initial values of the hyperparameters. For instance, when the algorithms are initialized using  $\beta = 1$  and  $\alpha = \frac{0.064}{2} \sigma^2$ , as in [27], the resulting ISNR values are similar to the ones reported in Table 3.3. For instance, for the 40 dB BSNR case with the Lena image, the ISNR values are 4.64 (4.75) dB and 4.34 (4.42) dB, and for the 20 dB BSNR case, the ISNR values are 2.88 (3.06) dB and 2.45 (2.51) dB for the *ALG1* and *ALG2* methods, respectively. These results show the robustness of the methods to parameter initialization.

Although the results in Table 3.4 are similar to the Gaussian blur case, we note some interesting differences. It is clear that *ALG2* outperforms *ALG1* in high BSNRs, but it results in a lower ISNR in the low BSNR case. We can conclude that in the high BSNR case, where the noise level is low, exploiting additional information using the full variational formulation actually results in lower performance. However, using the full variational algorithm, i.e., *ALG1*, provides better image estimates in the low BSNR case. Another remark is that both algorithms fail to accurately estimate the noise variance when the noise level is very low at 40dB BSNR, although the estimated noise variance is very close to the true value at high noise levels.



Figure 3.3. Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) *MOL* method (ISNR = 3.90 dB), (b) *BFO1* method (ISNR = 4.72 dB), (c) *BFO2* method (ISNR = 4.5 dB), (d) *ALG1* method (ISNR = 4.84 dB), and (e) *ALG2* method (ISNR = 4.64 dB).



Figure 3.4. Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) *MOL* method (ISNR = 2.45 dB), (b) *BFO1* method (ISNR = 3.02 dB), (c) *BFO2* method (ISNR = 2.47 dB), (d) *ALG1* method (ISNR = 3.06 dB), and (e) *ALG2* method (ISNR = 2.58 dB).

The results obtained with the use of GD and CG are comparable, although in most cases GD results in fewer iterations.

A fair comparison between *BFO1* and the proposed approaches can be made by looking at the performances of *ALG1-True* and *ALG2-True*. In most cases *ALG1-True* and *ALG2-True* outperform *BFO1*, while a smaller number of iterations is adequate for convergence. Additionally, *ALG1-TrueU* and *ALG2-TrueU* provide the upper bound in ISNR that can be achieved by TV-based restoration methods represented here. Clearly, knowledge of the true value of the matrix  $\mathbf{u}$  provides a significant advantage to the methods.

We next examine the effect of the introduction of additional information about the unknown hyperparameters through the use of the confidence parameters  $\gamma_\alpha$  and  $\gamma_\beta$  on the performance of the algorithms. As we have already explained before, in the case of  $\gamma_\alpha = \gamma_\beta = 0$ , no information about the hyperparameters is available, and the observed image is responsible for the estimation of the hyperparameters and the image. However, one usually has some information about the original image and the degradation process. For example, off-line estimates of the image and noise variance can be computed, and provided to the algorithms. In our experiments, we provided the true image and noise variance to the algorithms and run the algorithms while varying the confidence parameters  $\gamma_\alpha$  and  $\gamma_\beta$  from 0 to 1 in 0.1 intervals.

Table 3.5 shows the means of the posterior distributions of the hyperparameters, ISNR values, and the number of iterations obtained using *ALG1* for selected values of the confidence parameters. The confidence values are selected to demonstrate the behavior of the algorithm in the following cases: 1) when full information about the image and noise variance is available, 2) when no information is provided, i.e., the observation is fully responsible for the restoration, 3)

when some information about the image prior variance  $\alpha$  is provided, and 4) when some information about the noise variance is provided. Moreover, the evolution of ISNR for the full set of confidence parameters is depicted in Fig. (3.5). A similar ISNR evolution is obtained for *ALG2* so its corresponding plot is not shown. It can be observed that the noise level changes the effect of the confidence parameters. In the low noise case (BSNR = 40 dB), information about the noise variance affects the final ISNR more than the information about the image variance; there is almost no ISNR variance when  $\gamma_\beta = 1$  and  $\gamma_\alpha$  changes from 0 to 1. However, in the 20 dB BSNR case information about the image variance is more valuable than the noise variance. For a fixed  $\gamma_\alpha$ , the ISNR value remains fixed for varying  $\gamma_\beta$ , whereas increasing the image variance confidence increases the obtained ISNR. It can be stated as a final remark that the algorithm is less successful at estimating the noise variance in low noise conditions, and less successful at estimating the image variance in high noise conditions. Therefore information about the poorly estimated parameter helps to further increase the ISNR values. However, we should also state that the ISNR variation in these plots is small compared to the ISNR values (difference between the maximum and minimum ISNR values are 0.13 dB at 40 dB BSNR and 0.19 dB at 20 dB BSNR), therefore we can see that the algorithm is robust to the estimated hyperparameter values in terms of the final restored image quality.

We will now examine the additional information provided by the variational approach and study how the distributions on the hyperparameters can be used to improve the results already obtained. We start our experiments by assuming flat hyperpriors, and applied algorithm *ALG2* to the Lena image degraded by a Gaussian blur of variance 9 and additive Gaussian noise at 40 dB and 20 dB BSNR, as we had before. This provides estimates of the noise and image variance, denoted by  $\hat{\beta}$  and  $\hat{\alpha}$ , respectively. Next we run algorithm *ALG2* multiple times on the

same degraded image with different initial hyperparameters: The final estimated noise variance of the algorithm is used without update, i.e.,  $\bar{\beta}^o = \hat{\beta}$  and  $\gamma_\beta = 1$ . By moving  $\gamma_\alpha$  in  $[0, 1]$  and selecting the hyperprior mean as  $\bar{\alpha}^o = d \cdot \hat{\alpha}$ , where  $d$  is in the range  $[10^{-1}, 10^2]$ , we obtain the ISNR evolution graphs shown in Fig. 3.6(a) for the 40 dB BSNR case and Fig. 3.6(b) for the 20 dB BSNR case. It should be noted that the range of ISNR values obtained by this experiment includes the best ISNR achieved with known hyperparameter values, depicted in Table 3.3, corresponding to *ALG2-True*. Thus, as expected, the results by *ALG1-True* and *ALG2-True* are included in the case when different selections of the gamma hyperpriors on the hyperparameters are used. A few remarks can be made by examining at Fig. (3.6): First, the algorithms are very robust with respect to the parameter  $\alpha$ , since even in the case  $\bar{\alpha}^o = 100 \cdot \hat{\alpha}$  the resulting ISNR value is very close to the highest achievable value. Second, one can conclude that the distribution of  $\alpha$  is not sharply peaked at one value, and therefore multiple values of this parameter can be used in the restoration process without greatly affecting the performance of the algorithm.

Overall, the experimental results demonstrate that algorithms *ALG1* and *ALG2* provide comparable performance to the existing TV-based approaches even though no prior knowledge about the image and degradation process is assumed, and outperform them if prior knowledge is utilized. It is also clear that TV-based approaches result in higher quality restorations than non-spatially adaptive restoration methods. Another important point to be made is that with the developed framework, we can draw different estimates for the unknown hyperparameters from their estimated distributions and thus assign a degree of trust to the results and potentially achieve improved restoration results. The major distinction between the proposed algorithms *ALG1* and *ALG2* is that *ALG1* provides the approximation to the posterior distribution of the

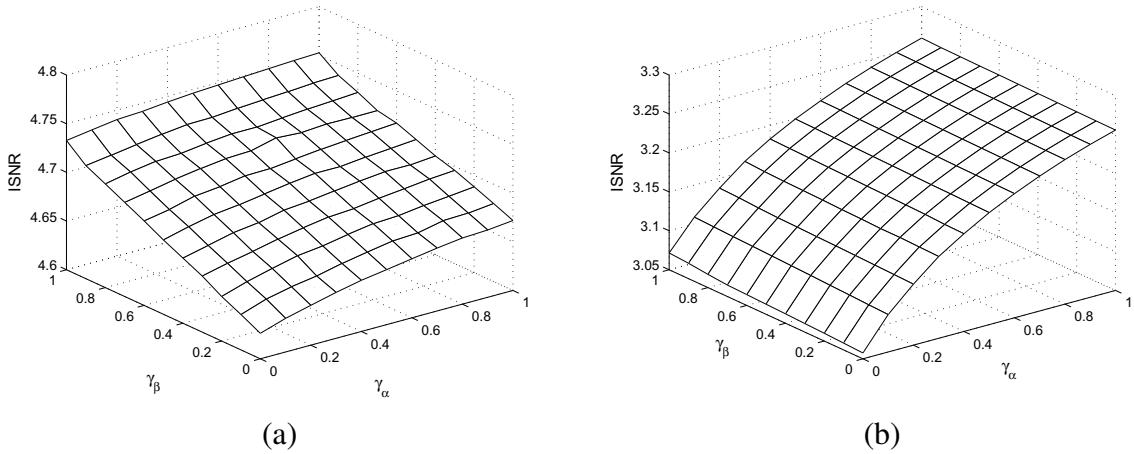


Figure 3.5. Evolution of ISNR using *ALG1* for different values of  $\gamma_\alpha$  and  $\gamma_\beta$  for the restoration of the Lena image blurred with a Gaussian with variance 9, and (a) BSNR = 40 dB; (b) BSNR = 20 dB.

unknown image. For scientific applications for which a confidence value for a restoration is important (i.e., restoration of astronomical or medical images), *ALG1* can provide such information through the use of this posterior distribution. On the other hand, when images are restored for, for example, consumer applications *ALG2* can be the algorithm of choice.

The proposed algorithms are computationally more intensive than non-spatially adaptive restoration methods since (3.33) and (3.48) cannot be solved by direct inversion in the frequency domain and numerical approaches are utilized. Typically, the MATLAB implementations of our algorithms required on the average about 2-5 minutes on a 3.20 GHz Xeon PC for 256x256 images. Note that the running time of the algorithms can be improved by utilizing preconditioning methods (see, for example, [52] [239] [57] [178]).

Table 3.5. Posterior means of the distributions of the hyperparameters, ISNR, and number of iterations using *ALG1* for the Lena image with 40 dB and 20 dB BSNR using  $\bar{\alpha}^o = 1/23.84$ , and  $\bar{\beta}^o = 1/0.16$  and  $\bar{\beta}^o = 1/16$ , respectively, for different values of  $\gamma_\beta$  and  $\gamma_\alpha$ .

BSNR	$\gamma_\beta$	$\gamma_\alpha$	$E[\beta]^{-1}$	$E[\alpha]^{-1}$	ISNR (dB)	iterations
40dB	1.0	1.0	0.16	23.84	4.75	13
	0.0	0.5	0.24	22.26	4.66	16
	0.5	0.0	0.20	20.30	4.68	16
	0.0	0.0	0.24	19.89	4.63	18
20dB	1.0	1.0	16.00	23.84	3.07	26
	0.0	0.5	16.17	20.83	3.01	28
	0.5	0.0	16.16	16.29	2.89	30
	0.0	0.0	16.34	16.29	2.88	30

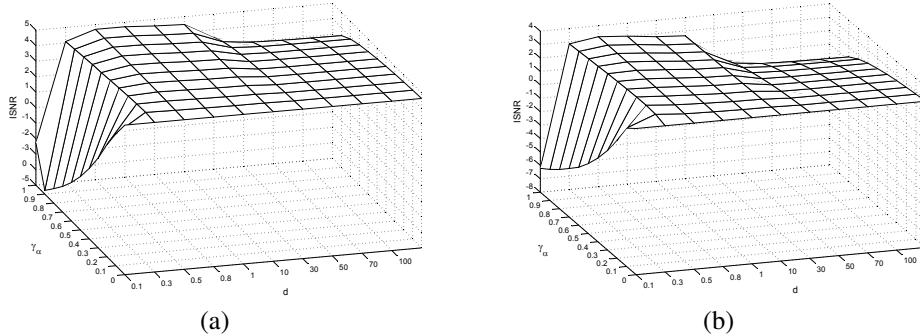


Figure 3.6. Evolution of ISNR with varying  $\gamma_\alpha$  and  $\bar{\alpha}^o$  for Lena image degraded with Gaussian blur with variance 9 at (a) 40 dB BSNR and (b) 20 dB BSNR (Note that  $\bar{\alpha}^o = d \cdot \hat{\alpha}$ ).

### 3.6. Conclusions

We have presented two new methods for the simultaneous estimation of the image and the unknown hyperparameters in TV-based image restoration problems. We adopt a variational approach to provide approximations to the posterior distributions of the unknown variables. Utilizing this variational framework, different values from the posterior distributions can be drawn as estimates to the latent variables and prior information about the degradation process and the

unknowns can be incorporated into the estimation process to increase the performance of the algorithms. We have analyzed the proposed methods and demonstrated that some of the current methods in TV-based image restoration are special cases of our formulation. Experimental results are provided to show the performance of the methods in the case where information about the degradation process and the unknown variables is not available, and when some information can be provided for improved performance.

## CHAPTER 4

# **Generalized Gaussian Markov Random Field Image Restoration Using Variational Distribution Approximation**

### **4.1. Introduction**

In this chapter we present novel algorithms for image restoration and parameter estimation with a Generalized Gaussian Markov Random Field (GGMRF) [37] [146] image prior utilizing variational distribution approximation. The restored image and the unknown hyperparameters for both the image prior and the image degradation noise are simultaneously estimated within a hierarchical Bayesian framework. Two algorithms are developed using this formulation that jointly provide estimates of the posterior distributions of the restored image and the hyperparameters.

We utilize a GGMRF as the image prior. In addition to the unknown image and noise, the hyperparameters are also cast into the Bayesian framework and simultaneously estimated. This is in contrast to the methods in literature utilizing GGMRF priors. For instance, in [37] and [146] point estimates for the unknown image are found and the hyperparameters are not estimated explicitly, but they are instead marginalized (evidence approach). In addition, in [146] a Poisson noise model is utilized.

This chapter is organized as follows. The hierarchical Bayesian model is presented in Sec. 4.2. Section 4.3 describes the variational approach to distribution approximation and the

---

<sup>0</sup>This work has appeared in [16]

derivation of our algorithms. We present the experimental results in Sec. 4.4 and conclude in Sec. 4.5.

## 4.2. Bayesian Modeling

The Bayesian modeling of the GGMRF restoration problem requires first the definition of a joint distribution  $p(\alpha, \beta, \mathbf{x}, \mathbf{y})$  of the observation,  $\mathbf{y}$ , the unknown image,  $\mathbf{x}$ , and the hyperparameters  $\alpha$  (to be defined below) and  $\beta$ . We utilize the hierarchical Bayesian paradigm where in the first stage we form the prior distributions  $p(\mathbf{y}|\mathbf{x}, \beta)$  and  $p(\mathbf{x}|\alpha)$  for the unknowns, and in the second stage we define hyperpriors on the hyperparameters. The joint probability model is shown in graphical form in Fig. 4.1(a) using a directed acyclic graph.

### 4.2.1. First stage: prior models on image and observation

The probability distribution corresponding to the observation model in (2.4) is given by

$$(4.1) \quad p(\mathbf{y}|\mathbf{x}, \beta) \propto \beta^{N/2} \exp \left[ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2 \right]$$

As the image model we use the GGMRF prior, given by

$$(4.2) \quad p(\mathbf{x}|\alpha) \propto \frac{1}{Z_{\mathbf{GG}}(\alpha)} \exp[-\alpha \mathbf{GG}(\mathbf{x})],$$

where  $Z_{\mathbf{GG}}(\alpha)$  is the partition function and

$$\mathbf{GG}(\mathbf{x}) = \sum_i \sum_{d=1}^4 \left[ |\Delta_i^d(\mathbf{x})|^p \right],$$

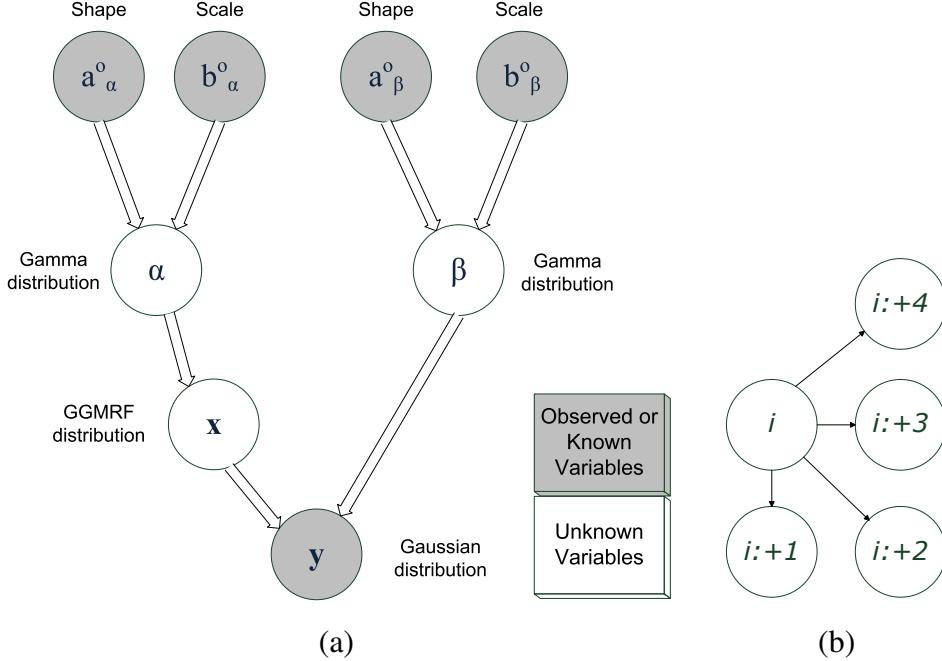


Figure 4.1. (a) Graphical model showing relationships between variables, (b) the directions for the first order differences around the pixel  $i$ .

where the first summation is over all pixels  $i$ ,  $p \in [1, 2]$ , and  $\Delta_i^d(\mathbf{x})$  denotes the first order difference in the  $d$  direction, such that

$$\Delta_i^d(\mathbf{x}) = x_i - x_{i:+d}, \quad d = 1, \dots, 4$$

Figure 4.1(b) shows the directions  $d = 1, \dots, 4$  along which the first order differences are taken.

Using  $u^p = v$  and taking into account that

$$\int_0^\infty \exp[-\alpha u^p] du = \frac{1}{p} \int_0^\infty \exp[-\alpha v] v^{\frac{1-p}{p}} dv \propto \alpha^{-\frac{1}{p}},$$

we use the approximation  $\alpha^{-N/p}$  to the partition function to obtain

$$(4.3) \quad p(\mathbf{x}|\alpha) \propto \alpha^{N/p} \exp[-\alpha \mathbf{G}\mathbf{G}(\mathbf{x})].$$

#### 4.2.2. Second stage: hyperprior on the hyperparameters

We use Gamma distributions as our model for the hyperparameters  $\omega \in \{\alpha, \beta\}$ , given by

$$(4.4) \quad p(\omega) = \Gamma(\omega | a_\omega^o, b_\omega^o) = \frac{(b_\omega^o)^{a_\omega^o}}{\Gamma(a_\omega^o)} \omega^{a_\omega^o - 1} \exp[-\omega b_\omega^o].$$

Combining the first and second stage, the joint distribution can be written as

$$(4.5) \quad p(\alpha, \beta, \mathbf{x}, \mathbf{y}) = p(\alpha)p(\beta)p(\mathbf{x}|\alpha)p(\mathbf{y}|\mathbf{x}, \beta).$$

#### 4.3. Inference and Variational Approximation

The Bayesian inference on  $(\alpha, \beta, \mathbf{x})$  should be based on

$$(4.6) \quad p(\alpha, \beta, \mathbf{x} | \mathbf{y}) = \frac{p(\alpha, \beta, \mathbf{x}, \mathbf{y})}{p(\mathbf{y})}.$$

However, since the posterior  $p(\alpha, \beta, \mathbf{x} | \mathbf{y})$  cannot be found in closed form, we approximate it by a simpler parametric form  $q(\alpha, \beta, \mathbf{x}) = q(\alpha, \beta)q(\mathbf{x})$ . This distribution can be found in a variational framework by minimizing the Kullback-Leibner (KL) distance, that is,

$$(4.7) \quad \begin{aligned} C_{KL}(q(\alpha, \beta)q(\mathbf{x}) \| p(\alpha, \beta, \mathbf{x} | \mathbf{y})) &= \int \int \int q(\alpha, \beta)q(\mathbf{x}) \log \left( \frac{q(\alpha, \beta)q(\mathbf{x})}{p(\alpha, \beta, \mathbf{x} | \mathbf{y})} \right) d\alpha d\beta d\mathbf{x} \\ &= \int \int \int q(\alpha, \beta)q(\mathbf{x}) \log \left( \frac{q(\alpha, \beta)q(\mathbf{x})}{p(\alpha, \beta, \mathbf{x}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x} + \text{const}, \end{aligned}$$

which is always non negative and equal to zero only when  $q(\alpha, \beta)q(\mathbf{x}) = p(\alpha, \beta, \mathbf{x} | \mathbf{y})$ .

Due to the form of our image prior, the KL distance cannot be minimized directly. We define the functional  $\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u})$  for  $\alpha, x$  and  $v \in (R^4+)^N$ , with components  $(u_{i,1}, \dots, u_{i,4}), i = 1, \dots, N$

$$\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u}) = \alpha^{N/p} \exp \left[ -\frac{\alpha p}{2} \sum_i \sum_{d=1}^4 \left[ \frac{(\Delta_i^d(\mathbf{x}))^2 + \frac{2-p}{p} u_{i,d}}{u_{i,d}^{1-p/2}} \right] \right].$$

Next, using the following inequality for  $w \geq 0, z > 0$ , and  $p \in [1, 2]$

$$(4.8) \quad w^{p/2} \leq z^{p/2} + \frac{p}{2z^{1-p/2}}(w-z) = \frac{p}{2} \frac{(w + \frac{2-p}{p}z)}{z^{1-p/2}},$$

we find a lower bound for the image prior, given by

$$p(\mathbf{x}|\alpha) \geq c \cdot \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u}),$$

where  $c$  is a constant. This inequality can be used to find a lower bound for the joint probability distribution

$$(4.9) \quad \begin{aligned} p(\alpha, \beta, \mathbf{x}, \mathbf{y}) &\geq c \cdot p(\alpha)p(\beta)\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\alpha, \mathbf{x}, \mathbf{u})p(\mathbf{y}|\mathbf{x}, \beta) \\ &= \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y}). \end{aligned}$$

Using these lower bounds in (4.7), we can find an upper bound for the KL distance as follows:

$$(4.10) \quad \begin{aligned} C_{KL}(q(\alpha, \beta)q(\mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y})) \\ \leq \min_{\mathbf{u}} \int \int \int q(\alpha, \beta)q(\mathbf{x}) \log \left( \frac{q(\alpha, \beta)q(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}. \end{aligned}$$

Table 4.1. Proposed Algorithm

**Algorithm 3.** *Posterior parameter and image distributions estimation by approximating  $p(\alpha, \beta, \mathbf{x} | \mathbf{y})$  by  $q(\alpha, \beta)q(\mathbf{x})$ .*

*Given  $v^1 \in (R^4+)^N$  and  $q^1(\alpha, \beta)$ ,*

*For  $k = 1, 2, \dots$  until convergence:*

(1) *Find*

$$(4.11) \quad q^k(\mathbf{x}) = \operatorname{argmin}_{q(\mathbf{x})} \int \int \int q^k(\alpha, \beta) q(\mathbf{x}) \times \log \left( \frac{q^k(\alpha, \beta) q(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^k, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

(2) *Find*

$$(4.12) \quad \mathbf{u}^{k+1} = \operatorname{argmin}_{\mathbf{u}} \int \int \int q^k(\alpha, \beta) q^k(\mathbf{x}) \log \left( \frac{q^k(\alpha, \beta) q^k(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

(3) *Find*

$$(4.13) \quad q^{k+1}(\alpha, \beta) = \operatorname{argmin}_{q(\alpha, \beta)} \int \int \int q(\alpha, \beta) q^k(\mathbf{x}) \log \left( \frac{q(\alpha, \beta) q^k(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^{k+1}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x}$$

Finally, we employ a minimization of the right-hand side of (4.10) and obtain the following iterative procedure to estimate the unknowns:

Now we proceed to give the explicit solutions at each step of the algorithm. Note that in the first step we have

$$(4.14) \quad q^k(\mathbf{x}) \propto \exp \left\{ E_{q^k(\alpha, \beta)} [\ln \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^k)] \right\},$$

which corresponds to a multivariate Gaussian distribution with the mean and the covariance given by

$$(4.15) \quad E_{q^k(\mathbf{x})}[\mathbf{x}] = \operatorname{cov}_{q^k(\mathbf{x})}[\mathbf{x}] E_{q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{y},$$

$$(4.16) \quad \text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] = [\mathbf{E}_{\mathbf{q}^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} + p \mathbf{E}_{\mathbf{q}^k(\alpha)}[\alpha] \sum_{d=1}^4 (\Delta^d)^t W_d(\mathbf{u}^k)(\Delta^d)]^{-1} = [\mathbf{C}^k(\mathbf{u}^k)]^{-1},$$

where

$$W_d(\mathbf{u}^k) = \text{diag} \left( \frac{1}{u_{i,d}^{1-p/2}} \right), \quad d = 1, \dots, 4, \quad i = 1, \dots, N.$$

In the second step, we have

$$\mathbf{u}_d^{k+1} = \arg \min_{\mathbf{u}_d} \sum_i \frac{\mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^d(\mathbf{x}))^2] + \frac{2-p}{p} u_{i,d}}{u_{i,d}^{1-p/2}} \quad d = 1, \dots, 4$$

and therefore

$$(4.17) \quad \mathbf{u}_{i,d}^{k+1} = \mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^d(\mathbf{x}))^2], \quad i = 1, \dots, N \quad d = 1, \dots, 4$$

where

$$\mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^d(\mathbf{x}))^2] = (\Delta_i^d(\mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2 + \frac{1}{N} \text{trace} \left[ \text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] \times ((\Delta^d)^t (\Delta^d)) \right].$$

Finally to find  $\mathbf{q}^{k+1}(\alpha, \beta)$  we differentiate the integral on the right hand side of (4.13) with respect to  $\mathbf{q}(\alpha, \beta)$  and set it equal to zero to obtain

$$\mathbf{q}^{k+1}(\alpha, \beta) \propto \exp \left\{ \mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[\ln \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\alpha, \beta, \mathbf{x}, \mathbf{u}^{k+1})] \right\}.$$

Therefore,  $\mathbf{q}^{k+1}(\alpha)$  and  $\mathbf{q}^{k+1}(\beta)$  are both Gamma distributions, given by

$$\mathbf{q}^{k+1}(\alpha) \propto \alpha^{N/p + a_\alpha^o - 1} \exp \left[ -\alpha \left( \sum_i \sum_{d=1}^4 ([u_{i,d}^{k+1}]^{p/2}) + b_\alpha^o \right) \right],$$

$$q^{k+1}(\beta) \propto \beta^{N/2+a_\beta^o-1} \exp \left[ -\beta \left( \frac{E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2]}{2} + b_\beta^o \right) \right].$$

As the estimates to these hyperparameters, we use the means of these distributions, which can be given as

$$(4.18) \quad (E_{q^{k+1}(\alpha)}[\alpha])^{-1} = \gamma_\alpha \frac{1}{\bar{\alpha}^o} + (1 - \gamma_\alpha) \frac{p \sum_{d=1}^4 \sum_i [u_{i,d}^{k+1}]^{p/2}}{N},$$

$$(4.19) \quad (E_{q^{k+1}(\beta)}[\beta])^{-1} = \gamma_\beta \frac{1}{\bar{\beta}^o} + (1 - \gamma_\beta) \frac{E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2]}{N},$$

where  $\bar{\alpha}^o = a_\alpha^o/b_\alpha^o$ ,  $\bar{\beta}^o = a_\beta^o/b_\beta^o$ ,  $\gamma_\alpha = \frac{a_\alpha^o}{a_\alpha^o + \frac{N}{p}}$ , and  $\gamma_\beta = \frac{a_\beta^o}{a_\beta^o + \frac{N}{2}}$ . The parameters  $\gamma_\alpha$  and  $\gamma_\beta$ , both taking values in the interval  $[0, 1)$ , can be understood as normalized confidence parameters. According to (4.18) and (4.19), when they are equal to zero, no confidence is placed on the inverse of the mean of the corresponding hyperprior, while when they are asymptotically equal to one, the prior knowledge of the mean is fully enforced, i.e., no estimation of the hyperparameters is performed.

The only remaining task is the calculation of  $E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2]$  which can be given as

$$E_{q^k(x)} [\| \mathbf{y} - \mathbf{Hx} \|^2] = \| \mathbf{y} - \mathbf{H} E_{q^k(x)}[\mathbf{x}] \|^2 + \text{trace} \left( \text{cov}_{q^k(x)}[\mathbf{x}] \mathbf{H}' \mathbf{H} \right).$$

The estimate  $q^k(\mathbf{x})$  in Algorithm 3 is the best approximation to the posterior in the KL divergence sense. However, we can also consider a suboptimal case where we assume a degenerate distribution for  $q(\mathbf{x})$ , that is,  $q(\mathbf{x})$  takes one value,  $\mathbf{x}^k$ , with probability one and the rest of the values with probability zero. This approach leads to an alternative algorithm, referred to as

**Algorithm 4.**

, where the expectations involving the parameter  $q^k(\mathbf{x})$  are removed. Thus, the covariances in (4.17), (4.18), and (4.19) are set equal to zero.

As the estimate to the unknown image  $\mathbf{x}$ , we use the mean of  $q^k(\mathbf{x})$  shown in (4.15) in both algorithms, which requires the inversion of a very large matrix  $\mathbf{C}^k(\mathbf{u}^k)$ . This, however, introduces a big computational challenge since the last terms in (4.16) cannot be represented as block-circulant matrices with circulant blocks (BCCB), and therefore the inverse cannot be computed in the Fourier domain. We therefore employ a gradient descent approach to compute the image estimates without explicitly calculating the image covariance.

Note, however, that the explicit form of  $\text{cov}_{q^k(\mathbf{x})}[\mathbf{x}]$  is needed in (4.18)-(4.19) in Algorithm 3. To overcome this computational difficulty, we use the following approximation

$$\text{cov}_{q^k(\mathbf{x})}[\mathbf{x}] \approx [\mathbf{E}_{q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} + p \mathbf{E}_{q^k(\alpha)}[\alpha] \sum_{d=1}^4 z_d(\mathbf{u}^k) (\Delta^d)^t (\Delta^d)]^{-1} = \mathbf{B}^{-1},$$

where  $W_d(\mathbf{u}^k) \approx z_d(\mathbf{u}^k) \mathbf{I}$  and  $z_d(\mathbf{u}^k) = \frac{1}{N} \sum_i \frac{1}{[u_{i,d}^k]^{1-p/2}}$ . Note that in this approximation, matrix  $\mathbf{B}$  is BCCB, and therefore its inversion can be carried out very efficiently in the Fourier domain.

#### 4.4. Experimental Results

We performed a number of experiments with the proposed algorithms using several images and several types of blurring functions. The results of some of them are presented here. Since we developed two different algorithms resulting from our framework, we will present results for both of them.

For the experiments presented here, the ‘‘Lena’’ image (shown in Fig. 4.2(a)) is blurred with a Gaussian shaped blur with variance 9 and a 9x9 uniform blur. Gaussian noise is added to the blurred images to obtain degraded images with blurred-signal-to-noise (BSNR) ratios of 20 and

40dB. An example degraded image is shown in Fig. 4.2(b) where the blur is Gaussian-shaped with variance 9 and BSNR = 40dB.

The parameters of both algorithms are initialized as follows: The observed image is used as initial estimate for the unknown image  $\mathbf{x}$ . The initial values of the hyperparameters and  $\mathbf{u}$  are determined using this initial  $\mathbf{x}$  in (4.17)-(4.19). Note that all parameters of the algorithms are initialized using the observation  $\mathbf{y}$  so that no manual input is needed, i.e., both algorithms are initialized and run automatically. For all experiments, the criterion  $\| \mathbf{x}^k - \mathbf{x}^{k-1} \|^2 / \| \mathbf{x}^{k-1} \|^2 < 10^{-4}$  is used to terminate the iterative procedure, where  $\mathbf{x}^k$  is the mean of  $\mathbf{q}^k(\mathbf{x})$  in Algorithm 3 and the point estimate in Algorithm 4.

The restoration results of the Lena image in the case of Gaussian blur with 40dB BSNR are shown in Fig. 4.2(c) for Algorithm 3 and 4.2(d) for Algorithm 4. Note that Algorithm 3 is more successful at removing the blur whereas the restored image has less pronounced ringing artifacts in Algorithm 4. In both cases the restoration quality is good considering that the parameters of both algorithms are estimated using only the degraded observation without any prior knowledge about the noise. Also, in all cases the estimated value of  $\beta^{-1}$  was very close to the original noise variance.

Figure (4.3) shows the ISNR evolution in the case of Gaussian and uniform blurs with Algorithm 3 and BSNR = 40dB and 20dB with varying  $p$ -values, where ISNR is defined as  $10\log_{10}(\| \mathbf{x} - \mathbf{y} \|^2 / \| \mathbf{x} - \hat{\mathbf{x}} \|^2)$ , where  $\hat{\mathbf{x}}$  is the estimated image. We experimented with two cases, where in the first case we initialize and estimate the hyperparameters from the observation, and in the second case we compute them using the original unknown image and noise. As can be seen from Fig. (4.3), the highest ISNR values are achieved with different  $p$ -values for different noise levels and blur functions, and the ISNR values are comparable for both cases. It



Figure 4.2. (a) Original Lena Image, (b) Image degraded by a Gaussian shaped PSF with variance 9 and Gaussian noise of variance 0.16 (BSNR=40dB), (c) Restored image using Algorithm 3 with  $p = 1.8$  (ISNR = 4.15dB), (c) Restored image using Algorithm 4 with  $p = 1.6$  (ISNR = 3.78dB).

can also be seen that with fixed parameters the performance of the algorithms as a function of  $p$  is in agreement with the results reported in [37] and [146].

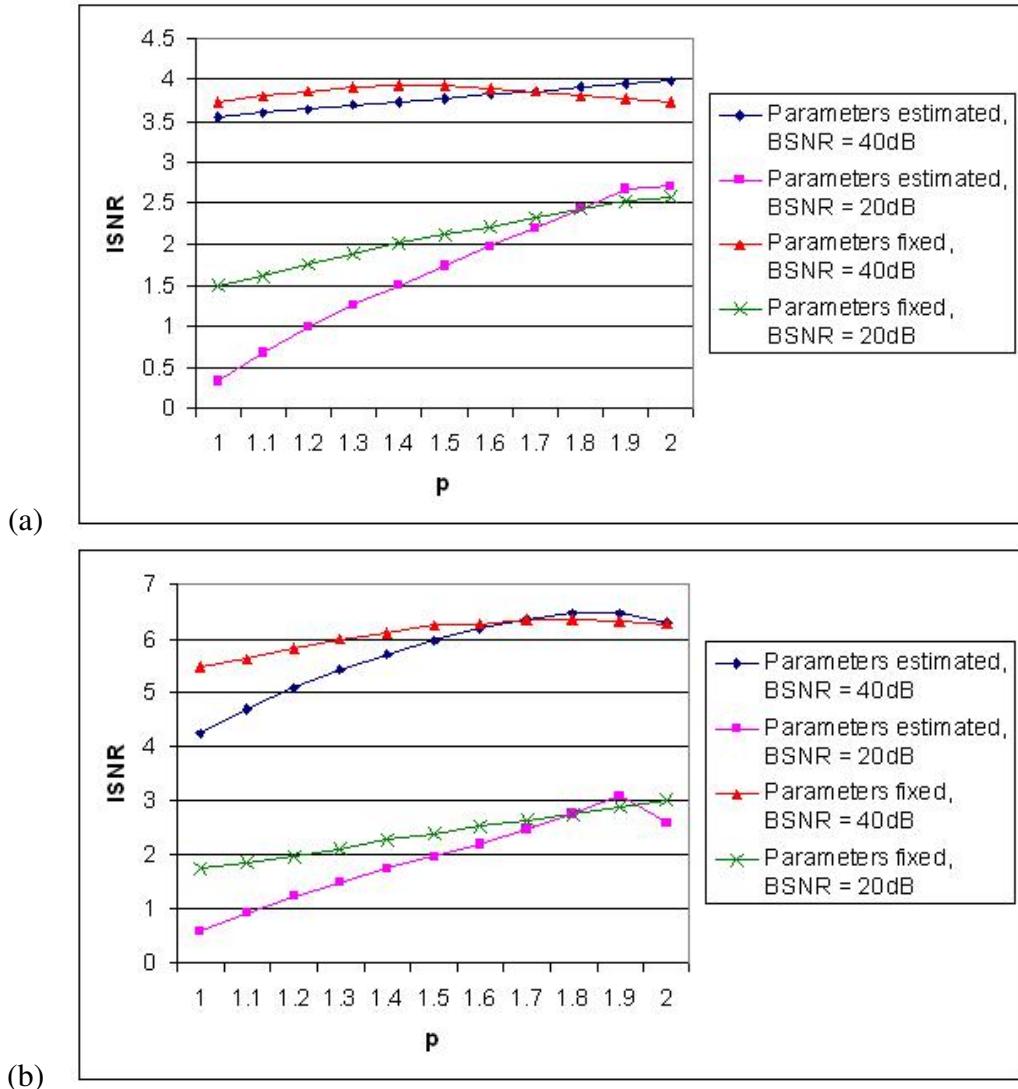


Figure 4.3. ISNR values obtained by different  $p$  values with Lena image degraded by (a) a Gaussian blur with variance 9 and (b) a 9x9 uniform blur with Gaussian noise (BSNR = 40dB and 20dB).

#### 4.5. Conclusions

A novel GGMRF based image restoration methodology has been presented that simultaneously estimates the reconstructed image and the hyperparameters of the Bayesian formulation.

We have adopted a variational approach to approximate the posterior distributions of the unknown parameters to estimate the posterior distributions of unknowns so that the uncertainty of the estimates can be evaluated and different values from these distributions can be used in the restoration process. Two algorithms are provided resulting from this approach. We have shown that the unknown parameters of the Bayesian formulation can be calculated automatically using only the observation or initial knowledge can be incorporated with different confidence values. Experimental results demonstrated the performance of the proposed algorithms.

## CHAPTER 5

### **Total Variation Blind Deconvolution Using A Variational Approach**

#### **5.1. Introduction**

Recently there has been an interest in applying variational methods to the blind deconvolution problem. These methods attempt to obtain approximations to the posterior distributions on the unknowns with the use of the Kullback-Leibner cross entropy [133]. Miskin and Mackay [163], Adami [4], Likas and Galatsanos [144], and Molina *et. al.* [169] employ this variational methodology to the blind deconvolution problem in a Bayesian formulation.

In chapter we present a blind deconvolution method that uses variational methods for the blind deconvolution problem by incorporating a Total Variation (TV) function as the image prior, and a SAR model for the blur prior. Although the TV model has been used in a regularization formulation in blind deconvolution before (see, for example, [58]), to the best of our knowledge, no work has been reported on the simultaneous estimation of the model parameters, image, and blur. Previous works attempted to solve for the unknown image and the blur, but the model parameters are manually selected [58, 57]. Moreover, we cast the TV-based blind deconvolution into a Bayesian estimation problem, which provides advantages in blind deconvolution such as means to estimate the uncertainties of the estimates. We develop two novel variational methods based on our hierarchical Bayesian formulation, and provide approximations to the posterior distributions of the image, blur, and model parameters rather than point estimates.

---

<sup>0</sup>This work has appeared in [12, 10]

## 5.2. Hierarchical Bayesian Modeling

We will utilize the standard formulation of the image degradation model, shown in (3.1). In blind deconvolution,  $\mathbf{H}$  represents the unknown block-circulant blurring matrix. As in the previous chapters, all unknown and observable parameters are treated as unknown stochastic quantities, and probability distributions are assigned based on subjective beliefs. The unknown parameters  $\mathbf{x}$  and  $\mathbf{h}$  are assigned *prior* distributions  $p(\mathbf{x}|\alpha_{im})$  and  $p(\mathbf{h}|\alpha_{bl})$ , which model the knowledge about the nature of the original image and the blur, respectively. The observation  $\mathbf{y}$  is also a random variable with the corresponding *conditional* distribution  $p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \beta)$ , also called the *likelihood* function. Clearly, these distributions depend on the model parameters  $\alpha_{im}$ ,  $\alpha_{bl}$  and  $\beta$ , which are called *hyperparameters*. The meaning of the hyperparameters will become clear when the prior distributions and the likelihood are defined below. In this chapter, we will denote the set of hyperparameters as  $\Omega = (\alpha_{im}, \alpha_{bl}, \beta)$ .

We again define the joint probability distribution function of all unknown and observed quantities, which can be written by Bayes' theorem as

$$(5.1) \quad p(\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, \mathbf{h}, \mathbf{y}) = p(\alpha_{im}, \alpha_{bl}, \beta)p(\mathbf{x}|\alpha_{im})p(\mathbf{h}|\alpha_{bl})p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \beta).$$

As in the image restoration problem, to alleviate the ill-posed nature of the blind deconvolution problem, prior knowledge about the degradation process, the unknown image and the blur can be incorporated through the use of the prior distributions and the likelihood function. We utilize a hierarchical Bayesian formulation by assuming that the hyperparameters are unknown. In the first step of this hierarchical model, the *a priori* probability distributions  $p(\mathbf{h}|\alpha_{bl})$  and

$p(\mathbf{x}|\alpha_{im})$  and the likelihood  $p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \beta)$  are formed that model the structure of the PSF, the original image and the noise, respectively. In the second stage, *hyperpriors* on the hyperparameters  $\beta$ ,  $\alpha_{im}$  and  $\alpha_{bl}$  are defined to model the prior knowledge of their values.

In the next subsections we first describe the prior models for the image and the PSF as well as the observation model we use in the first stage of the hierarchical Bayesian paradigm. We then proceed to explain the hyperprior distributions on the hyperparameters.

### 5.2.1. First stage: Prior models on the observation, PSF and the image

We assume that the degradation noise is Gaussian with zero mean and variance equal to  $\beta^{-1}$ , so that the likelihood function can be expressed as

$$(5.2) \quad p(\mathbf{y}|\mathbf{x}, h, \beta) \propto \beta^{N/2} \exp\left[-\frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2\right].$$

For the image prior we adopt the TV function, and utilize the same approximation in (3.13), given by

$$(5.3) \quad p(\mathbf{x}|\alpha_{im}) = c \alpha_{im}^{N/2} \exp[-\alpha_{im} \text{TV}(\mathbf{x})].$$

It is clear from (5.3) that the hyperparameter  $\alpha_{im}$  is the inverse variance of the unknown image  $\mathbf{x}$ .

We utilize the SAR model for the blur prior, that is,

$$(5.4) \quad p(\mathbf{h}|\alpha_{bl}) \propto \alpha_{bl}^{M/2} \exp\left\{-\frac{1}{2} \alpha_{bl} \|\mathbf{Ch}\|^2\right\},$$

where  $\mathbf{C}$  denotes the discrete Laplacian operator,  $\alpha_{\text{bl}}^{-1}$  is the variance of the Gaussian distribution, and  $M$  is the support of the blur, which is assumed to be the same as the image support. Note that in (5.4),  $M$  should in practice be replaced by  $M - 1$ , because  $\mathbf{C}^T \mathbf{C}$  is singular.

Finally, we assume that the degradation noise is Gaussian with zero mean and variance equal to  $\beta^{-1}$ , so that the likelihood function can be expressed as

$$(5.5) \quad p(\mathbf{y}|\mathbf{x}, h, \beta) \propto \beta^{N/2} \exp \left[ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{Hx}\|^2 \right].$$

### 5.2.2. Second stage: Hyperpriors on the hyperparameters

As in Chapter 3, we utilize Gamma priors for the hyperparameters  $\alpha_{im}$ ,  $\alpha_{bl}$  and  $\beta$ , since it is the conjugate prior for the inverse variance of the Gaussian distribution. The gamma distribution is defined in (3.16).

Finally, by combining the first and second stage of the hierarchical Bayesian model, the joint distribution in (5.1) can be defined. The dependencies in this joint probability model are shown in graphical form in Fig.(5.1) using a directed acyclic graph.

## 5.3. Bayesian Inference and Variational Approximation of the posterior distributions

We will denote the set of all unknowns by  $\Theta = (\Omega, \mathbf{x}, \mathbf{h}) = (\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, \mathbf{h})$ . As widely known, Bayesian inference should be based on the posterior distribution

$$p(\Theta | \mathbf{y}) = p(\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, \mathbf{h} | \mathbf{y}) = \frac{p(\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, \mathbf{h}, \mathbf{y})}{p(\mathbf{y})},$$

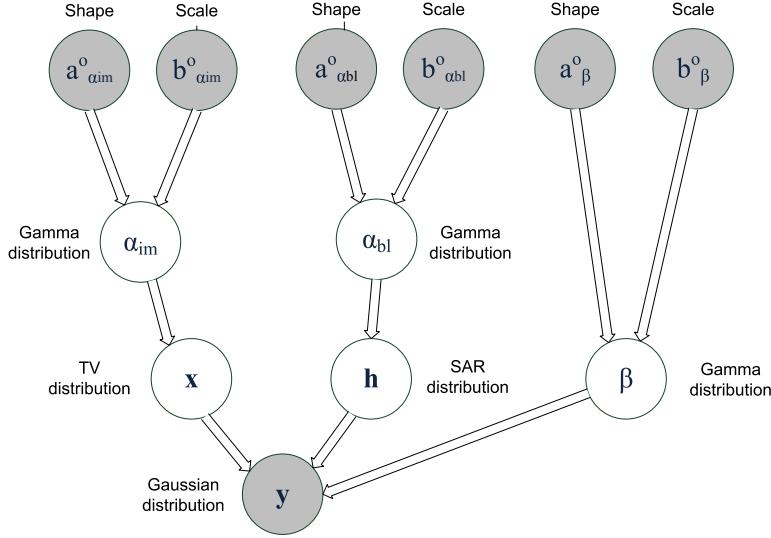


Figure 5.1. Graphical model showing relationships between variables.

where  $p(\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, \mathbf{h}, \mathbf{y})$  is given by (5.1). However, the posterior  $p(\Theta | \mathbf{y})$  is intractable, since

$$(5.6) \quad p(\mathbf{y}) = \int \int \int \int \int p(\alpha, \beta, \mathbf{x}, \mathbf{h}, \mathbf{y}) d\mathbf{h} d\mathbf{x} d\beta d\alpha_{bl} d\alpha_{im}$$

can not be calculated analytically. Therefore, we consider an approximation to it by a simpler tractable distribution  $q(\Theta)$  by following the variational methodology. The distribution  $q(\Theta)$  can be found by minimizing the Kullback-Leibler divergence, given by [134, 133]

$$(5.7) \quad C_{KL}(q(\Theta) \| p(\Theta|\mathbf{y})) = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta|\mathbf{y})} \right) d\Theta = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta, \mathbf{y})} \right) d\Theta + \text{const},$$

which is always nonnegative and equal to zero only when  $q(\Theta) = p(\Theta|\mathbf{y})$ , which corresponds to the expectation-maximization result. In order to obtain a tractable approximation, the family of distributions  $q(\Theta)$  are restricted utilizing the mean field approximation [187] so that  $q(\Theta) =$

$q(\Omega)q(\mathbf{x})q(\mathbf{h})$ , where  $q(\Omega) = q(\alpha_{im})q(\alpha_{bl})q(\beta)$ . Therefore, the latent variables are assumed to be statistically independent *a priori*. Note that no particular functional forms are placed on the distributions  $q(\alpha_{im})$ ,  $q(\alpha_{bl})$  and  $q(\beta)$ .

However, the use of the TV prior makes the calculation of the KL distance difficult to evaluate even with this factorization, as explained in Chapter 3. Therefore, we utilize the same majorization of the TV prior to find an upper bound of the KL divergence. We define the following functional  $\mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{u})$ , for  $\alpha_{im}$ ,  $\mathbf{x}$ , and an  $N$ -dimensional vector  $\mathbf{u} \in (R^+)^N$

$$(5.8) \quad \mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{u}) = \alpha_{im}^{N/2} \exp \left[ -\frac{\alpha_{im}}{2} \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + u_i}{\sqrt{u_i}} \right].$$

Following the same steps in Chapter 3, we obtain the following lower bound for the joint probability distribution

$$(5.9) \quad \begin{aligned} p(\Theta, \mathbf{y}) &\geq c \cdot p(\Omega) \mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{u}) p(\mathbf{h}|\alpha_{bl}) p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \beta) \\ &= \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{u}, \mathbf{y}). \end{aligned}$$

For  $\theta \in \{\alpha_{im}, \alpha_{bl}, \beta, \mathbf{x}, h\}$  let us denote by  $\Theta_\theta$  the subset of  $\Theta$  with  $\theta$  removed; for instance, if  $\theta = \mathbf{x}$ ,  $\Theta_\mathbf{x} = (\alpha_{im}, \alpha_{bl}, \beta, h)$ . Then, utilizing the lower bound  $\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{u}, \mathbf{y})$  for the joint probability distribution in (5.7) we obtain an upper bound for the KL divergence as follows

$$(5.10) \quad \begin{aligned} C_{KL}(q(\Theta) \| p(\Theta|\mathbf{y})) &\leq C_{KL}(q(\Theta) \| \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{u}, \mathbf{y})) \\ &= \int q(\theta) \left( \int q(\Theta_\theta) \log \left( \frac{q(\theta)q(\Theta_\theta)}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{u}, \mathbf{y})} \right) d\Theta_\theta \right) d\theta. \end{aligned}$$

Therefore, one can minimize this upper bound instead of minimizing the KL divergence. Note that the form of the inequality in (5.10) suggests an alternating (cyclic) optimization strategy where the algorithm cycles through the unknown distributions and replaces each with a revised estimate given by the minimum of (5.10) with other distributions held constant. Thus, given  $q(\Theta_\theta)$ , the posterior  $q(\theta)$  can be computed by solving

$$(5.11) \quad q(\theta) = \arg \min_{q(\theta)} C_{KL}(q(\Theta_\theta)q(\theta) \parallel F(\Theta, w, y_1, y_2)(\Theta, u, y)).$$

Differentiation of the integral on the right hand side in (5.10) with respect to  $q(\theta)$  results in (see Eq. (2.28) in [162]),

$$(5.12) \quad \hat{q}(\theta) = \text{const} \times \exp \left( E[ \log F(\Theta, w, y_1, y_2)(\Theta, u, y) ]_{q(\Theta_\theta)} \right),$$

where

$$E[ \log F(\Theta, w, y_1, y_2)(\Theta, u, y) ]_{q(\Theta_\theta)} = \int \log F(\Theta, w, y_1, y_2)(\Theta, u, y) q(\Theta_\theta) d\Theta_\theta.$$

We obtain the iterative procedure shown in Table (5.1) to find  $q(\Theta)$  by applying this minimization to each unknown in an alternating way.

Now we proceed to state the solutions at each step of the algorithm ((5.13)-(5.16)) explicitly. In estimating  $q(x)$  and  $q(h)$  we assume that the hyperparameters  $\Omega$  are known. From (5.12) it is clear that  $q^k(x)$  is an  $N$ -dimensional Gaussian distribution, rewritten as,

$$q^k(x) = \mathcal{N} \left( x \mid E_{q^k(x)}[x], \text{cov}_{q^k(x)}(x) \right).$$

Table 5.1. Proposed Algorithm I

**Algorithm 5.** Given  $q^1(\mathbf{h})$ ,  $q^1(\alpha_{\text{im}})$ ,  $q^1(\alpha_{\text{bl}})$ , and  $q^1(\beta)$ , the initial estimates of the distributions  $q(\mathbf{h})$ ,  $q(\alpha_{\text{im}})$ ,  $q(\alpha_{\text{bl}})$  and  $q(\beta)$ , for  $k = 1, 2, \dots$  until a stopping criterion is met:

(1) Find

$$(5.13) \quad q^k(\mathbf{x}) = \operatorname{argmin}_{q(\mathbf{x})} \int \int q^k(\Theta_{\mathbf{x}}) q(\mathbf{x}) \times \log \left( \frac{q^k(\Theta_{\mathbf{x}}) q(\mathbf{x})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta_{\mathbf{x}}^k, \mathbf{x}, \mathbf{u}^k, \mathbf{y})} \right) d\Theta_{\mathbf{x}} d\mathbf{x}$$

(2) Find

$$(5.14) \quad q^{k+1}(\mathbf{h}) = \operatorname{argmin}_{q(\mathbf{h})} \int \int q^k(\Theta_{\mathbf{h}}) q(\mathbf{h}) \times \log \left( \frac{q^k(\Theta_{\mathbf{h}}) q(\mathbf{h})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta_{\mathbf{h}}^k, \mathbf{h}, \mathbf{u}^k, \mathbf{y})} \right) d\Theta_{\mathbf{h}} d\mathbf{h}$$

(3) Find

$$(5.15) \quad \mathbf{u}^{k+1} = \operatorname{argmin}_{\mathbf{u}} \int q^k(\Theta_{\mathbf{h}}) q^{k+1}(\mathbf{h}) \times \log \left( \frac{q^k(\Theta_{\mathbf{h}}) q^{k+1}(\mathbf{h})}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta_{\mathbf{h}}^k, \mathbf{h}^{k+1}, \mathbf{u}, \mathbf{y})} \right) d\Theta$$

(4) Find

$$(5.16) \quad q^{k+1}(\Omega) = \operatorname{argmin}_{q(\Omega)} \int \int q^k(\Theta_{\Omega}) q(\Omega) \times \log \left( \frac{q^k(\Theta_{\Omega}) q(\Omega)}{\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta_{\Omega}^k, \Omega, \mathbf{u}^k, \mathbf{y})} \right) d\Theta_{\Omega} d\Omega$$

The covariance and mean of this normal distribution can be calculated from (5.13) as

(5.17)

$$\operatorname{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] = (\beta^k \mathbf{E}^k(\mathbf{H})^t \mathbf{E}^k(\mathbf{H}) + \alpha_{\text{im}}^k (\Delta^h)^t W(\mathbf{u}^k) (\Delta^h) + \alpha_{\text{im}}^k (\Delta^v)^t W(\mathbf{u}^k) (\Delta^v) + N \beta^k \operatorname{cov}_{\mathbf{q}^k(\mathbf{h})}[\mathbf{h}])^{-1},$$

$$(5.18) \quad \mathbf{E}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] = \operatorname{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] \beta^k \mathbf{E}^k(\mathbf{H})^t \mathbf{y},$$

where the matrix  $W(\mathbf{u})$  is the same as in (3.34). Clearly, the system in (5.18) can not be solved analytically because of the high dimensionality of the matrices. We again developed a gradient descent algorithm to solve to tackle this problem, similar to Chapter 3.

Similarly to  $q^k(\mathbf{x})$ ,  $q^k(\mathbf{h})$  is an  $M$ -dimensional Gaussian distribution, given by

$$(5.19) \quad q^{k+1}(\mathbf{h}) = \mathcal{N} \left( \mathbf{h} \mid E_{q^{k+1}(\mathbf{h})}[\mathbf{h}], \text{cov}_{q^{k+1}(\mathbf{h})}[\mathbf{h}] \right),$$

with

$$(5.20) \quad \text{cov}_{q^{k+1}(\mathbf{h})}[\mathbf{h}] = (\alpha_{\text{bl}}^k \mathbf{C}^t \mathbf{C} + \beta^k E_{q^k(\mathbf{x})}[\mathbf{x}]^t E_{q^k(\mathbf{x})}[\mathbf{x}] + N \beta^k \text{cov}_{q^k(\mathbf{x})}[\mathbf{x}])^{-1},$$

and

$$(5.21) \quad E_{q^{k+1}(\mathbf{h})}[\mathbf{h}] = \text{cov}_{q^{k+1}(\mathbf{h})}[\mathbf{h}] \beta^k E_{q^k(\mathbf{x})}[\mathbf{x}]^t \mathbf{y},$$

We will denote  $E_{q^k(\mathbf{x})}[\mathbf{x}]$  by  $E^k(\mathbf{x})$ ,  $\text{cov}_{q^k(\mathbf{x})}[\mathbf{x}]$  by  $\text{cov}^k(\mathbf{x})$ ,  $E_{q^k(\mathbf{h})}[\mathbf{h}]$  by  $E^k(\mathbf{h})$  and  $\text{cov}_{q^k(\mathbf{h})}[\mathbf{h}]$  by  $\text{cov}^k(\mathbf{h})$  for clarity. It is worth emphasizing here that we did not assume *a priori* that  $q^k(\mathbf{x})$  and  $q^k(\mathbf{h})$  are Gaussian distributions. This result is derived due to the minimization of the KL divergence with respect to all possible distributions according to the factorization  $q(\Theta) = q(\alpha_{\text{im}})q(\alpha_{\text{bl}})q(\beta)q(\mathbf{x})q(\mathbf{h})$ . Note also that the image estimate in (5.18) is very similar to the image estimated proposed in [58] in a regularization framework except the uncertainty term  $N \beta^k \text{cov}_{q^k(\mathbf{h})}[\mathbf{h}]$ . As we will see in experimental results, this formulation will provide improvements in the restoration.

In the step 4 of the algorithm, we find  $\mathbf{u}^{k+1}$  from (5.15), given by

$$(5.22) \quad \mathbf{u}^{k+1} = \arg \min_{\mathbf{u}} \sum_i \frac{E_{q^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2] + u_i}{\sqrt{u_i}}.$$

Therefore,  $\mathbf{u}^{k+1}$  can be obtained as

$$(5.23) \quad \mathbf{u}_i^{k+1} = E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2], \quad i = 1, \dots, N.$$

After finding estimates of the posterior distributions of the image and blur, we find the estimates for the hyperpriors at the last step of the algorithm. For  $\omega \in \{\alpha_{im}, \alpha_{bl}, \beta\}$ , evaluating (5.16) using (5.12) results in

$$q^{k+1}(\omega) \propto \exp E_{\mathbf{q}^k(\mathbf{x})\mathbf{q}^{k+1}(\mathbf{h})\mathbf{q}(\Omega_\omega)}[\log \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Omega_\omega^k, \omega, \mathbf{x}^k, \mathbf{h}^{k+1}, \mathbf{u}^{k+1}, \mathbf{y})].$$

Evaluating this explicitly we obtain

$$\begin{aligned} E_{\mathbf{q}^k(\mathbf{x})\mathbf{q}^{k+1}(\mathbf{h})}[\log \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta)] &= \text{const} + \sum_{\omega \in \{\alpha_{im}, \alpha_{bl}, \beta\}} ((a_\omega^o - 1) \log \omega - \omega / b_\omega^o) \\ &+ \frac{N}{2} \log \alpha_{im} + \frac{M}{2} \log \alpha_{bl} + \frac{N}{2} \log \beta \\ &- \frac{1}{2} \alpha_{im} E_{\mathbf{q}^k(\mathbf{x})} \left[ \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + u_i}{\sqrt{u_i}} \right] \\ &- \frac{1}{2} \alpha_{bl} E_{\mathbf{q}^{k+1}(\mathbf{h})} [\|\mathbf{Ch}\|^2] - \frac{1}{2} \beta E_{\mathbf{q}^k(\mathbf{x})\mathbf{q}^{k+1}(\mathbf{h})} [\|\mathbf{y} - \mathbf{Hx}\|^2], \end{aligned} \quad (5.24)$$

where

$$(5.25) \quad E_{\mathbf{q}^k(\mathbf{x})} \left[ \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + u_i}{\sqrt{u_i}} \right] = 2 \sum_i \sqrt{u_i^{k+1}},$$

$$(5.26) \quad E_{\mathbf{q}^{k+1}(\mathbf{h})} [\|\mathbf{Ch}\|^2] = \|\mathbf{CE}^{k+1}(\mathbf{h})\|^2 + \text{trace}(\mathbf{C}^t \mathbf{C} \text{cov}^k(\mathbf{h})),$$

and

$$\begin{aligned}
 E_{q^k(\mathbf{x})q^{k+1}(\mathbf{h})} [\|\mathbf{y} - \mathbf{Hx}\|^2] &= \|\mathbf{y} - E^{k+1}(\mathbf{h})E^k(\mathbf{x})\|^2 + \text{trace}(N \text{cov}^k(\mathbf{x}) \text{cov}^{k+1}(\mathbf{h})) \\
 &\quad + \text{trace}(E^k(\mathbf{x})^t E^k(\mathbf{x}) \text{cov}^{k+1}(\mathbf{h})) \\
 (5.27) \quad &\quad + \text{trace}(E^{k+1}(\mathbf{H})^t E^{k+1}(\mathbf{H}) \text{cov}^k(\mathbf{x})).
 \end{aligned}$$

The details of the calculation of  $\text{cov}_{q^k(\mathbf{x})}[\mathbf{x}]$ ,  $\text{cov}_{q^k(\mathbf{h})}[\mathbf{h}]$ ,  $E_{q^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2]$ ,  $E_{q^{k+1}(\mathbf{h})}[\|\mathbf{Ch}\|^2]$  and  $E_{q^k(\mathbf{x})q^{k+1}(\mathbf{h})}[\|\mathbf{y} - \mathbf{Hx}\|^2]$  are shown in Appendix C.

It can be seen from (5.24) that all hyperparameters have Gamma distributions, given by

$$(5.28) \quad q^{k+1}(\alpha_{im}) \propto \alpha_{im}^{N/2+a_{\alpha_{im}}^o-1} \exp \left[ -\alpha_{im}(1/b_{\alpha_{im}}^o + \sum_i \sqrt{u_i^{k+1}}) \right],$$

$$(5.29) \quad q^{k+1}(\alpha_{bl}) \propto \alpha_{bl}^{M/2+a_{\alpha_{bl}}^o-1} \exp \left[ -\alpha_{bl}(1/b_{\alpha_{bl}}^o + \frac{E_{q^{k+1}(\mathbf{h})}[\|\mathbf{Ch}\|^2]}{2}) \right],$$

$$(5.30) \quad q^{k+1}(\beta) \propto \beta^{N/2+a_{\beta}^o-1} \exp \left[ -\beta(1/b_{\beta}^o + \frac{E_{q^k(\mathbf{x})q^{k+1}(\mathbf{h})}[\|\mathbf{y} - \mathbf{Hx}\|^2]}{2}) \right],$$

The means of these gamma distributions can be found using (3.16) and be represented as follows

$$(5.31) \quad (\alpha_{im}^k)^{-1} = (E_{q^{k+1}(\alpha_{im})}[\alpha_{im}])^{-1} = \gamma_{\alpha_{im}} \frac{1}{\bar{\alpha}_{im}^o} + (1 - \gamma_{\alpha_{im}}) \frac{\sum_i \sqrt{u_i^{k+1}}}{N/2},$$

$$(5.32) \quad (\alpha_{bl}^k)^{-1} = (E_{q^{k+1}(\alpha_{bl})}[\alpha_{bl}])^{-1} = \gamma_{\alpha_{bl}} \frac{1}{\bar{\alpha}_{bl}^o} + (1 - \gamma_{\alpha_{bl}}) \frac{E_{q^{k+1}(\mathbf{h})} [\|\mathbf{Ch}\|^2]}{M},$$

$$(5.33) \quad (\beta^k)^{-1} = (E_{q^{k+1}(\beta)}[\beta])^{-1} = \gamma_{\beta} \frac{1}{\bar{\beta}^o} + (1 - \gamma_{\beta}) \frac{E_{q^k(\mathbf{x})q^{k+1}(\mathbf{h})} [\|\mathbf{y} - \mathbf{Hx}\|^2]}{N},$$

where  $\bar{\alpha}_{im}^o = a_{\alpha_{im}}^o / b_{\alpha_{im}}^o$ ,  $\bar{\alpha}_{bl} = a_{\alpha_{bl}}^o / b_{\alpha_{bl}}^o$  and  $\bar{\beta}^o = a_{\beta}^o / b_{\beta}^o$  and

$$(5.34) \quad \gamma_{\alpha_{im}} = \frac{a_{\alpha_{im}}^o}{a_{\alpha_{im}}^o + \frac{N}{2}}, \quad \gamma_{\alpha_{bl}} = \frac{a_{\alpha_{bl}}^o}{a_{\alpha_{bl}}^o + \frac{M}{2}}, \quad \gamma_{\beta} = \frac{a_{\beta}^o}{a_{\beta}^o + \frac{N}{2}}.$$

As in Chapter 3, the parameters  $\gamma_{\alpha_{im}}$ ,  $\gamma_{\alpha_{bl}}$ , and  $\gamma_{\beta}$  can be understood as normalized confidence parameters, as can be seen from (5.31)-(5.33) and they take values in the interval  $(0, 1)$ .

In Algorithm 5 no assumptions were imposed on the posterior approximations  $q(\mathbf{x})$  and  $q(\mathbf{h})$ . We can, however, assume that these distributions are *degenerate*, i.e., distributions which take one value with probability one and the rest of the values with probability zero. We can obtain another algorithm under this assumption which is similar to algorithm 5. In this second algorithm, the value of the KL divergence is again decreased at each update step, but not by the maximum possible amount as was the case in algorithm 5.

Utilizing the fact that the distributions on  $\mathbf{x}$  and  $\mathbf{h}$  are degenerate, that is,

$$(5.35) \quad q(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} = \underline{\mathbf{x}} \\ 0 & \text{otherwise} \end{cases}$$

$$(5.36) \quad q(\mathbf{h}) = \begin{cases} 1 & \text{if } \mathbf{h} = \underline{\mathbf{h}} \\ 0 & \text{otherwise} \end{cases}$$

we obtain the following algorithm 6, where we use  $\mathbf{x}^k$  and  $\mathbf{h}^k$  to denote the values  $q^k(\mathbf{x})$  and  $q^k(\mathbf{h})$  take with probability one, respectively.

The update equations for the distribution of the hyperparameters are then obtained from (5.31)-(5.33) as follows:

$$(5.45) \quad (E_{q^{k+1}(\alpha_{im})}[\alpha_{im}])^{-1} = \gamma_{\alpha_{im}} \frac{1}{\bar{\alpha}_{im}^o} + (1 - \gamma_{\alpha_{im}}) \frac{\sum_i \sqrt{u_i^{k+1}}}{N/2},$$

$$(5.46) \quad (E_{q^{k+1}(\alpha_{bl})}[\alpha_{bl}])^{-1} = \gamma_{\alpha_{bl}} \frac{1}{\bar{\alpha}_{bl}^o} + (1 - \gamma_{\alpha_{bl}}) \frac{\|\mathbf{Ch}\|^2}{M},$$

$$(5.47) \quad (E_{q^{k+1}(\beta)}[\beta])^{-1} = \gamma_{\beta} \frac{1}{\bar{\beta}^o} + (1 - \gamma_{\beta}) \frac{\|\mathbf{y} - \mathbf{Hx}\|^2}{N}.$$

## 5.4. Experimental Results

A number of experiments have been performed with the proposed methods. We will denote Algorithm 5 as *TV1*, and Algorithm 6, where the distributions  $q(\mathbf{x})$  and  $q(\mathbf{h})$  are both degenerate, as *TV2*. We present both synthetic and real deconvolution examples to demonstrate the performance of the algorithms.

For the first set of our experiments, the images “Lena”, “Cameraman” and the “Shepp-Logan” phantom are blurred with a Gaussian-shaped function with variance 9, and white Gaussian noise is added to obtain degraded images with blurred-signal-to-noise ratios (BSNR) of 20dB and 40dB. The degraded “Lena” images as well as the original are shown in Fig. (3.2). The degraded “Shepp-Logan” phantom as well as the original are shown in Fig. (5.2). We

Table 5.2. Proposed Algorithm II

**Algorithm 6.** Given  $q^1(\mathbf{h})$ ,  $q^1(\alpha_{\text{im}})$ ,  $q^1(\alpha_{\text{bl}})$ , and  $q^1(\beta)$  the initial estimates of the distributions  $q(\mathbf{h})$ ,  $q(\alpha_{\text{im}})$ ,  $q(\alpha_{\text{bl}})$  and  $q(\beta)$ ,  
for  $k = 1, 2, \dots$  until a stopping criterion is met:

(1) Calculate

$$(5.37) \quad \begin{aligned} \mathbf{x}^k &= \left( \mathbb{E}_{q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} + \mathbb{E}_{q^k(\alpha_{\text{im}})}[\alpha_{\text{im}}] (\Delta^h)^t W(\mathbf{u}^k) (\Delta^h) + \mathbb{E}_{q^k(\alpha_{\text{im}})}[\alpha_{\text{im}}] (\Delta^v)^t W(\mathbf{u}^k) (\Delta^v) \right)^{-1} \\ &\quad \mathbb{E}_{q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{y} \end{aligned}$$

(2) Calculate

$$(5.38) \quad \mathbf{h}^k = \left( \mathbb{E}_{q^k(\alpha_{\text{bl}})}[\alpha_{\text{bl}}] \mathbf{C}^t \mathbf{C} + \mathbb{E}_{q^k(\beta)}[\beta] \mathbf{x}^t \mathbf{x} \right)^{-1} \mathbb{E}_{q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{x}$$

(3) Calculate

$$(5.39) \quad \mathbf{u}_i^{k+1} = (\Delta_i^h(\mathbf{x}^k))^2 + (\Delta_i^v(\mathbf{x}^k))^2, \quad i = 1, \dots, N.$$

(4) Calculate

$$(5.40) \quad q^{k+1}(\alpha_{\text{im}}, \alpha_{\text{bl}}, \beta) = q^{k+1}(\alpha_{\text{im}}) q^{k+1}(\alpha_{\text{bl}}) q^{k+1}(\beta),$$

where  $q^{k+1}(\alpha_{\text{im}})$ ,  $q^{k+1}(\alpha_{\text{bl}})$  and  $q^{k+1}(\beta)$  are gamma distributions given respectively by

$$(5.41) \quad q^{k+1}(\alpha_{\text{im}}) \propto \alpha_{\text{im}}^{N/2+a_{\text{im}}^o-1} \exp \left[ -\alpha_{\text{im}}(1/b_{\alpha_{\text{im}}}^o + \sum_i \sqrt{u_i^{k+1}}) \right],$$

$$(5.42) \quad q^{k+1}(\alpha_{\text{bl}}) \propto \alpha_{\text{bl}}^{M/2+a_{\text{bl}}^o-1} \exp \left[ -\alpha_{\text{bl}}(1/b_{\alpha_{\text{bl}}}^o + \frac{\|\mathbf{Ch}\|^2}{2}) \right],$$

$$(5.43) \quad q^{k+1}(\beta) \propto \beta^{N/2+a_{\beta}^o-1} \exp \left[ -\beta(1/b_{\beta}^o + \frac{\|\mathbf{y} - \mathbf{Hx}\|^2}{2}) \right].$$

Set

$$(5.44) \quad q(\alpha_{\text{im}}, \alpha_{\text{bl}}, \beta) = \lim_{k \rightarrow \infty} q^k(\alpha_{\text{im}}, \alpha_{\text{bl}}, \beta), \quad \hat{\mathbf{x}} = \lim_{k \rightarrow \infty} \mathbf{x}^k, \quad \hat{\mathbf{h}} = \lim_{k \rightarrow \infty} \mathbf{h}^k.$$

compare our algorithms to two other blind deconvolution algorithms based on variational approximations, which use SAR models for both the image and the blur (see [169] for details).

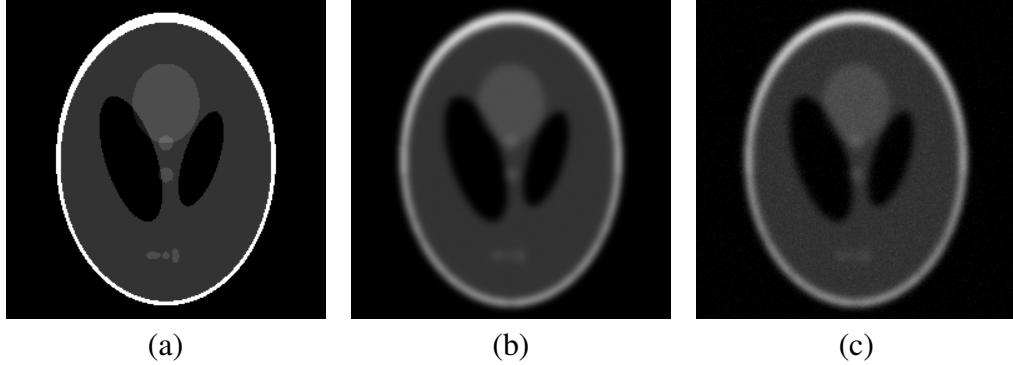


Figure 5.2. (a) Shepp-Logan phantom image; degraded with a Gaussian shaped PSF with variance 9 and Gaussian noise of variance: (b) 0.18 (BSNR = 40 dB), (c) 18 (BSNR = 20 dB).

We also include the results from the non-blind versions of our algorithms, presented in Chapter 3, where the blur function is assumed known and only the image and the hyperparameters are estimated during iterations. These algorithms will be denoted as *TV1-NB* and *TV2-NB*.

The initial values for the *TV1* and *TV2* algorithms are chosen as follows: The observed image is used as the initial estimate of  $\mathbf{x}^1$ , and as for the initial estimate of  $\mathbf{h}^1$  we chose a Gaussian function with variance 4. The covariance matrices  $\text{cov}^1(\mathbf{h})$  and  $\text{cov}^1(\mathbf{x})$  are set equal to zero. The initial values  $\beta^1$ ,  $\alpha_{\text{im}}^1$ , and  $\alpha_{\text{bl}}^1$  are calculated according to (5.31)–(5.33), assuming degenerate distributions. It should be emphasized that except the initial value of the blur, all parameters are automatically estimated from the observed image. For the *SAR1* and *SAR2* algorithms, the same initial blur is used, and other parameters are found also automatically from the observed image [169].

For this set of experiments, we set all confidence parameters equal to zero, i.e., the observation is made fully responsible for the estimation process. The quantitative results are shown in Table 5.3, where ISNR is defined as  $10 \log_{10}(\|x - y\|^2 / \|x - \hat{x}\|^2)$ , where  $x$ ,  $y$  and  $\hat{x}$  are the original, observed, and estimated images, respectively. The corresponding restoration results

Table 5.3. ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 9.

BSNR	Method	Lena		Cameraman		Shepp-Logan	
		ISNR (dB)	iterations	ISNR (dB)	iterations	ISNR (dB)	iterations
40dB	<i>TV1</i>	2.53	85	1.82	92	3.07	200
	<i>TV2</i>	2.95	200	1.73	200	3.36	200
	<i>SAR1-CAR</i>	2.10	116	1.54	137	1.64	192
	<i>SAR1-SAR</i>	1.35	63	1.03	66	1.20	121
	<i>SAR2-CAR</i>	2.42	200	1.74	200	1.81	200
	<i>SAR2-SAR</i>	1.43	78	1.01	89	1.35	180
	<i>TV1-NB</i>	4.33	9	2.96	11	4.16	28
	<i>TV2-NB</i>	4.31	9	2.95	11	4.15	28
20dB	<i>TV1</i>	2.62	81	1.70	5	2.47	8
	<i>TV2</i>	-2.54	7	-3.39	7	-1.41	10
	<i>SAR1-CAR</i>	1.06	200	0.85	200	1.56	200
	<i>SAR1-SAR</i>	1.62	80	1.16	98	1.53	146
	<i>SAR2-CAR</i>	-0.29	9	-0.27	6	-0.15	11
	<i>SAR2-SAR</i>	-0.60	5	-0.71	5	-0.52	6
	<i>TV1-NB</i>	3.31	11	2.42	12	4.28	17
	<i>TV2-NB</i>	3.29	11	2.41	12	4.27	17

for the “Lena” image are shown in Fig. (5.3) for the 40dB BSNR case, and in Fig. (5.3) for the 20dB BSNR case. We also show the restoration results for the Shepp-Logan phantom in Fig. (5.7) for the 40dB BSNR case, and in Fig. (5.8) for the 20dB BSNR case.

A few remarks can be made by examining the ISNR values in Table 5.3 and the restorations visually. First, note that the nonblind algorithms *TV1-NB* and *TV2-NB* result in higher ISNR values than the blind algorithms as expected. Although the nonblind algorithms clearly outperform the blindly restored images in terms if ISNR, the blind algorithms result in visually comparable results even with these severely blurred observations in both noise levels. Second, algorithms *TV1* and *TV2* result in higher ISNR values in all images and noise levels than SAR-based algorithms. Visually, the TV-based algorithms result in smoother restorations with



Figure 5.3. Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.53 dB), (b) *TV2* algorithm (ISNR = 2.95 dB), (c) *SAR1* algorithm (ISNR = 2.10 dB), (d) *SAR2* algorithm (ISNR = 2.42 dB), (e) *TV1-NB* algorithm (ISNR = 4.33 dB), and (f) *TV2-NB* algorithm (ISNR = 4.31 dB).

Table 5.4. ISNR values, and number of iterations for the Lena, Cameraman and Shepp-Logan images degraded by a Gaussian blur with variance 5.

BSNR	Method	Lena		Cameraman		Shepp-Logan	
		ISNR (dB)	iterations	ISNR (dB)	iterations	ISNR (dB)	iterations
40dB	<i>TV1</i>	3.19	200	1.66	73	2.05	137
	<i>TV2</i>	3.29	115	2.49	58	3.79	200
	<i>SAR1-CAR</i>	2.35	165	1.70	184	1.91	200
	<i>SAR1-SAR</i>	1.26	53	0.90	53	1.24	157
	<i>SAR2-CAR</i>	2.57	182	1.89	200	1.30	108
	<i>SAR2-SAR</i>	1.45	77	0.99	86	1.51	200
	<i>TV1-NB</i>	4.98	10	3.50	12	7.57	43
	<i>TV2-NB</i>	4.93	10	3.48	12	7.29	39
20dB	<i>TV1</i>	1.39	189	1.43	136	2.09	200
	<i>TV2</i>	-4.39	7	-5.19	7	-2.89	9
	<i>SAR1-CAR</i>	0.36	200	0.55	200	1.46	200
	<i>SAR1-SAR</i>	1.14	87	0.87	76	1.24	200
	<i>SAR2-CAR</i>	-0.23	7	-0.31	5	-0.17	9
	<i>SAR2-SAR</i>	-0.81	5	-0.91	5	-0.72	6
	<i>TV1-NB</i>	2.92	10	2.40	12	4.68	16
	<i>TV2-NB</i>	2.83	11	2.37	12	4.65	16

reduced ringing artifacts. Especially in high noise conditions, i.e., 20dB BSNR, algorithms *TV1* and *TV2* give acceptable restorations whereas SAR algorithms result in very noisy restorations or flat images [169].

The difference between the TV-based and SAR-based algorithms are clearer in the restoration of the Shepp-Logan phantom which consists of few but strong edges. The algorithms *TV1* and *TV2* clearly outperform the SAR algorithms in terms of preserving and recovering the edges, whereas the ringing artifacts are more visible. Note also that even in the case of 20dB BSNR, where the noise variance is very high (15.6), the restoration result of *TV1* is quite acceptable.

We note here that the proposed algorithms are relatively robust to the initial selected value of the blur. When a Gaussian with variance 2 is chosen as  $\mathbf{h}^1$ , the ISNR values are 1.80dB

for  $TV1$  and 2.75dB for  $TV2$  for 40dB BSNR, and 1.71 for  $TV1$  and -3.36dB for  $TV2$  for 20dB BSNR, similar to Table (5.3).

One dimensional slices through the origins of the estimated blurs for all algorithms corresponding to the first column of Table 5.3 are shown in Fig (5.11) for the 40dB BSNR case, and in Fig. (5.12) for the 20dB BSNR case. It is clear that all algorithms provide close estimates to the true PSF in the 40dB BSNR case, whereas the estimates are not as good in the case of high noise (20dB BSNR) observation. An important remark is that all methods, except  $TV1$ , provide PSF estimates with a larger support than the true PSF, which explains the increased ringing artifacts in the restoration results.

In the second set of experiments, we tested the algorithms with a less severe blur. The images are blurred with a Gaussian shaped PSF with variance 5, and the initial value for the blur,  $\mathbf{h}^1$  is set to 2. The corresponding ISNR values of the restorations are shown in Table (5.4). The restored “Lena” images are shown in Figs. (5.5) and (5.6), and the restored “Shepp-Logan” phantoms are shown in Figs. (5.9) and (5.9). As expected, all algorithms provide better restorations with less severe blur, although the noise variances are higher compared to the first set of experiments. Similar to the first experiment, algorithms  $TV1$  and  $TV2$  result in better restoration performance both in terms of ISNR and visual quality. Especially, the restorations by  $TV1$  are of very high quality for both images. The corresponding estimated PSFs for the restoration of “Lena” are shown in Figs. (5.13) and (5.14). A behavior similar to the first experiment is observed: Although close estimates are obtained in 40dB BSNR, the support of the PSF is overestimated by all algorithms in the high noise case.

We noticed in our experiments that the quality of the estimation of  $\mathbf{u}$  is a very important factor in the performance of the algorithms. For example, in the case of “Lena” image with

$\text{BSNR} = 40\text{dB}$ , if we run the algorithms  $TV1$  and  $TV2$  by calculating  $\mathbf{u}$  from the original image, we obtain  $\text{ISNR} = 3.16\text{dB}$ . Other cases showed similar improvements. Thus knowledge about this parameter greatly improves the ISNR performance. This also confirms that the poor performance of the algorithms in the presence of high noise, e.g.,  $\text{BSNR} = 20\text{dB}$ , is due to the fact that the parameter  $\mathbf{u}$  cannot be estimated well. As shown above, a simple smoothing of the gradient of the image largely improves the performance and the convergence of the algorithms. Therefore, it is safe to assume that in high noise-cases, incorporating robust gradient estimation methods, will improve the performance of the blind deconvolution process.

We examine the effect of prior information on the performance of the proposed algorithms in the third set of experiments. Generally, some information about the values of the hyperparameters is available and can be utilized in the restoration to improve the performance. For instance, the noise variance can be estimated quite accurately when a part of the image has uniform color. The image variance is more difficult to estimate from a single degraded observation. In this case, a set of images with similar characteristics can be used to acquire an estimate for this parameter. If an estimate of the image prior can be provided, the PSF variance can be approximated using this value (see [250] for details).

For simulation purposes, we calculated the true values for the hyperparameters from the original image and the PSF. Then, we applied  $TV1$  to the “Lena” image degraded by a Gaussian PSF and  $40\text{dB}$  BSNR with varying confidence parameters and obtained the ISNR evolution graphs shown in Fig. (5.19). To show the improved restoration performance and the best achievable ISNR, we also applied positivity and symmetry constraints to the estimated blur at each iteration as in [58]. Selected ISNR values from these graphs with the estimated hyperparameters are shown in Table 5.5. We included cases corresponding to best ISNR values when (a)

information about the noise variance is available, (b) information about only the PSF variance is available, (c) information about only the image variance is available, and (d) information about all hyperparameters is available. It is clear that if some information on the hyperparameters is available, biasing the algorithm towards these hyperparameters leads to improved ISNR values. However, it is interesting that incorporating the knowledge about the true noise variance is actually decreasing the performance, thus, it is better to put no confidence on this parameter and let the algorithms adaptively select this parameter at each iteration. On the other hand, we note that at convergence, the  $E[\beta]^{-1}$  almost always converged to the true noise variance.

It should also be emphasized that the most critical hyperparameter is  $\alpha_{bl}$  which determines the support of the estimated blur. It is clear from Fig. (5.19) and Table 5.5 that incorporating information about this parameter greatly increases the performance of the algorithm, and that the best ISNR is achieved when the true value of  $\alpha_{bl}$  is used. Some restoration results with these confidence parameters are shown in Fig. (5.20). Note that the restoration quality is almost as high as non-blind restoration results. One dimensional slices of the estimated blur corresponding to these cases are shown in Fig. (5.21). Overall, it is clear from the results that the algorithms provide better results in terms of ISNR when some information about these hyperparameters is provided.

In our last set of experiments, we experimented with a real image of Saturn, which is taken at the Calar Alto Observatory (Spain) on July, 1991, shown in Fig. 5.22(a). There is no exact expression the shape of the PSF for this image, however, the following approximation is suggested by [164, 171]

$$(5.48) \quad h(r) \propto \left(1 + \frac{r^2}{R^2}\right)^{-\delta},$$

where  $\delta \sim 3$  and  $R \sim 3.4$  are used. The non-blind restoration result using *TV2-NB* with this theoretical PSF is shown in Fig. 5.22(b). By running *TV2-NB* we also obtained estimates for the hyperparameters, as  $\bar{\beta}^o = 8.16$ ,  $\bar{\alpha}_{im}^o = 0.24$ , and  $\bar{\alpha}_{bl}^o = 1.6 \times 10^8$ .

In our experiments, we first run the proposed methods *TV1* and *TV2* with zero confidences placed on the prior values, thus,  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ . Our experiments show that, even in this case, *TV2* gives reasonable good restoration results, whereas *TV1* fails to remove the blur enough, so that the restoration is blurry. The restoration result by *TV2* is shown in Fig. 5.22(c). As shown before, utilizing prior knowledge about the parameters, we achieve better restoration results. For instance, by selecting  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ , we obtain the restoration results shown in Fig. 5.22(d) with *TV1* and Fig. 5.22(d) with *TV2*. As a comparison, the restoration result with *SAR1* with the same confidence parameters is shown in Fig. 5.22(e). The estimated PSFs corresponding to these cases as well as the theoretical PSF is shown in Fig. (5.23).



Figure 5.4. Restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.62 dB), (b) *TV2* algorithm (ISNR = -2.54 dB), (c) *SAR1* algorithm (ISNR = 1.06 dB), (d) *SAR2* algorithm (ISNR = -0.29 dB), (e) *TVI-NB* algorithm (ISNR = 3.31 dB), and (f) *TV2-NB* algorithm (ISNR = 3.29 dB).



Figure 5.5. Restorations of the Lena image blurred with a Gaussian PSF with variance 5 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 3.19 dB), (b) *TV2* algorithm (ISNR = 3.29 dB), (c) *SAR1* algorithm (ISNR = 2.35 dB), (d) *SAR2* algorithm (ISNR = 2.57 dB), (e) *TV1-NB* algorithm (ISNR = 4.98 dB), and (f) *TV2-NB* algorithm (ISNR = 4.93 dB).



Figure 5.6. Restorations of the Lena image blurred with a Gaussian PSF with variance 5 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 1.39 dB), (b) *TV2* algorithm (ISNR = -4.39 dB), (c) *SAR1* algorithm (ISNR = 0.36 dB), (d) *SAR2* algorithm (ISNR = -0.23 dB), (e) *TVI-NB* algorithm (ISNR = 2.92 dB), and (f) *TV2-NB* algorithm (ISNR = 2.83 dB).

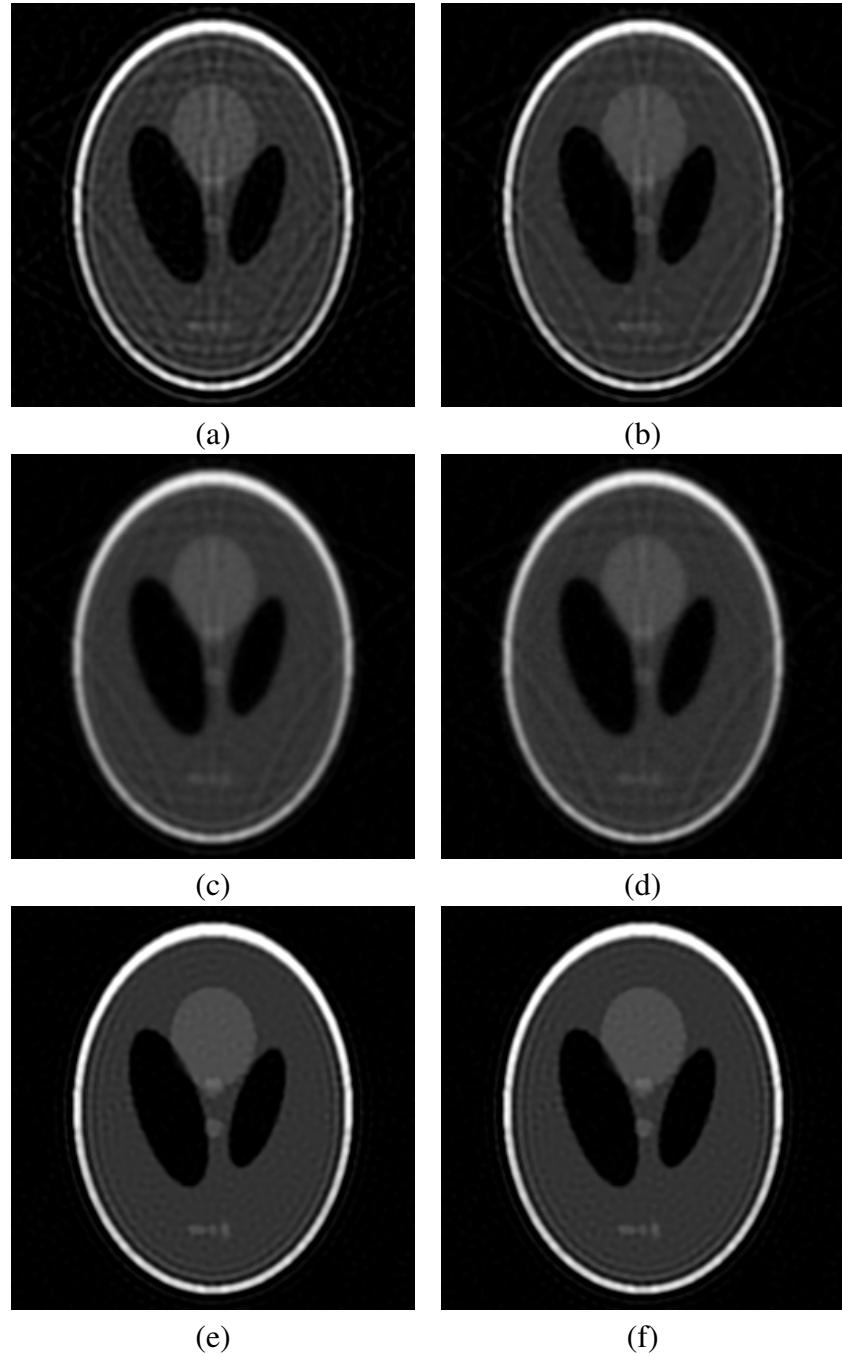


Figure 5.7. Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 3.07 dB), (b) *TV2* algorithm (ISNR = 3.36 dB), (c) *SAR1* algorithm (ISNR = 1.64 dB), (d) *SAR2* algorithm (ISNR = 1.81 dB), (e) *TV1-NB* algorithm (ISNR = 4.16 dB), and (f) *TV2-NB* algorithm (ISNR = 4.15 dB).

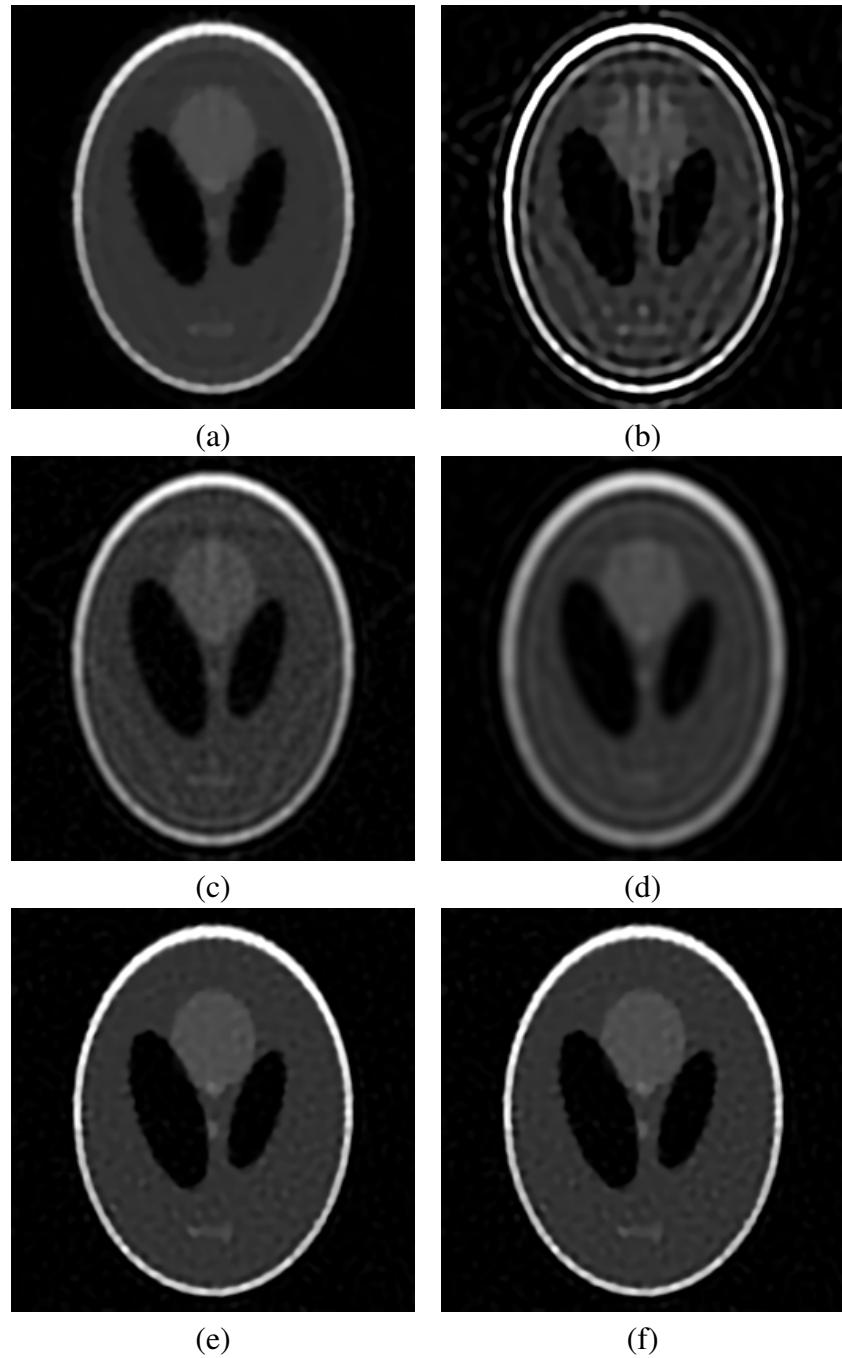


Figure 5.8. Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 9 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.47 dB), (b) *TV2* algorithm (ISNR = -1.41 dB), (c) *SAR1* algorithm (ISNR = 1.56 dB), (d) *SAR2* algorithm (ISNR = -0.15 dB), (e) *TV1-NB* algorithm (ISNR = 4.28 dB), and (f) *TV2-NB* algorithm (ISNR = 4.27 dB).

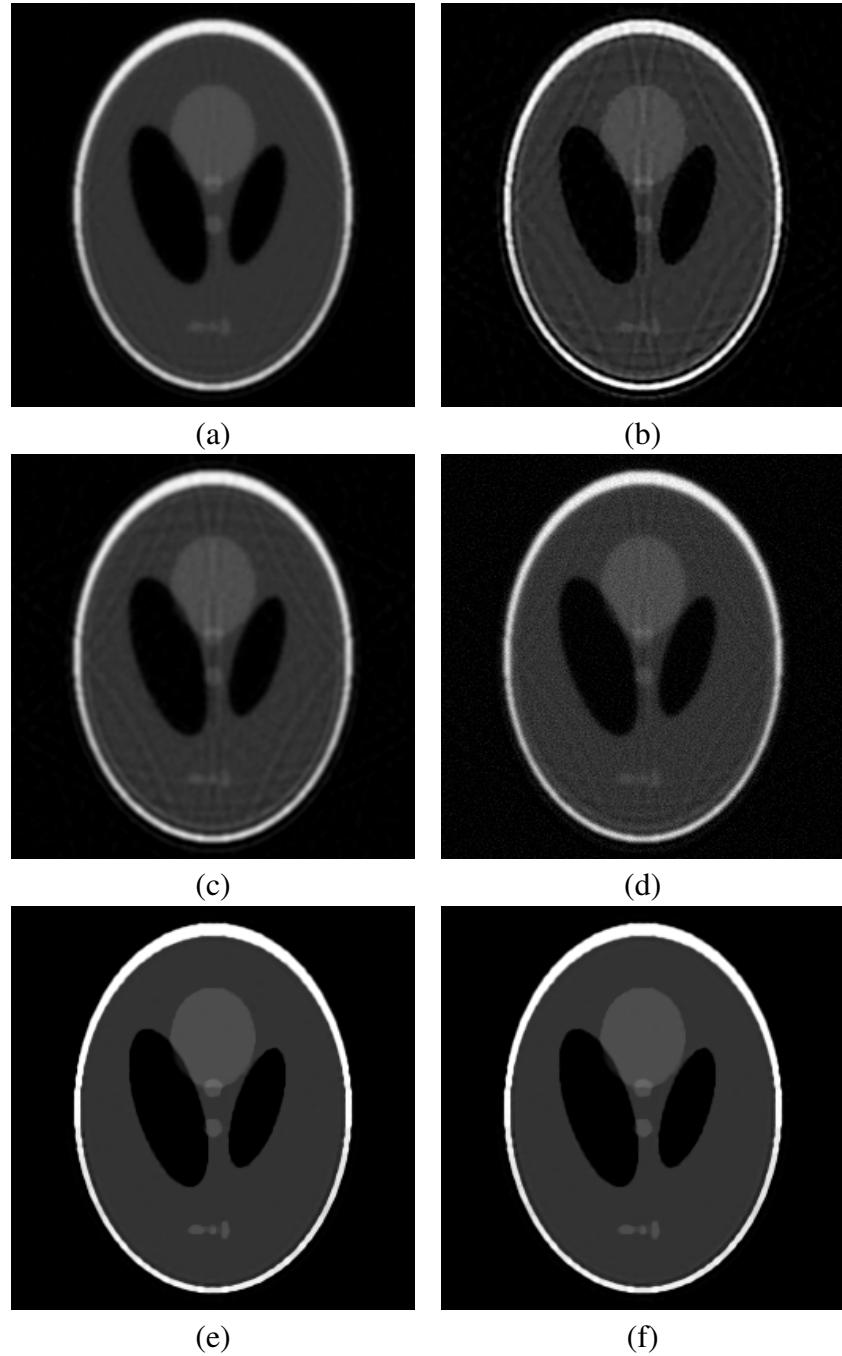


Figure 5.9. Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 5 and 40 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.05 dB), (b) *TV2* algorithm (ISNR = 3.79 dB), (c) *SAR1* algorithm (ISNR = 1.91 dB), (d) *SAR2* algorithm (ISNR = 1.30 dB), (e) *TV1-NB* algorithm (ISNR = 7.57 dB), and (f) *TV2-NB* algorithm (ISNR = 7.29 dB).

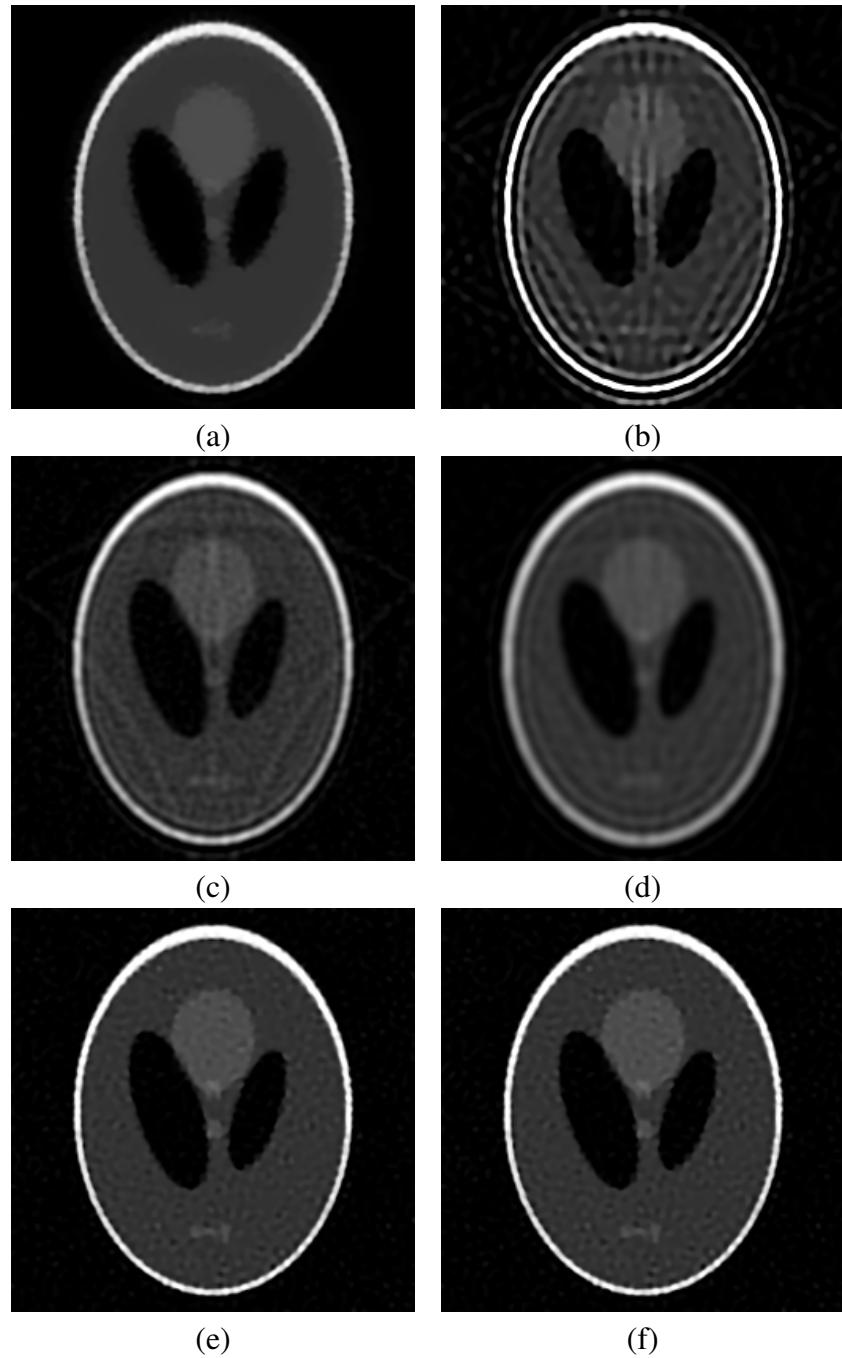


Figure 5.10. Restorations of the Shepp-Logan phantom blurred with a Gaussian PSF with variance 5 and 20 dB BSNR using the (a) *TV1* algorithm (ISNR = 2.09 dB), (b) *TV2* algorithm (ISNR = -2.89 dB), (c) *SAR1* algorithm (ISNR = 1.46 dB), (d) *SAR2* algorithm (ISNR = -0.17 dB), (e) *TV1-NB* algorithm (ISNR = 4.68 dB), and (f) *TV2-NB* algorithm (ISNR = 4.65 dB).

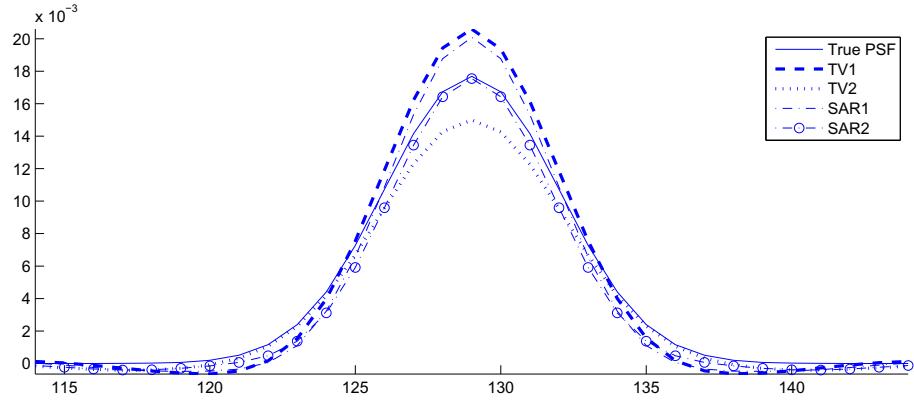


Figure 5.11. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 9 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

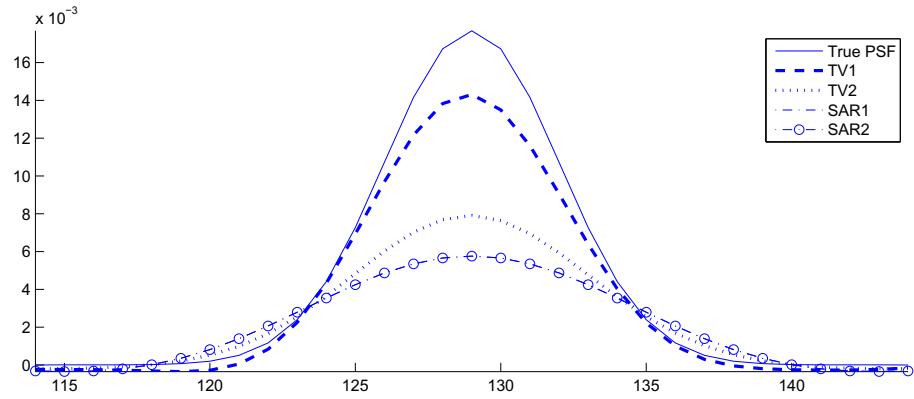


Figure 5.12. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 9 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

## 5.5. Conclusions

A novel total variation based blind deconvolution methodology has been proposed which simultaneously estimates the reconstructed image, the blur, and the hyperparameters of the Bayesian formulation. We have adopted a variational approach to approximate the posterior

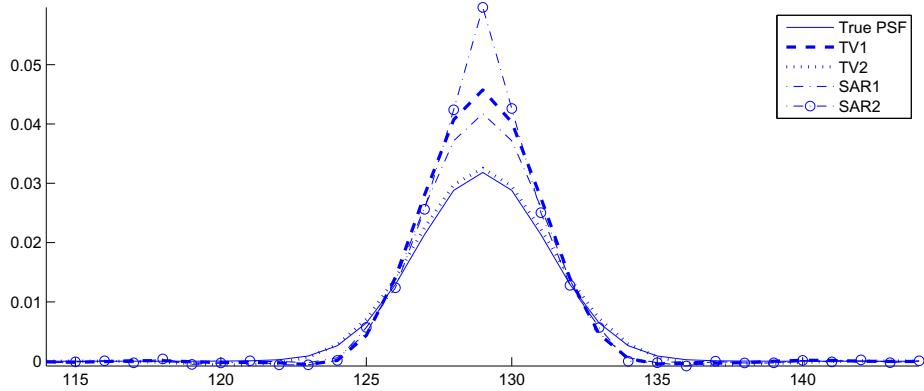


Figure 5.13. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 5 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

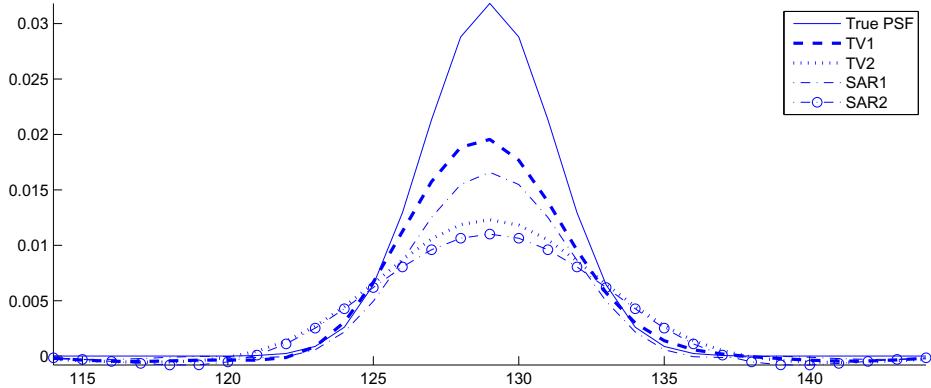


Figure 5.14. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of Lena image degraded by a Gaussian with variance 5 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

distributions of the unknown parameters, so that the uncertainty of the estimates can be evaluated and different values from these distributions can be used in the restoration process. Two algorithms are provided resulting from this approach. We have shown that the unknown parameters of the Bayesian formulation can be calculated automatically using only the observation or

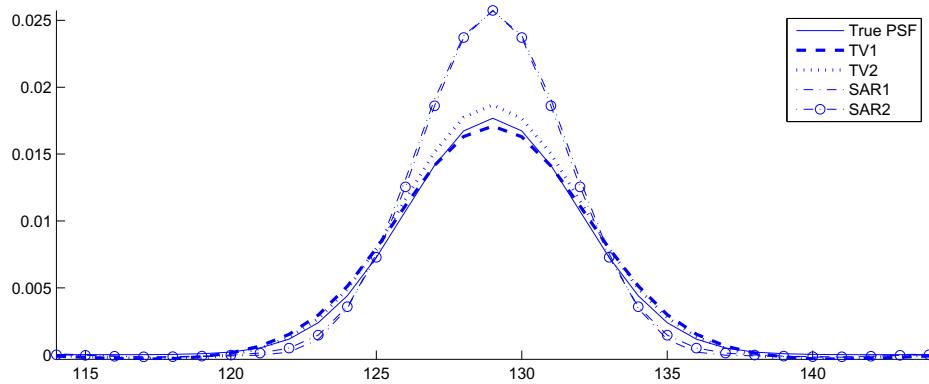


Figure 5.15. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Shepp-Logan phantom degraded by a Gaussian with variance 9 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

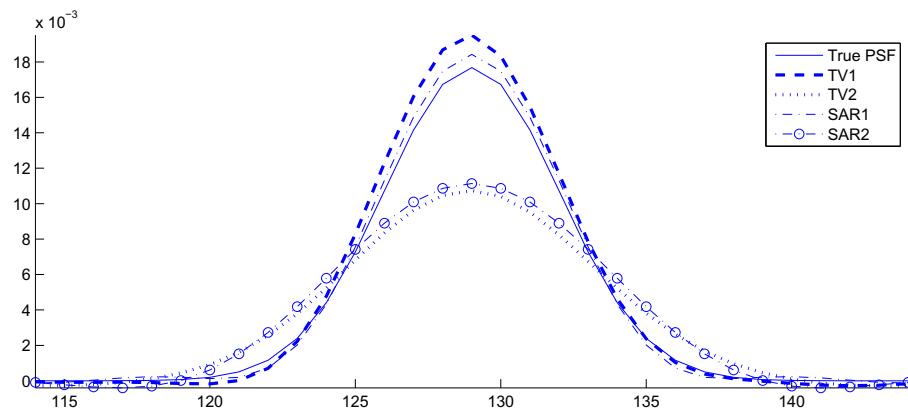


Figure 5.16. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Shepp-Logan phantom degraded by a Gaussian with variance 9 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

with different confidence values to improve the performance of the algorithms. Experimental results demonstrated the improved performance of the proposed algorithms.

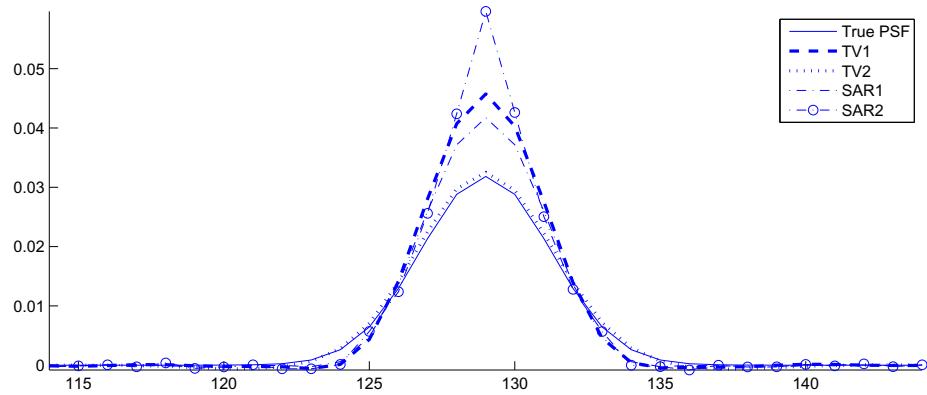


Figure 5.17. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Shepp-Logan phantom degraded by a Gaussian with variance 5 and BSNR = 40dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

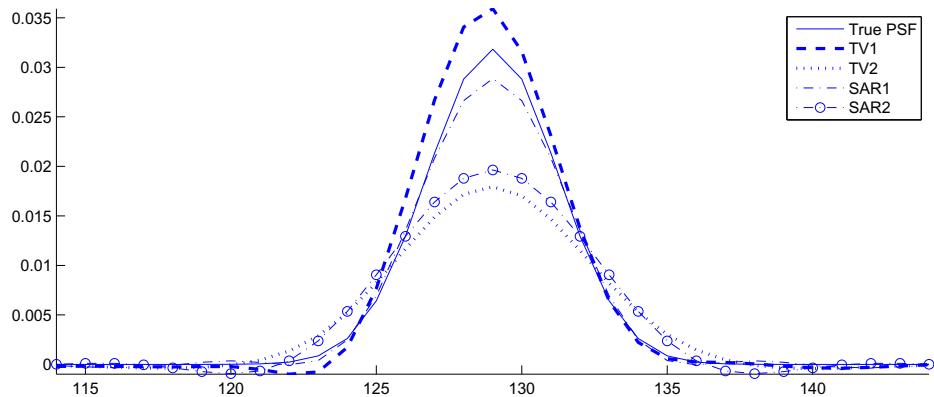


Figure 5.18. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Shepp-Logan phantom degraded by a Gaussian with variance 5 and BSNR = 20dB, with algorithms *TV1*, *TV2*, *SAR1* and *SAR2*.

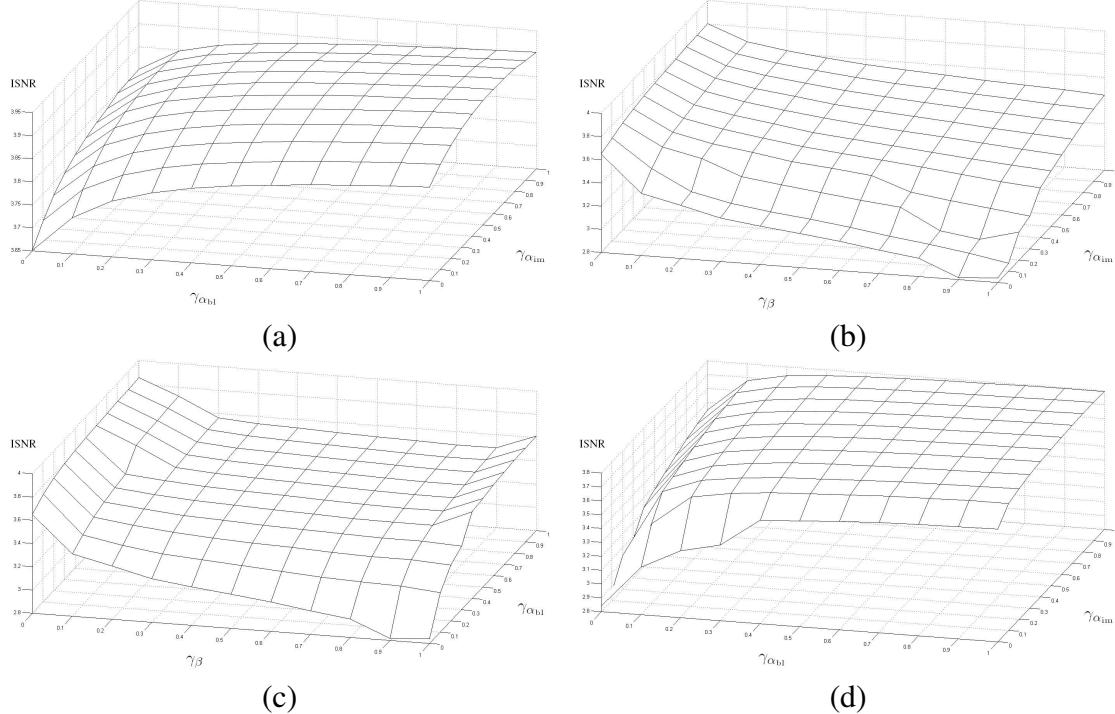


Figure 5.19. ISNR evolution for different values of confidence parameters for the Algorithm 1 (*TVI*) applied to “Lena” image degraded by a Gaussian with variance 9 and BSNR = 40dB. (a) For fixed  $\gamma_\beta = 0$ , (b) for fixed  $\gamma_{\alpha_{bl}} = 0$ , (c) for fixed  $\gamma_{\alpha_{im}} = 0$ , and (d) for fixed  $\gamma_\beta = 1$ .

Table 5.5. Posterior means of the distributions of the hyperparameters, ISNR, and number of iterations using *TVI* for the Lena image with 40 dB and 20 dB BSNR using  $\overline{\alpha_{im}^o} = 0.042$ ,  $\overline{\alpha_{bl}^o} = 4.6 \times 10^8$ , and  $\overline{\beta^o} = 6.25$ , respectively, for different values of  $\gamma_{\alpha_{im}}$ ,  $\gamma_{\alpha_{im}}$  and  $\gamma_\beta$ .

$\gamma_{\alpha_{im}}$	$\gamma_{\alpha_{bl}}$	$\gamma_\beta$	$E[\alpha_{im}]$	$E[\alpha_{bl}]$	$E[\beta]$	ISNR (dB)	iterations
0	0	0	0.088	$3.3 \times 10^8$	5.63	3.65	32
0	1	0	0.086	$4.6 \times 10^8$	5.62	3.85	38
1	1	0	0.041	$4.6 \times 10^8$	5.75	3.90	51
1	0	0	0.041	$3.7 \times 10^8$	5.76	3.80	51
0.6	1	0	0.051	$4.6 \times 10^8$	5.72	3.92	45
0.8	1	1	0.046	$4.6 \times 10^8$	6.25	3.80	82



Figure 5.20. Some restorations of the Lena image blurred with a Gaussian PSF with variance 9 and 40 dB BSNR using the *TV1* algorithm utilizing prior knowledge through confidence parameters and positivity and symmetry constraints on the estimated blur. (a)  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (b)  $\gamma_{\alpha_{im}} = 0, \gamma_{\alpha_{bl}} = 1, \gamma_{\beta} = 0$ , (c)  $\gamma_{\alpha_{im}} = 0.6, \gamma_{\alpha_{bl}} = 1, \gamma_{\beta} = 0$ , and (d)  $\gamma_{\alpha_{im}} = 0.8, \gamma_{\alpha_{bl}} = 1, \gamma_{\beta} = 1$ .

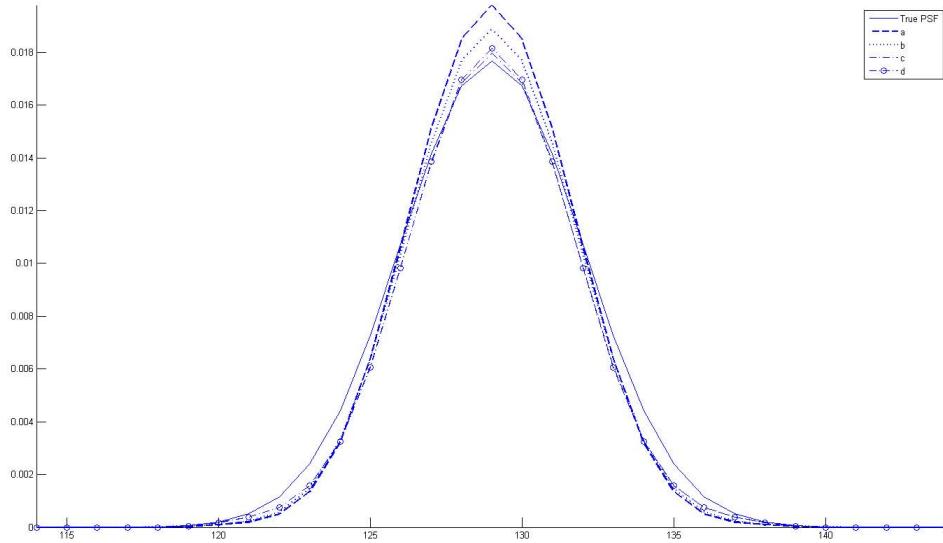


Figure 5.21. One-dimensional slice through the origin of the original and estimated PSFs in the restoration of the Lena image degraded by a Gaussian with variance 8 and BSNR = 40dB with algorithm *TV1*. (a) True PSF, Estimated PSF with (b)  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (c)  $\gamma_{\alpha_{im}} = 0$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , (d)  $\gamma_{\alpha_{im}} = 0.6$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 0$ , and (e)  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 1$ ,  $\gamma_{\beta} = 1$ ,

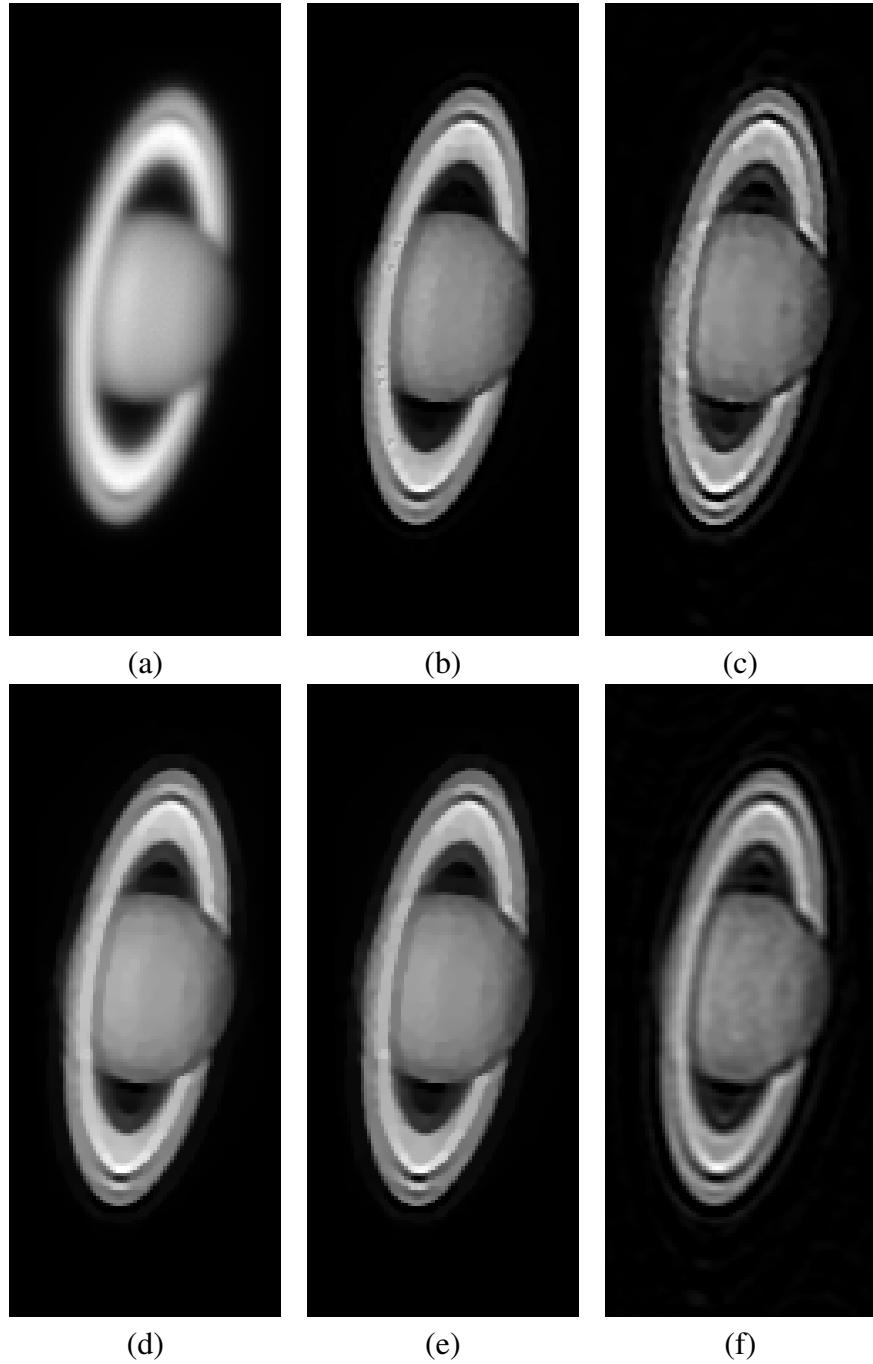


Figure 5.22. (a) Observed Saturn image. (b) Non-blind Restoration with *TV2-NB*, (c) Restoration with *TV2* with  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , (d) Restoration with *TVI* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ; (e) Restoration with *TV2* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ; (f) Restoration with *SARI* with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$ ;

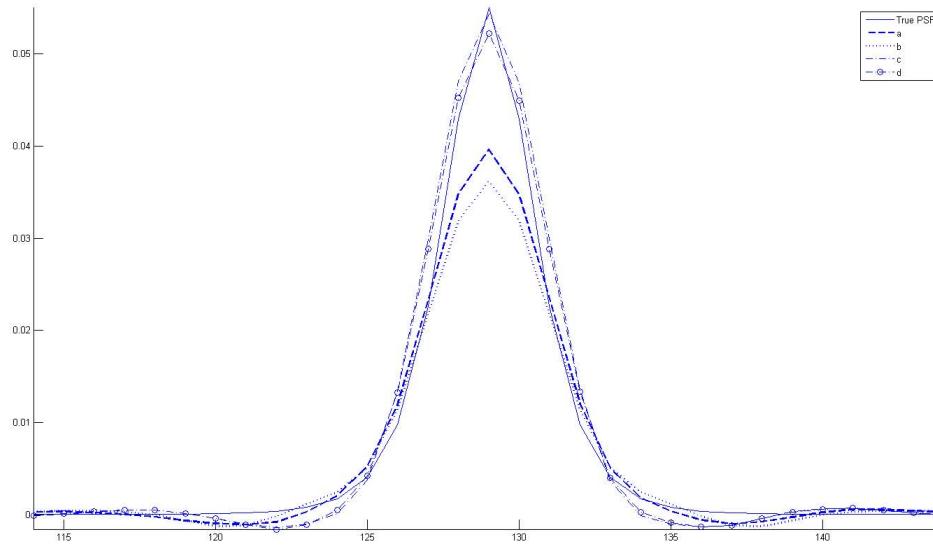


Figure 5.23. One-dimensional slice through the origin of the theoretical and estimated PSFs in the restoration of the Saturn image. (a) Theoretical PSF, estimated PSF (b) using  $TV2$  with  $\gamma_{\alpha_{im}} = \gamma_{\alpha_{bl}} = \gamma_{\beta} = 0.0$ , and with  $\gamma_{\alpha_{im}} = 0.8$ ,  $\gamma_{\alpha_{bl}} = 0.1$ , and  $\gamma_{\beta} = 0.8$  using (c)  $SARI$ , (d)  $TVI$  and (e)  $TV2$

## CHAPTER 6

### **Bayesian Blind Deconvolution from Differently Exposed Image Pairs**

#### **6.1. Introduction**

Taking high-quality photographs under low-lighting conditions is a major challenge. A longer exposure time than usual is required to obtain an image with low-noise, but any motion of the camera during exposure causes blur in the recorded image. On the other hand, a short exposure time will result in an image with a very high level of noise. Possible hardware-based solutions include increasing the light sensitivity (ISO) of the camera sensor, which increases the noise level; increasing the aperture, which results in a smaller depth of field in the acquired image; and using a tripod to stabilize the camera which is not practical in many cases. Additionally, many digital cameras incorporate optical image stabilizers, either inside the camera body or inside the lens, that significantly help in reducing degradations caused by hand-held photography. However, when the exposure times are too long, as might be required in some conditions, these hardware based solutions can not provide satisfactory results. In these cases, digital image stabilization methods, applied at a post-processing stage, provide a powerful means to obtain high-quality images using the low-quality observations.

In case a photograph is taken using a long exposure time under dim lighting, the resulting blur in the image can be removed by utilizing a single-image blind deconvolution algorithm. A number of methods are proposed for blind deconvolution of a single observation (see, for

example, [33] for a recent review), and the specific case of restoring images degraded by camera shake is addressed in [81, 214]. However, due to the challenging nature of the problem, obtaining a high-quality restoration result is very hard in most cases and requires significant user-supervision. Additionally, due to the difficulty in estimating the camera shake degradation, the restored images generally exhibit deconvolution artifacts such as ringing.

Another possible approach is to use a short exposure time to prevent blur at the expense of high noise, and then apply denoising algorithms to the sharp short-exposed image to remove the noise. Many advanced denoising methods are available in the literature (see, for instance, [192, 40, 129]). However, the noise level in such short-exposed images is generally so high that features of the underlying image are concealed, and the denoising algorithms cannot easily separate image and noise. An additional, and possibly more important problem is that due to the short exposure time, the images generally have low contrast and the colors might be partially lost.

Recently, deconvolution methods have been proposed that utilize an image pair instead of a single observation, where two images are taken with different exposure times [254]. Some digital cameras have exposure bracketing features which allow the user to acquire consecutive photographs with different exposure settings, which was mainly developed for high dynamic range applications [154, 68], but can also be used for image stabilization. Utilizing two images reduces the ill-posedness of the deconvolution problem, and generally results in much higher quality restorations than methods utilizing single observations [123]. A wide range of algorithms exists for the general problem of multi-frame blind deconvolution [198, 217, 33]. The specific case of blind deconvolution from a pair of short- and long-exposure images has been considered in [254, 224, 62]. In [254], the blur point spread function (PSF) is first identified

using the long-exposure image and the denoised version of the short-exposure image, where Tikhonov regularization and hysteresis thresholding is utilized to regularize the solution. This PSF estimate is then used in a classical image restoration method [199, 149] in order to obtain an estimate of the original image from the long-exposure image. A joint identification method is proposed in [224], where the unknown image and the PSF are estimated simultaneously. The image is modeled using a total-variation (TV) based prior, and the blur PSFs are estimated by imposing the constraint that the blur in the short exposure image is very small. No explicit blur model is utilized in this work, and denoising is applied to the blur estimates by thresholding in an ad hoc manner to enhance the estimates. Finally, sparsity priors with continuity constraints on the blurs are utilized in [62], and the image is modeled using a mixture-of-Gaussians prior on the image derivatives. However, the model is derived in a somewhat ad hoc manner, and the resulting algorithm has many parameters to tune, which makes it hard to apply to a wide range of images.

This work addresses the problem of blind deconvolution from a short- and long-exposed image pair. We provide a systematic modeling of the unknowns within a novel hierarchical Bayesian formulation and develop a blind deconvolution algorithm which jointly estimates the unknown image and blur. We utilize a TV-prior on the image to model natural image statistics and to achieve robustness in the algorithm. The blur in the long-exposed image is modeled using a mixture prior which imposes both sparsity and positivity on the estimated blur PSF. We also model the coupling between the long- and short-exposed images using an additional observation model. Moreover, we incorporate a fully-Bayesian approach, where all required model parameters are estimated along with the unknowns. As a result, the proposed algorithm does not require user-intervention and the restoration process is adaptively steered between

the long- and short-exposed images. Finally, we incorporate a variational Bayesian analysis, which provides estimates of the distributions of the unknowns. These distributions are implicitly used to incorporate the uncertainties of the estimates into the algorithm and to compensate for the estimation errors. We demonstrate with both synthetic and real image experiments that the proposed method provides very high quality restoration results and compares favorably to existing methods.

The rest of this chapter is organized as follows. In Sec. 6.2 we formulate the image acquisition processes mathematically. The unknown variables in our model are cast into a hierarchical Bayesian framework as presented in Sec. 6.3. The variational inference to estimate the unknowns and the proposed algorithm are presented in Sec. 6.4. Experimental results are presented in Sec. 6.6 and conclusions are drawn in Sec. 6.7.

## 6.2. Problem Formulation

The degradations in the image pair can be modeled using a linear and space invariant degradation model, by assuming that the blur is mainly caused by the shake of the camera during the long exposure time. Under this assumption, the observation processes can mathematically be expressed as follows

$$(6.1) \quad \mathbf{y}_1 = \mathbf{H}\mathbf{x} + \mathbf{n}_1$$

$$(6.2) \quad \mathbf{y}_2 = \lambda_1 \mathbf{C}\mathbf{x} + \lambda_2 \mathbf{1} + \mathbf{n}_2,$$

where  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are the observed images,  $\mathbf{x}$  the unknown original image,  $\mathbf{n}_1$  and  $\mathbf{n}_2$  the noise components, and the rest of the quantities are explained in the following. We use matrix-vector notation throughout the chapter, so that the images  $\mathbf{y}_1$ ,  $\mathbf{y}_2$ ,  $\mathbf{x}$ ,  $\mathbf{n}_1$ , and  $\mathbf{n}_2$  are  $N \times 1$  vectors,

where  $N$  is the number of pixels in each image. The  $N \times N$  matrix  $\mathbf{H}$  models the blur point spread function (PSF)  $\mathbf{h}$ , which has support  $M \leq N$ . The selection of the support  $M$  of the PSF is important and will be explained in the experimental results section. Note that the explicit construction of the matrix  $\mathbf{H}$  is not needed but it is used for notation purposes only.

Generally, the average luminance levels of the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are significantly different due to the different exposure times, and furthermore the images have to be geometrically registered. These geometric and photometric differences between the observed images are represented in (6.2) by the matrix  $\mathbf{C}$  and the parameters  $\lambda_1$  and  $\lambda_2$ , respectively. The geometric registration (or warping) matrix  $\mathbf{C}$  represents the motion between the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$ , and the parameters  $\lambda_1$  and  $\lambda_2$  represent the illumination differences and therefore, the photometric registration between the observations, and  $\mathbf{1}$  is an  $N \times 1$  vector of ones.

In this work we assume that the photometric and geometric calibration between the images  $\mathbf{y}_1$  and  $\mathbf{y}_2$  is calculated in a pre-processing stage, as is the case with the existing methods [224, 254, 62, 147]. The geometric registration can be performed using methods specifically designed for blurred/non-blurred image pairs (see, for example, [253]). In this work, we performed the geometric registration using publicly available registration software [2]. Alternatively, the registration algorithm in [241] can be used, as suggested by [147]. As demonstrated by our experimental results, crude initial registrations still result in high quality results due to the blur estimation process, which compensates for possible misalignments by appropriately shifting the estimated blur kernels. Photometric registration between the images is performed in a similar fashion by estimating the parameters  $\lambda_1$  and  $\lambda_2$  using the least squares solution utilizing the approximation  $\mathbf{y}_2 \approx \lambda_1 \mathbf{C} \mathbf{y}_1 + \lambda_2 \mathbf{1}$  and the RANSAC algorithm [48, 191].

After the observed images are corrected using the computed geometric and photometric registration, the observation models in (6.1) and (6.2) can be simplified using  $\mathbf{C} = \mathbf{I}$ ,  $\lambda_1 = 1$ , and  $\lambda_2 = 0$ , that is,

$$(6.3) \quad \mathbf{y}_1 = \mathbf{Hx} + \mathbf{n}_1$$

$$(6.4) \quad \mathbf{y}_2 = \mathbf{x} + \mathbf{n}_2.$$

These observation models will be utilized in the rest of the chapter. Using (6.3) and (6.4), the blind deconvolution problem is then to find estimates of  $\mathbf{x}$  and  $\mathbf{h}$  utilizing  $\mathbf{y}_1$  and  $\mathbf{y}_2$  and prior knowledge about  $\mathbf{x}$ ,  $\mathbf{h}$ ,  $\mathbf{n}_1$ , and  $\mathbf{n}_2$ .

### 6.3. Hierarchical Bayesian Model

The proposed hierarchical model is composed of two stages. In the first stage, prior distributions are utilized on the unknown image  $\mathbf{x}$  and blur  $\mathbf{h}$ , and conditional distributions are utilized for the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$ . These distributions in the first stage depend on certain parameters, called *hyperparameters*, which are modeled by hyperprior distributions in the second stage. The explicit forms of these distributions are presented in the following subsections.

#### 6.3.1. First Stage: Observation models

We assume that the observation noise in both observed images follows independent Gaussian distributions, that is, from (6.3) and (6.4),

$$(6.5) \quad p(\mathbf{y}_1 | \mathbf{x}, \mathbf{h}, \beta_1) \propto \beta_1^{N/2} \exp \left[ -\frac{\beta_1}{2} \| \mathbf{y}_1 - \mathbf{Hx} \|^2 \right],$$

and

$$(6.6) \quad p(\mathbf{y}_2|\mathbf{x}, \beta_2) \propto \beta_2^{N/2} \exp\left[-\frac{\beta_2}{2} \|\mathbf{y}_2 - \mathbf{x}\|^2\right],$$

where  $\beta_1$  and  $\beta_2$  are the precisions (inverse variances) of the noises, with  $\beta_1 \gg \beta_2$ .

Note that the dependency between the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$  is very high, as they are images of the same scene. To exploit this dependency, we incorporate an additional observation model making use of the coprimeness condition employed in some multichannel blind deconvolution methods (see, for example, [218]). Note first that combining (6.3) and (6.4) we obtain

$$(6.7) \quad \mathbf{y}_1 - \mathbf{H}\mathbf{y}_2 | \mathbf{h}, \beta_1, \beta_2 \sim \mathcal{N}(\mathbf{0}, \beta_1^{-1}\mathbf{I} + \beta_2^{-1}\mathbf{H}\mathbf{H}^T).$$

We then modify this observation model by considering that the noise in this model is uncorrelated, and obtain the following third independent observation model

$$(6.8) \quad \mathbf{y}_1 - \mathbf{H}\mathbf{y}_2 | \mathbf{h}, \beta_{12} \sim \mathcal{N}(\mathbf{0}, \beta_{12}^{-1}\mathbf{I})$$

where  $\beta_{12}^{-1} > 0$ . We have experimentally observed that using this additional observation model and defining the general observation model as

$$(6.9) \quad \begin{aligned} p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}, \mathbf{h}, \beta_1, \beta_2, \beta_{12}) &\propto \beta_1^{N/2} \beta_2^{N/2} \beta_{12}^{N/2} \exp\left[-\frac{\beta_1}{2} \|\mathbf{y}_1 - \mathbf{H}\mathbf{x}\|^2\right] \exp\left[-\frac{\beta_2}{2} \|\mathbf{y}_2 - \mathbf{x}\|^2\right] \\ &\times \exp\left[-\frac{\beta_{12}}{2} \|\mathbf{y}_1 - \mathbf{H}\mathbf{y}_2\|^2\right], \end{aligned}$$

produces both a better restored image and a better estimate of the blur. This additional observation model incorporates the strong dependency between the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$  into the inference procedure (see [216] for a similar approach).

### 6.3.2. First Stage: Prior model on the blur

Since the blur is mainly caused by the shaking of the camera during the long exposure time, it exhibits the characteristics of the nonuniform motion blur. Hence, it is expected to be very sparse, i.e., most of the PSF coefficients being zero or very small. In order to exploit this information, we utilize a mixture prior of  $D$  exponential distributions on each PSF coefficient, that is,

$$(6.10) \quad p(\mathbf{h} | \{\tau_{jd}\}, \{\sigma_{jd}\}) = \prod_{j=1}^M \left( \sum_{d=1}^D \tau_{jd} \text{Expon}(h_j | \sigma_{jd}) \right)$$

with  $\tau_{jd}$  the mixture coefficients for each pixel  $j$  and

$$(6.11) \quad \text{Expon}(h_j | \sigma_{jd}) = \begin{cases} \sigma_{jd} \exp(-\sigma_{jd} h_j) & \text{if } h_j \geq 0, \\ 0 & \text{if } h_j < 0. \end{cases}$$

with  $\sigma_{jd}$  the parameters of each exponential distribution.

Note that this blur prior enforces sparsity to a great extent, and the degree of sparsity is increased by increasing the number of mixture coefficients  $D$  [162]. In addition to imposing sparsity, note that (6.11) also imposes positivity on the blur coefficients  $h_j$ . This property makes the prior especially useful, since unlike most previous works the positivity constraint is imposed during the formulation and subsequent optimization process, and not artificially after the optimization, which can move the estimates away from their optimal values. Note that this mixture-of-exponentials prior has also been utilized before for modeling PSFs resulting from camera shake [81, 162] and in independent component analysis [162].

### 6.3.3. First Stage: Prior model on the image

The unknown image  $\mathbf{x}$  is expected to be mostly smooth except at the locations of discontinuities (e.g., edges). Therefore, as the prior model on the image  $\mathbf{x}$ , we utilize the total variation function because it preserves the edges in the image by not over-penalizing discontinuities while imposing smoothness [204]. Specifically, we utilize the quadratic approximation of the TV prior, as developed in Chapter 3, given by

$$(6.12) \quad p(\mathbf{x}|\alpha_{im}) = c \alpha_{im}^{N/2} \exp\left[-\frac{1}{2}\alpha_{im} \text{TV}(\mathbf{x})\right],$$

where  $c$  is a constant and

$$(6.13) \quad \text{TV}(\mathbf{x}) = \sum_{j=1}^N \sqrt{(\Delta_j^h(\mathbf{x}))^2 + (\Delta_j^v(\mathbf{x}))^2}.$$

The operators  $\Delta_j^h(\mathbf{x})$  and  $\Delta_j^v(\mathbf{x})$  correspond to, respectively, horizontal and vertical first order differences, at pixel  $j$ , that is,  $\Delta_j^h(\mathbf{x}) = x_j - x_{l(j)}$  and  $\Delta_j^v(\mathbf{x}) = x_j - x_{a(j)}$ , where  $l(j)$  and  $a(j)$  denote the nearest neighbors of  $j$ , to the left and above, respectively.

### 6.3.4. Second Stage: Hyperpriors on the hyperparameters

In the second stage of the hierarchical model, we model the hyperparameters  $\alpha_{im}$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_{12}$ ,  $\{\sigma_{jd}\}$  and  $\{\tau_{jd}\}$  by hyperprior distributions. In Bayesian models, hyperprior distributions are generally chosen to be conjugate distributions, i.e., they have the same functional form as the product of the conditional distribution and the priors. This choice of hyperpriors simplifies the analytical derivation of the inference procedure. Based on this, we employ conjugate Gamma distributions for the hyperpriors  $p(\alpha_{im})$ ,  $p(\beta_1)$ ,  $p(\beta_2)$ ,  $p(\beta_{12})$ ,  $p(\sigma_{jd})$  and Dirichlet distributions

on the mixture coefficients  $p(\tau_{jd})$ , that is,

$$(6.14) \quad p(\alpha_{im}) = \text{Gamma}(\alpha_{im}|a_{\alpha_{im}}^o, b_{\alpha_{im}}^o)$$

$$(6.15) \quad p(\beta_1) = \text{Gamma}(\beta_1|a_{\beta_1}^o, b_{\beta_1}^o)$$

$$(6.16) \quad p(\beta_2) = \text{Gamma}(\beta_2|a_{\beta_2}^o, b_{\beta_2}^o)$$

$$(6.17) \quad p(\beta_{12}) = \text{Gamma}(\beta_{12}|a_{\beta_{12}}^o, b_{\beta_{12}}^o)$$

$$(6.18) \quad p(\sigma_{jd}) = \text{Gamma}(\sigma_{jd}|a_{\sigma_{jd}}^o, b_{\sigma_{jd}}^o), \quad j = 1, \dots, M, d = 1, \dots, D$$

$$(6.19) \quad p(\{\tau_{jd}\}_{d=1}^D) = \text{Dirichlet}(\{\tau_{jd}\}_{d=1}^D|c_{\tau_{jd}}^o), \quad j = 1, \dots, M,$$

with shape parameters  $a_{\alpha_{im}}^o, a_{\beta_1}^o, a_{\beta_2}^o, a_{\beta_{12}}^o$  and scale parameters  $b_{\alpha_{im}}^o, b_{\beta_1}^o, b_{\beta_2}^o, b_{\beta_{12}}^o$ . The shape and scale parameters of the Gamma distributions are set to a small common value (e.g., 0.001), and  $c_{\tau_{jd}}^o$  is set to 1 to obtain vague hyperpriors which make the estimation process rely more on the observations than on prior knowledge. Note, however, that these hyperprior distributions are very flexible in incorporating additional information provided by the user. If some prior knowledge on the value of some of the hyperparameters (for instance, about the noise variances in the observed images) is available, this information can easily be incorporated into the estimation procedure by choosing the scale and shape parameters of the corresponding distributions accordingly (see, for example, [169, 13] for such incorporation of prior knowledge). Moreover, note that using nonzero values for the shape and scale parameters aid in avoiding trivial solutions such as delta PSF estimates. Note also that the additional observation model in (6.8) has an important role in preventing the blur to be estimated as a delta function unless  $\mathbf{y}_1 - \mathbf{y}_2$  can be considered as Gaussian independent noise.

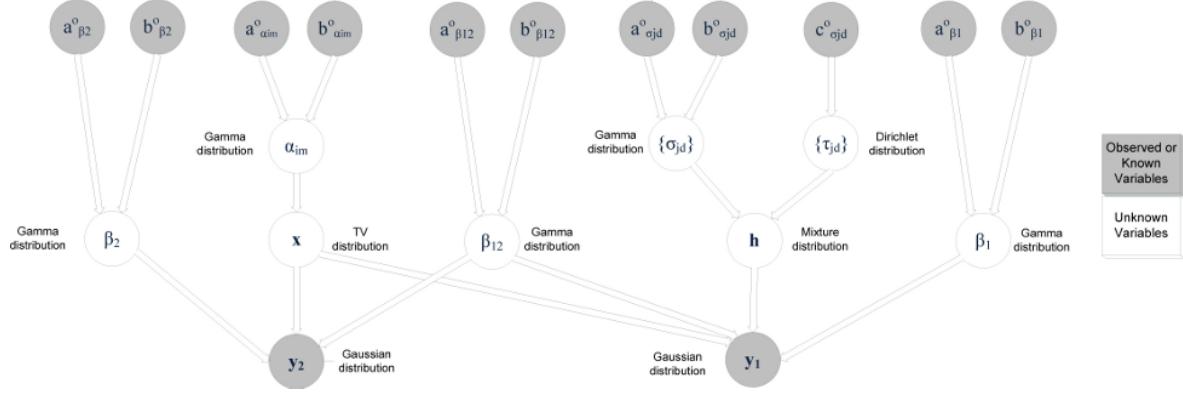


Figure 6.1. Graphical model showing relationships between variables.

Finally, combining the first and second stage of the hierarchical model we obtain the following global distribution

$$\begin{aligned}
 p(\mathbf{x}, \mathbf{h}, \mathbf{y}_1, \mathbf{y}_2, \alpha_{im}, \beta_1, \beta_2, \beta_{12}, \{\tau_{jd}\}, \{\sigma_{jd}\}) = & p(\mathbf{x}|\alpha_{im}) p(\mathbf{h}|\{\tau_{jd}\}, \{\sigma_{jd}\}) p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}, \mathbf{h}, \beta_1, \beta_2, \beta_{12}) \\
 (6.20) \quad & \times p(\alpha_{im}) p(\beta_1) p(\beta_2) p(\beta_{12}) \prod_{j=1}^M \left[ p(\{\tau_{jd}\}_{d=1}^D) \prod_{d=1}^D p(\sigma_{jd}) \right].
 \end{aligned}$$

The proposed hierarchical model is shown in Fig. (6.1) using a directed acyclic graph demonstrating the dependencies between the unknown and observed variables.

#### 6.4. Variational Bayesian Inference

In Bayesian formulations, the inference is based on the posterior distribution, which in our case is intractable. Therefore, in this work we utilize variational distribution approximations. Let us denote by  $\Theta$  the set of unknowns, i.e.,  $\Theta = \{\mathbf{x}, \mathbf{h}, \alpha_{im}, \beta_1, \beta_2, \beta_{12}, \{\sigma_{jd}\}, \{\tau_{jd}\}\}$ . The goal is to approximate the posterior distribution  $p(\Theta|\mathbf{y}_1, \mathbf{y}_2)$  by another distribution  $q(\Theta)$  which allows a tractable analysis. The approximating distribution  $q(\Theta)$  is found by minimizing the

Kullback-Leibler (KL) divergence between  $q(\Theta)$  and  $p(\Theta|y_1, y_2)$ , which is given by

$$(6.21) \quad C_{KL}(q(\Theta) \| p(\Theta|y_1, y_2)) = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta|y_1, y_2)} \right) d\Theta$$

$$(6.22) \quad = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta, y_1, y_2)} \right) d\Theta + \text{const.}$$

Generally, the only assumption in variational Bayesian analysis is that the approximating distribution  $q(\Theta)$  is factorizable. In this work, we use the following factorization

$$(6.23) \quad q(\Theta) = q(\mathbf{x}) q(\mathbf{h}) q(\alpha_{im}) q(\beta_1) q(\beta_2) q(\beta_{12}) \prod_{j=1}^M \left[ q(\{\tau_{jd}\}_{d=1}^D) \prod_{d=1}^D q(\sigma_{jd}) \right]$$

with

$$(6.24) \quad q(\mathbf{h}) = \prod_{j=1}^M q(h_j)$$

Unfortunately the general results from variational Bayesian analysis cannot be directly utilized in this work, since the TV and mixture priors in our model render the calculation of the KL divergence in (6.22) not possible. The problems caused by the TV prior can be avoided by utilizing a majorization-minimization approach, whose details are given Chapter 3. We will provide a brief overview here for completeness. Let us consider again the geometric-arithmetic mean inequality for any real numbers  $a \geq 0$  and  $b > 0$

$$(6.25) \quad \sqrt{ab} \leq \frac{a+b}{2} \Rightarrow \sqrt{a} \leq \frac{a+b}{2\sqrt{b}}.$$

Let us also define for  $\alpha_{\text{im}}$ ,  $\mathbf{x}$ , and an  $N$ -dimensional vector  $\mathbf{w} \in (R^+)^N$ , with components  $w_i$ ,  $i = 1, \dots, N$ , the following functional

$$(6.26) \quad \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w}) = c \alpha_{\text{im}}^{N/2} \exp \left[ -\frac{\alpha_{\text{im}}}{2} \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + w_i}{\sqrt{w_i}} \right],$$

where  $c$  is the same constant as in (6.12). Using  $a = (\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2$  and  $b = w_i$  in the inequality (6.25), it can be seen that the functional  $\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})$  is a lower bound of the image prior  $p(\mathbf{x}|\alpha_{\text{im}})$ , that is,

$$(6.27) \quad p(\mathbf{x}|\alpha_{\text{im}}) \geq \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w}).$$

The quadratic form of the bounding functional  $\mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})$  renders the analytical derivation of the Bayesian inference tractable. Using the lower bound in (6.26), a lower bound of the joint probability distribution in (6.20) can be found, that is,

$$\begin{aligned} p(\Theta) &\geq \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w}) p(\mathbf{h}|\{\tau_{jd}\}, \{\sigma_{jd}\}) p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}, \mathbf{h}, \beta_1, \beta_2, \beta_{12}) \\ &\times p(\alpha_{\text{im}}) p(\beta_1) p(\beta_2) p(\beta_{12}) \prod_{j=1}^M \left[ p(\{\tau_{jd}\}_{d=1}^D) \prod_{d=1}^D p(\sigma_{jd}) \right] \\ (6.28) \quad &= \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2), \end{aligned}$$

which leads to the following upper bound for the KL divergence in (6.22)

$$(6.29) \quad C_{KL}(q(\Theta) \| p(\Theta | \mathbf{y}_1, \mathbf{y}_2)) \leq C_{KL}(q(\Theta) \| \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)) + \text{const.}$$

The upper bound  $C_{KL}(q(\Theta) \parallel F(\Theta, w, y_1, y_2))$  can be made tighter by minimizing it with respect to  $w$ , since

$$(6.30) \quad C_{KL}(q(\Theta) \parallel p(\Theta|y_1, y_2)) \leq \min_w C_{KL}(q(\Theta) \parallel F(\Theta, w, y_1, y_2)) + \text{const.}$$

Therefore, by minimizing the upper bound  $C_{KL}(q(\Theta) \parallel F(\Theta, w, y_1, y_2))$  with respect to both  $q(\Theta)$  and  $w$ , the upper bound can iteratively be made closer to the KL distance  $C_{KL}(q(\Theta) \parallel p(\Theta|y_1, y_2))$ . Thus, this upper bound can be used as an approximation to the original KL distance, and variational Bayesian analysis can be performed using this upper bound instead (see [13] and Chapters 3 and 5 for details on the theoretical justification). For each unknown  $\theta \in \Theta$ , the distribution approximation  $q(\theta)$  can then be found by alternating the minimization of  $C_{KL}(q(\Theta) \parallel F(\Theta, w, y_1, y_2))$  with respect to each  $q(\theta)$  by holding  $q(\Theta_\theta)$  constant, where  $\Theta_\theta$  denotes the set  $\Theta$  with  $\theta$  removed from the set. This approach results in the following general solution [31]

$$(6.31) \quad q(\theta) = \text{const} \times \exp \left( E_{q(\Theta_\theta)} [\log F(\Theta, w, y_1, y_2)] \right),$$

where  $E_{q(\Theta_\theta)}[\cdot]$  denotes the expectation with respect to the distribution  $q(\Theta_\theta)$ . In order to solve (6.31), an additional approximation is needed when using mixture priors. Specifically, we utilize Jensen's inequality as follows [162]

$$(6.32) \quad \log \left[ \prod_{j=1}^M \left( \sum_{d=1}^D \tau_{jd} \text{Expon}(h_j | \sigma_{jd}) \right) \right] \leq \sum_{j=1}^M \sum_{d=1}^D \mu_{jd} \log \left( \frac{\tau_{jd}}{\mu_{jd}} \text{Expon}(h_j | \sigma_{jd}) \right),$$

with  $\sum_{d=1}^D \mu_{jd} = 1$ ,  $j = 1, \dots, M$ . An analysis of the closeness of this bound can be found in [162]. The auxiliary variables  $\mu_{jd}$  need to be computed along with the unknowns  $\Theta$ , as will

be shown in the next section. Using (6.32), we obtain an upper bound of  $\log \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)$  as follows. Let us denote by  $\bar{\mathbf{F}}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)$  the product of the terms in  $\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)$  except  $p(\mathbf{h} | \{\tau_{jd}\}, \{\sigma_{jd}\})$ , that is,  $\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2) = p(\mathbf{h} | \{\tau_{jd}\}, \{\sigma_{jd}\}) \bar{\mathbf{F}}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)$ . Then,

$$\begin{aligned} \log \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2) &= \log [p(\mathbf{h} | \{\tau_{jd}\}, \{\sigma_{jd}\}) \bar{\mathbf{F}}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)] = \log p(\mathbf{h} | \{\tau_{jd}\}, \{\sigma_{jd}\}) + \log \bar{\mathbf{F}}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2) \\ (6.33) \quad &\leq \sum_{j=1}^M \sum_{d=1}^D \mu_{jd} \log \left( \frac{\tau_{jd}}{\mu_{jd}} \text{Expon}(h_j | \sigma_{jd}) \right) + \log \bar{\mathbf{F}}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2), \end{aligned}$$

$$(6.34) \quad = \mathbf{B}(\Theta, \mathbf{w}, \boldsymbol{\mu}, \mathbf{y}_1, \mathbf{y}_2),$$

with  $\boldsymbol{\mu} = \{\mu_{jd} | j = 1, \dots, M, d = 1, \dots, D\}$ . Utilizing this upper bound, we replace the general solution in (6.31) by

$$(6.35) \quad q(\theta) = \text{const} \times \exp \left( E_{q(\Theta_\theta)} [\mathbf{B}(\Theta, \mathbf{w}, \boldsymbol{\mu}, \mathbf{y}_1, \mathbf{y}_2)] \right).$$

Applying this general solution (6.35) to each unknown results in an iterative procedure, which converges to the best approximation of the true posterior distribution  $p(\Theta | \mathbf{y}_1, \mathbf{y}_2)$  by distributions of the form in (6.23). This iterative procedure provides estimates  $q(\theta)$  to the distributions of the unknowns. In this work, we utilize the means of these distributions as the point estimates of the unknowns. Finally, note that in the case of  $D = 1$ , the solutions provided by (6.31) and (6.35) are equal.

## 6.5. Calculation of Posterior Distribution Approximations

In this section, we provide the explicit forms of each  $q(\cdot)$  distribution. In the following, the means of the distributions will be denoted by  $\langle \cdot \rangle = E_{q(\theta)} [\cdot]$ , when the corresponding distribution is clear from the context.

The distribution  $q(\mathbf{x})$  is calculated from (6.35) as a multivariate Gaussian distribution, that is,

$$(6.36) \quad q(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \langle \mathbf{x} \rangle, \Sigma_{\mathbf{x}})$$

where its mean and covariance are given by

$$(6.37) \quad \langle \mathbf{x} \rangle = \Sigma_{\mathbf{x}} (\langle \beta_1 \rangle \langle \mathbf{H} \rangle^T \mathbf{y}_1 + \langle \beta_2 \rangle \mathbf{y}_2)$$

$$(6.38) \quad \Sigma_{\mathbf{x}}^{-1} = \langle \alpha_{im} \rangle (\Delta^h)^T \mathbf{W} (\Delta^h) + \langle \alpha_{im} \rangle (\Delta^v)^T \mathbf{W} (\Delta^v) + \langle \beta_1 \rangle \langle \mathbf{H}^T \mathbf{H} \rangle + \langle \beta_2 \rangle \mathbf{I}$$

with

$$(6.39) \quad w_j = \langle (\Delta_j^h(\mathbf{x}))^2 + (\Delta_j^v(\mathbf{x}))^2 \rangle, j = 1, \dots, N,$$

$$(6.40) \quad \mathbf{W} = \text{diag} \left( \frac{1}{\sqrt{w_j}} \right), j = 1, \dots, N.$$

The mean  $\langle \mathbf{x} \rangle$  of the distribution  $q(\mathbf{x})$  is used as the image estimate, which is calculated by applying a conjugate gradient method in (6.37). It can be seen from (6.38) that the matrix  $\mathbf{W}$  in (6.40) is the spatial adaptivity matrix which controls the amount of smoothing at each pixel location depending on the intensity variation at that pixel, as expressed by the vector  $\mathbf{w}$  representing the total variation of the estimated image. It therefore controls the trade-off between the data fidelity and image smoothness. Additionally, the parameters  $\langle \beta_1 \rangle$ ,  $\langle \beta_2 \rangle$  and  $\langle \beta_{12} \rangle$  control the fidelity of the image estimate to the observed images  $\mathbf{y}_1$  and  $\mathbf{y}_2$ . Since they are also estimated simultaneously with the image (as shown below), the estimation process is automatically steered towards the more reliable observation. The reliability is expressed by the

constraints imposed on the image and blur estimates using their corresponding prior distributions. For instance, if the noise level in the short-exposed image is low, the estimation process relies more on  $\mathbf{y}_2$  by increasing  $\langle \beta_2 \rangle$ , as this provides smoother PSF and image estimates.

Next we find the distribution approximations  $q(h_j)$  of the blur PSF coefficients from (6.35).

Using (6.34) and (6.35), we have

$$(6.41) \quad q(h_j) \propto \exp \left[ -\frac{\langle \beta_1 \rangle}{2} \langle \| \mathbf{y}_1 - \mathbf{X}\mathbf{h} \|^2 \rangle - \frac{\langle \beta_{12} \rangle}{2} \langle \| \mathbf{y}_1 - \mathbf{Y}_2\mathbf{h} \|^2 \rangle \right] \\ \times \left[ \sum_{d=1}^D \mu_{jd} \log \left( \frac{\langle \tau_{jd} \rangle}{\mu_{jd}} \langle \sigma_{jd} \rangle \exp(-\langle \sigma_{jd} \rangle h_j) \right) \right].$$

The first and second terms in the exponent can be written as

$$(6.42) \quad \langle \| \mathbf{y}_1 - \mathbf{X}\mathbf{h} \|^2 \rangle = \left\langle \sum_{n=1}^N \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M X_{nm} h_m - X_{nj} h_j \right)^2 \right\rangle$$

$$(6.43) \quad \langle \| \mathbf{y}_1 - \mathbf{Y}_2\mathbf{h} \|^2 \rangle = \left\langle \sum_{n=1}^N \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M (Y_2)_{nm} h_m - (Y_2)_{nj} h_j \right)^2 \right\rangle$$

Substituting these identities in (6.41) and ignoring the terms not containing  $h_j$ , we obtain

$$(6.44) \quad q(h_j) \propto \exp \left[ -\frac{\langle \beta_1 \rangle}{2} \sum_{n=1}^N \langle X_{nj}^2 \rangle h_j^2 + \langle \beta_1 \rangle \sum_{n=1}^N \left\langle X_{nj} \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M X_{nm} h_m \right) \right\rangle h_j \right. \\ \left. - \frac{\langle \beta_{12} \rangle}{2} \sum_{n=1}^N \langle (Y_2)_{nj}^2 \rangle h_j^2 + \langle \beta_{12} \rangle \sum_{n=1}^N \left\langle (Y_2)_{nj} \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M (Y_2)_{nm} \langle h_m \rangle \right) \right\rangle h_j \right. \\ \left. - \sum_{d=1}^D \langle \sigma_{jd} \rangle \mu_{jd} h_j \right]$$

Note that (6.44) is in standard form of the rectified Gaussian distribution (see [162], Appendix A.3), given by

$$(6.45) \quad q(h_j) = \mathcal{N}^R(h_j | \hat{h}_j, \tilde{h}_j)$$

with parameters

$$(6.46) \quad \begin{aligned} \hat{h}_j &= (\tilde{h}_j)^{-1} \left[ - \sum_{d=1}^D \langle \sigma_{jd} \rangle \mu_{jd} + \sum_{n=1}^N \langle \beta_1 X_{nj} \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M X_{nm} h_m \right) \rangle \right. \\ &\quad \left. + \sum_{n=1}^N \langle \beta_{12} \rangle (Y_2)_{nj} \left( (y_1)_n - \sum_{\substack{m=1 \\ m \neq j}}^M (Y_2)_{nm} \langle h_m \rangle \right) \right] \end{aligned}$$

$$(6.47) \quad \tilde{h}_j = \sum_{n=1}^N \langle \beta_1 \rangle \langle X_{nj}^2 \rangle + \langle \beta_{12} \rangle \sum_{n=1}^N (Y_2)_{nj}^2$$

where  $(\cdot)_{ij}$  denotes the  $(i, j)^{th}$  element of a matrix. The mean  $\langle h_j \rangle$  and variance  $\text{var}(h_j)$  of the distributions  $q(h_j)$  are given by [162]

$$(6.48) \quad \langle h_j \rangle = \hat{h}_j + \sqrt{\frac{2}{\pi \tilde{h}_j}} \frac{1}{\text{erfcx}(-\hat{h}_j \sqrt{\frac{\tilde{h}_j}{2}})},$$

$$(6.49) \quad \text{var}(h_j) = \tilde{h}_j^{-1} + \sqrt{\frac{1}{\pi \tilde{h}_j}} \frac{\hat{h}_j}{\text{erfcx}(-\hat{h}_j \sqrt{\frac{\tilde{h}_j}{2}})},$$

where  $\text{erfcx}(\cdot)$  is the scaled complementary error function.

In the next step, we calculate the distributions of the hyperparameters from (6.35) as

$$(6.50) \quad q(\alpha_{im}) = \text{Gamma}(\alpha_{im} | \bar{a}_{\alpha_{im}}, \bar{b}_{\alpha_{im}})$$

$$(6.51) \quad q(\beta_1) = \text{Gamma}(\beta_1 | \bar{a}_{\beta_1}, \bar{b}_{\beta_1})$$

$$(6.52) \quad q(\beta_2) = \text{Gamma}(\beta_2 | \bar{a}_{\beta_2}, \bar{b}_{\beta_2})$$

$$(6.53) \quad q(\beta_{12}) = \text{Gamma}(\beta_{12} | \bar{a}_{\beta_{12}}, \bar{b}_{\beta_{12}})$$

$$(6.54) \quad q(\sigma_{jd}) = \text{Gamma}(\sigma_{jd} | \bar{a}_{\sigma_{jd}}, \bar{b}_{\sigma_{jd}})$$

$$(6.55) \quad q(\{\tau_{jd}\}_{d=1}^D) = \text{Dirichlet}(\{\tau_{jd}\}_{d=1}^D | \{\bar{c}_{\tau_{jd}}\}_{d=1}^D).$$

Note that the posterior distribution approximations have the same shapes as their corresponding prior distributions due to the use of conjugate priors. We utilize the means of these distributions

as their estimates, which are given by

$$(6.56) \quad \langle \alpha_{im} \rangle = \frac{\bar{b}_{\alpha_{im}}}{\bar{a}_{\alpha_{im}}} = \frac{b_{\alpha_{im}}^o + \frac{N}{2}}{a_{\alpha_{im}}^o + \sum_j \sqrt{w_j}}$$

$$(6.57) \quad \langle \beta_1 \rangle = \frac{\bar{b}_{\beta_1}}{\bar{a}_{\beta_1}} = \frac{b_{\beta_1}^o + \frac{N}{2}}{a_{\beta_1}^o + \frac{1}{2} \langle \| \mathbf{y}_1 - \mathbf{Hx} \|^2 \rangle}$$

$$(6.58) \quad \langle \beta_2 \rangle = \frac{\bar{b}_{\beta_2}}{\bar{a}_{\beta_2}} = \frac{b_{\beta_2}^o + \frac{N}{2}}{a_{\beta_2}^o + \frac{1}{2} \langle \| \mathbf{y}_2 - \mathbf{x} \|^2 \rangle}$$

$$(6.59) \quad \langle \beta_{12} \rangle = \frac{\bar{b}_{\beta_{12}}}{\bar{a}_{\beta_{12}}} = \frac{b_{\beta_{12}}^o + \frac{N}{2}}{a_{\beta_{12}}^o + \frac{1}{2} \langle \| \mathbf{y}_1 - \mathbf{Hy}_2 \|^2 \rangle}$$

$$(6.60) \quad \langle \sigma_{jd} \rangle = \frac{\bar{b}_{\sigma_{jd}}}{\bar{a}_{\sigma_{jd}}} = \frac{b_{\sigma_{jd}}^o + \mu_{jd}}{a_{\sigma_{jd}}^o + \mu_{jd} h_j}, \quad j = 1, \dots, M, d = 1, \dots, D$$

$$(6.61) \quad \bar{c}_{\tau_{jd}} = c_{\tau_{jd}}^o + \mu_{jd}$$

$$(6.62) \quad \langle \tau_{jd} \rangle = \frac{\bar{c}_{\tau_{jd}}}{\sum_{d=1}^D \bar{c}_{\tau_{jd}}}$$

Finally, the auxiliary variables  $\mu_{jd}$  are computed by first taking the expectation of (6.34) with respect to  $h_j$  and  $\sigma_{jd}$ , and then maximizing it with respect to the auxiliary variables. This results in the following update

$$(6.63) \quad \mu_{jd} \propto \langle \tau_{jd} \rangle \text{Expon}(\langle h_j \rangle | \langle \sigma_{jd} \rangle), \quad j = 1, \dots, M$$

with the condition

$$(6.64) \quad \sum_{d=1}^D \mu_{jd} = 1, \quad j = 1, \dots, M$$

The proposed algorithm is summarized in Algorithm 7. The explicit forms of the expectations  $\langle \mathbf{H}^T \mathbf{H} \rangle$  in (6.38),  $\langle (\Delta_j^h(\mathbf{x}))^2 + (\Delta_j^v(\mathbf{x}))^2 \rangle$  in (6.39), and  $\langle \| \mathbf{y}_1 - \mathbf{Hx} \|^2 \rangle$  in (6.57),

Table 6.1. Proposed algorithm

**Algorithm 7.** Bayesian Blind Deconvolution From Short- and Long-Exposure Image Pairs

Set initial image estimate  $\langle \mathbf{x} \rangle^{(0)} = \mathbf{y}_1$

Calculate initial estimates of  $\langle h_j \rangle$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_{12}$ ,  $\alpha_{\text{im}}$ ,  $\{\sigma_{jd}\}$  and  $\{\tau_{jd}\}$  using  $\langle \mathbf{x} \rangle^{(0)}$ ,  $\mathbf{y}_1$ , and  $\mathbf{y}_2$ .

For  $k = 1, 2, \dots$  until convergence:

- (1) Find image distribution  $q^k(\mathbf{x})$  using (6.37)-(6.38)
- (2) Find blur PSF coefficient distributions  $q^k(h_j)$  using (6.46), (6.47) and (6.48).
- (3) Find hyperparameter estimates using (6.56)-(6.62)
- (4) Find auxiliary variables  $\{\mu_{jd}\}$  using (6.63)

$\langle \| \mathbf{y}_2 - \mathbf{x} \|^2 \rangle$  in (6.58), and  $\langle \| \mathbf{y}_1 - \mathbf{H}\mathbf{y}_2 \|^2 \rangle$  in (6.59) are given by

$$(6.65) \quad \langle \mathbf{H}^T \mathbf{H} \rangle = \langle \mathbf{H} \rangle^T \langle \mathbf{H} \rangle + \Upsilon_{\mathbf{h}},$$

(6.66)

$$\langle (\Delta_j^h(\mathbf{x}))^2 + (\Delta_j^v(\mathbf{x}))^2 \rangle = (\Delta_j^h(\langle \mathbf{x} \rangle))^2 + (\Delta_j^v(\langle \mathbf{x} \rangle))^2 + \text{trace} \left( \Sigma_{\mathbf{x}} \left( (\Delta_j^h)(\Delta_j^h)^T + (\Delta_j^v)(\Delta_j^v)^T \right) \right),$$

$$\langle \| \mathbf{y}_1 - \mathbf{H}\mathbf{x} \|^2 \rangle = \| \mathbf{y}_1 - \langle \mathbf{H} \rangle \langle \mathbf{x} \rangle \|^2 + \text{trace} \left( \langle \mathbf{H} \rangle^T \langle \mathbf{H} \rangle \Sigma_{\mathbf{x}} \right)$$

$$(6.67) \quad + \text{trace} \left( \Upsilon_{\mathbf{h}} \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^T \right) + \text{trace} \left( \Sigma_{\mathbf{x}} \Upsilon_{\mathbf{h}} \right),$$

(6.68)

$$\langle \| \mathbf{y}_2 - \mathbf{x} \|^2 \rangle = \| \mathbf{y}_2 - \langle \mathbf{x} \rangle \|^2 + \text{trace} (\Sigma_{\mathbf{x}}),$$

(6.69)

$$\langle \| \mathbf{y}_1 - \mathbf{H}\mathbf{y}_2 \|^2 \rangle = \| \mathbf{y}_1 - \langle \mathbf{H} \rangle \mathbf{y}_2 \|^2 + \text{trace} (\Upsilon_{\mathbf{h}} \mathbf{Y}_2 \mathbf{Y}_2^T),$$

where

$$(6.70) \quad \Upsilon_{\mathbf{h}} = \text{diag} \left( \sum_{j=1}^M \text{var}(h_j) \right),$$

with  $\text{var}(h_j)$  given in (6.49).

Note that the explicit calculation of the covariance matrix  $\Sigma_{\mathbf{x}}$  is only needed in (6.66), (6.67), and (6.68), which is impractical due to its huge size of  $N \times N$ . As mentioned above, the image estimate in (6.37) is calculated using a conjugate gradient method, which does not require  $\Sigma_{\mathbf{x}}$  to be constructed. In order to avoid the high computational complexity of computing  $\Sigma_{\mathbf{x}}$  in (6.66), (6.67), and (6.68), in these equations we approximate  $\Sigma_{\mathbf{x}}$  as a diagonal matrix using the inverses of the diagonal elements of  $\Sigma_{\mathbf{x}}^{-1}$  in (6.38). We have conducted extensive experiments with small images which permit the explicit construction of  $\Sigma_{\mathbf{x}}$  to verify the validity of this approximation. We found out empirically that this approximation results in very similar estimates and has a minor effect in the estimation process. Note that similar approximations have also been utilized in other Bayesian recovery methods [228, 10, 118].

Finally we note the following. In the proposed framework, the distributions of the latent variables are estimated instead of their point estimates, which has several advantages. First, the uncertainty of the estimates can be calculated by examining the variances of the estimated distributions. Second, these uncertainties are incorporated into the estimation procedure using the expectations given in (6.65)-(6.69), so that when estimating an unknown, the algorithm accounts for the possible errors in the estimates of other variables. Note that if lower computational complexity is desired, degenerate distributions can be assumed for all unknowns, that is, all distributions are approximated by delta functions placed at their modes. This results in setting all covariances in the proposed method equal to zero, and it is equivalent to providing

*maximum a posteriori* (MAP) estimates of the unknowns. Finally, note that instead of using the means of the distributions as the estimates of the unknowns, one can apply a sampling algorithm to draw different values from the distributions. Although this approach will lead to a much slower procedure, it can be used to avoid local minima in cases where initial estimates of the unknowns are far from the desired solutions and the degradations are extremely severe.

It should also be noted that the proposed framework can easily be extended to handle a higher number of input images with possibly more than one blurred image. The proposed framework provides the main mathematical basis for more general cases with multiple input images and multiple blur PSFs, and the resulting algorithms are very similar to the one presented in this work.

## 6.6. Experimental Results

In this section, we demonstrate the performance of the proposed method with experiments with synthetic and real images. We first present experiments with synthetically generated degraded image pairs, to demonstrate the accuracy of the estimation of the unknown image and blur. We then show the application of the proposed method to real degraded image pairs and compare it to existing methods.

The following algorithm and experimental setup is utilized in all experiments. The observed image  $\mathbf{y}_1$  is used as the initial estimate of  $\mathbf{x}$ . The initial estimate of  $\mathbf{h}$  is obtained from (6.46) with the covariance matrix  $\Sigma_{\mathbf{x}}$  set equal to zero. This initial estimate corresponds to the maximum likelihood estimate of  $\mathbf{h}$ , and it can be obtained efficiently in the frequency domain by taking the ratio of the Fourier transforms of the observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$ . Although the initial PSF calculated in this fashion is a very crude estimate of the true PSF, it provides a very fast

initialization of the algorithm. The algorithm is able to provide very accurate results even with this crude initialization. The blur support  $M$  is chosen as the smallest support that covers the most significant entries of the initial PSF estimate. This operation can be performed manually or by simple thresholding followed by a convex hull algorithm. The blur support  $M$  chosen in this fashion is generally much smaller than the image support  $N$ , which improves the computational complexity of the method. Note that similar methods have been utilized by existing deconvolution methods, where the blur support is generally selected manually.

In all experiments, the number of mixture distributions is set to  $D = 3$ , but other values ( $D = 2$  or  $D = 4$ ) gave similar results. Utilizing single exponential distributions per pixel ( $D = 1$ ) generally resulted in less sparse PSF estimates with higher estimation noise. All other parameters are calculated using (6.56)-(6.63). Note that except possibly the PSF support  $M$ , all required parameters of the algorithm are initialized automatically. As convergence criterion we use  $\| \langle \mathbf{x} \rangle^{(k)} - \langle \mathbf{x} \rangle^{(k-1)} \|_2^2 / \| \langle \mathbf{x} \rangle^{(k-1)} \|_2^2 < 10^{-5}$ , where  $\langle \mathbf{x} \rangle^{(k)}$  and  $\langle \mathbf{x} \rangle^{(k-1)}$  are the image estimates at iterations  $k$  and  $k - 1$ , respectively. The convergence is generally achieved within 20 iterations, where each iteration takes approximately 20 seconds using our non-optimized Matlab code running on a Pentium Core2 CPU at 2.66 GHz, depending on the severity of the degradations in the input images.

For the synthetic image degradations, the image shown in Fig. 6.2(a) is used to create the observed images  $\mathbf{y}_1$  and  $\mathbf{y}_2$ . The observed image  $\mathbf{y}_2$ , shown in Fig. 6.2(b), is obtained by adding white Gaussian noise of variance 220 to the original image (SNR = 7 dB). This image suffers from a very high level of noise: the mean-squared-error (MSE) between this image and the original image is 196.20. We create five different observations  $\mathbf{y}_1$  by blurring the original image by five different blur PSFs shown at the bottom row of Fig. (6.3), which are typical examples

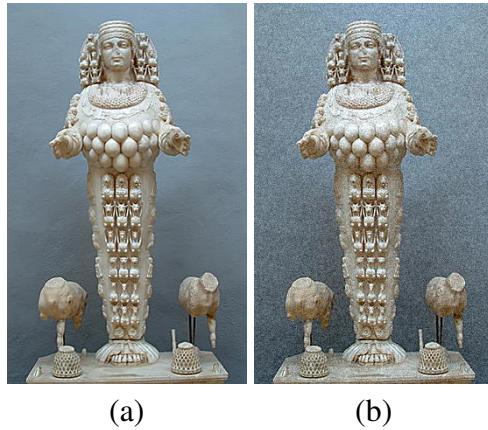


Figure 6.2. (a) Original Ephesus image, (b) observed noisy image simulating a short-exposure acquisition.

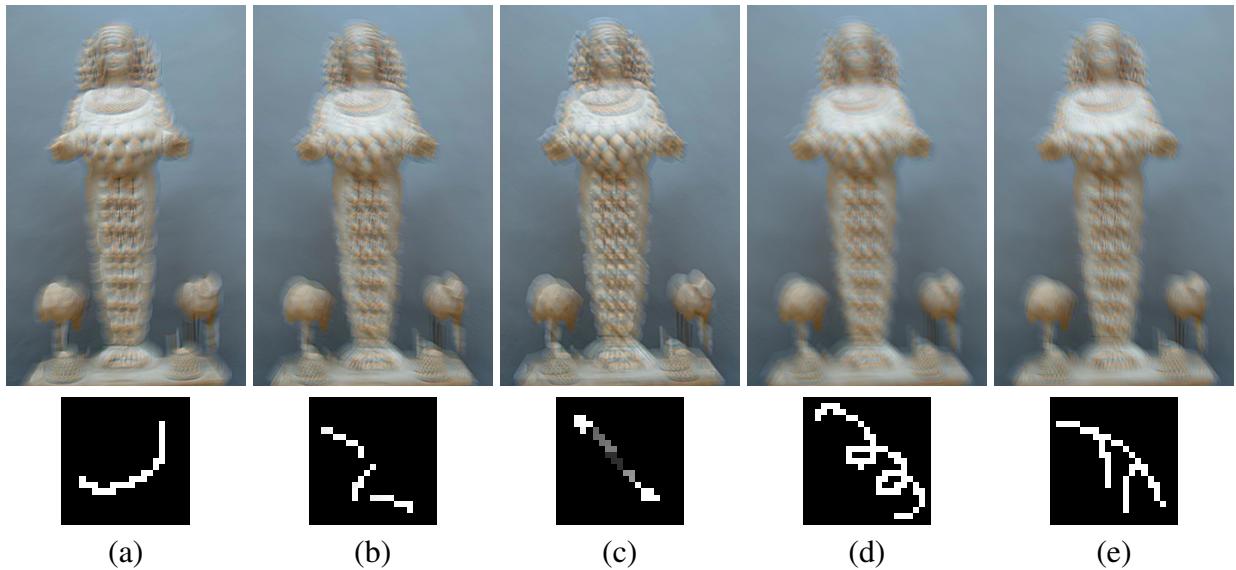


Figure 6.3. Blurred images simulating long-exposure photographs. The point spread function (PSF) used to generate each image is shown below the corresponding image. All PSFs have support  $21 \times 21$  pixels and the images are of size  $430 \times 270$  pixels. The values of the PSFs are linearly mapped to the  $[0,255]$  range for visualization purposes.

of PSFs resulting from the motion of the camera during long exposures. White Gaussian noise with a variance of 0.16 is added to the blurred images to obtain the final observed images

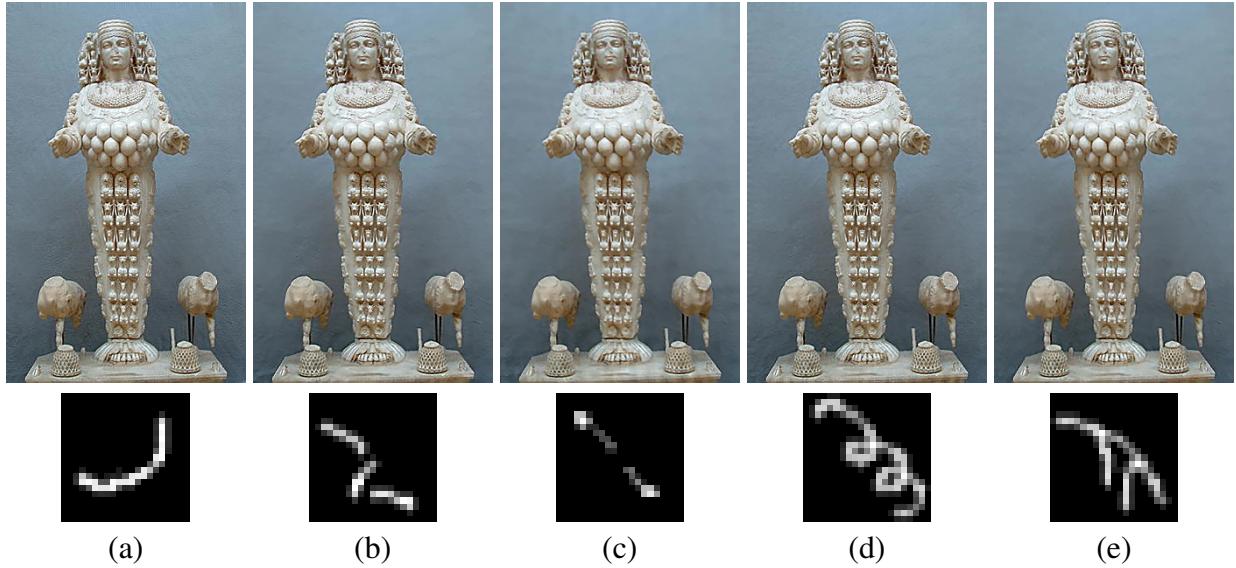


Figure 6.4. Restoration results using the proposed algorithm. The restored images are shown in the top row, and the corresponding recovered PSFs are shown below the images. The values of the PSFs are linearly mapped to the [0,255] range for visualization purposes.

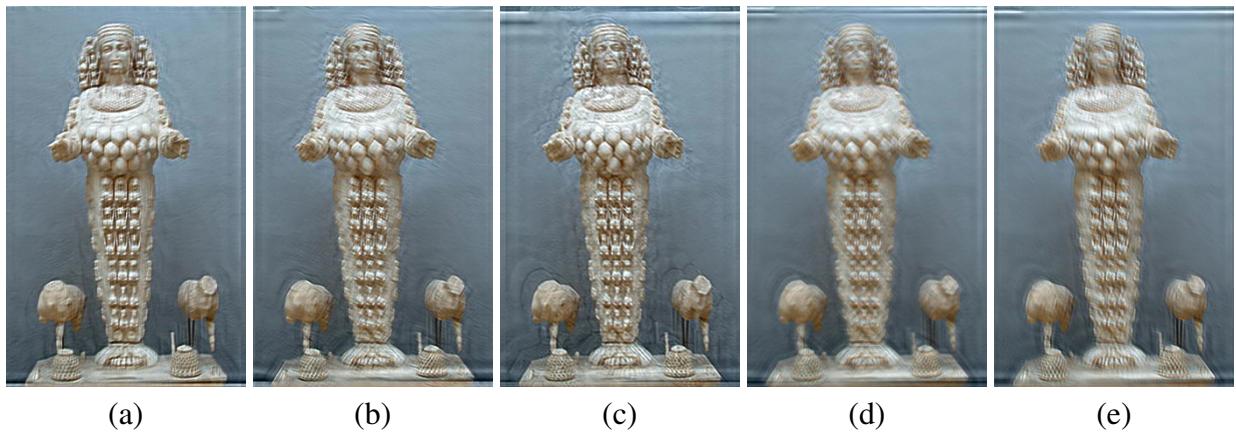


Figure 6.5. Restoration results using the *deconvblind* routine in MATLAB.

$\mathbf{y}_1$  with signal-to-noise-ratios (SNR) of 40 dB, which are shown at the top row of Fig. (6.3). Note that although the noise level is low, as is typically the case in long-exposure images, the images are severely degraded by PSFs with large-supports compared to the image size. The

MSEs between the original image and the observed images are 485.20, 451.50, 508.52, 412.44, 417.51, respectively. The support of the PSFs used in this experiment are  $21 \times 21$ , and the original image is of size  $430 \times 270$ .

Each observed image  $\mathbf{y}_1$  in Fig. (6.3) along with the observed image  $\mathbf{y}_2$  in Fig. 6.2(b) is provided to the proposed algorithm as an image pair. The restored images obtained by the proposed method corresponding to different PSFs are shown at the top row of Fig. (6.4), and the recovered PSF for each case is shown at the bottom row of Fig. (6.4). By comparing with the original image in Fig. 6.2(a), it is clear that the proposed algorithm provides restored images with very high visual quality in all cases. The mean-squared-error (MSE) between the original image and the restored images are 13.32, 24.9, 42.76, 17.68, and 23.14, respectively, which shows that the restored images are very close to the original image. The corresponding MSEs between the original and recovered PSFs are  $9 \times 10^{-7}$ ,  $8 \times 10^{-6}$ ,  $5 \times 10^{-6}$ ,  $9 \times 10^{-7}$  and  $3 \times 10^{-6}$ . Both quantitative MSE results and visual inspection of the recovered PSFs suggest that the algorithm is very successful in estimating the original PSFs. Moreover, note that the recovered images do not exhibit any deconvolution artifacts such as ringing or noise amplification, due to the accurate estimation of the PSFs and the spatially-varying smoothing due to the total variation prior.

We also compare the results of the proposed algorithm with the classical single-image blind deconvolution algorithm implemented by the *deconvblind* routine in MATLAB, which utilizes a modified form of the Richardson-Lucy algorithm [199, 149]. To achieve the best possible restoration results we provided the algorithm with the PSF estimates obtained by the proposed method shown at the bottom row of Fig. (6.3), to be used as initial PSF estimates. Even with this unrealistic scenario, the quality of the resulting restored images, shown in Fig. (6.5), is much

lower than that of the proposed method. The corresponding MSEs between the original and restored images are 135.21, 182.87, 198.27, 247.43, and 224.28, respectively. Note also that in most cases the blur is not completely removed, and significant ringing artifacts are present in the restored images. On the other hand, the proposed algorithm provides restored images of very high quality.

Next we apply the proposed method to a real image pair acquired by a compact digital camera. Figures 6.6(a)-(b) show an image pair acquired outdoors at ISO = 800 with exposure times 1/100 and 1/3 seconds, respectively. Due to low light conditions the noise level in the short exposure image is quite high. Moreover, it can be observed from Figure 6.6(a) that the noise in the short exposure image is colored and therefore does not follow the assumed observation model in (6.4). In addition, certain parts of the images are highly saturated (e.g., part of the window above the door), which introduces an additional difficulty in blur estimation due to its nonlinearity. The application of the denoising algorithm in [192] to the short exposure image is shown in Figs. 6.6(c). We manually tuned the parameters of the denoising algorithm for each color channel, and show the result with the highest visual quality. The image and blur estimates provided by the proposed algorithm are shown in Figs. 6.6(d)-(de), respectively. The support of the PSF estimate is  $51 \times 51$ . The center regions of the images are shown in their original size in Fig. 6.7 for a closer inspection. It can be observed that despite the challenging nature of the input images, the algorithm provides significant improvement both in removing the blur and revealing sharp details in the image, as well as, in correcting the color loss apparent in the short-exposure image. Although the denoising method is also successful in removing the acquisition noise, it can not correct the color loss in the short exposure image, and its result is softer than the one provided by the proposed method.

Finally, we present a comparison of the proposed method with the image stabilization method proposed in [224] and the denoising algorithm in [192] on a real image set. The images shown in Fig. 6.8(a) and Fig. 6.8(b) (published in [224]) are taken with exposure times 1/100 sec and 1 sec, respectively. The result of applying the denoising algorithm in [192] to the short exposure image is shown in Fig. 6.8(c). It is clear that although the noise level is significantly reduced, the contrast is very low, and there is a significant red color cast. The restored image in Fig. 6.8(d) is obtained by the algorithm in [224], which requires knowledge of the variances of the noise in the observations. Finally, the PSF and image estimates provided by the proposed algorithm are shown in Fig. 6.8(d) and (e), respectively, where the estimated support of the PSF is  $41 \times 41$ . Note that although the proposed method is fully-automated, the restored image is clearly sharper than that of [224] with almost no ringing artifacts. This is especially evident in the area around the letters. Moreover, the restored image by the proposed method is sharper than the denoised image and has a higher contrast with correctly restored colors.

In summary, experimental results with both synthetic and real image sets demonstrate that the proposed algorithm is very effective in providing high quality restored images, although no image- and observation-specific parameter tuning has been performed.

## 6.7. Conclusions

In this chapter we presented a novel Bayesian formulation for blind deconvolution from image pairs acquired using long- and short-exposure times. The unknown image, blur and all model parameters, including the noise variances, are estimated solely from the observations without prior knowledge or user intervention. On the other hand, the proposed framework is

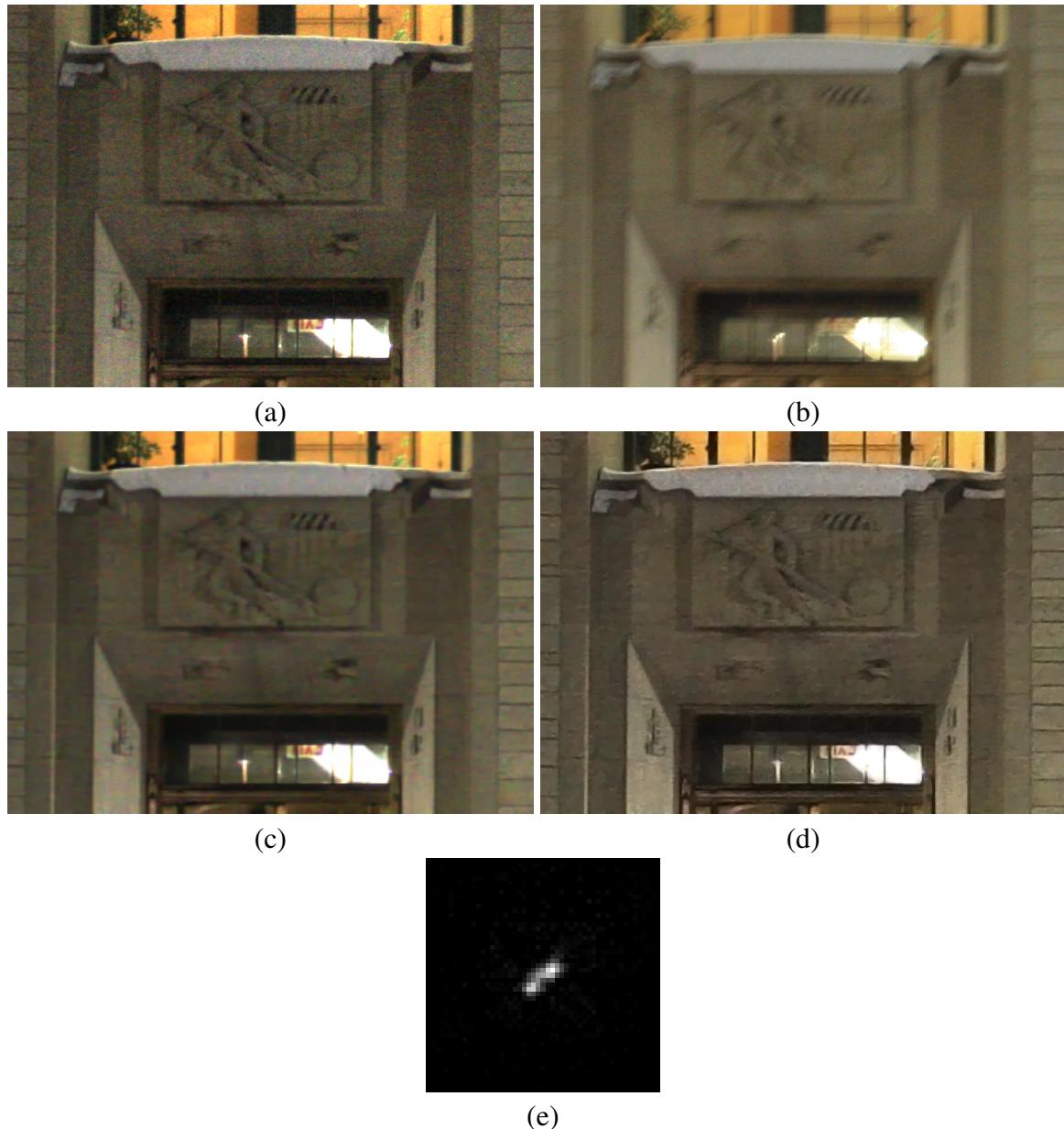


Figure 6.6. An outdoor image pair. (a) Short exposure image (brightness level is corrected), (b) long exposure image, (c) denoised short exposure image, (d) restored image using the proposed algorithm, and (e) recovered PSF (support:  $51 \times 51$ ).

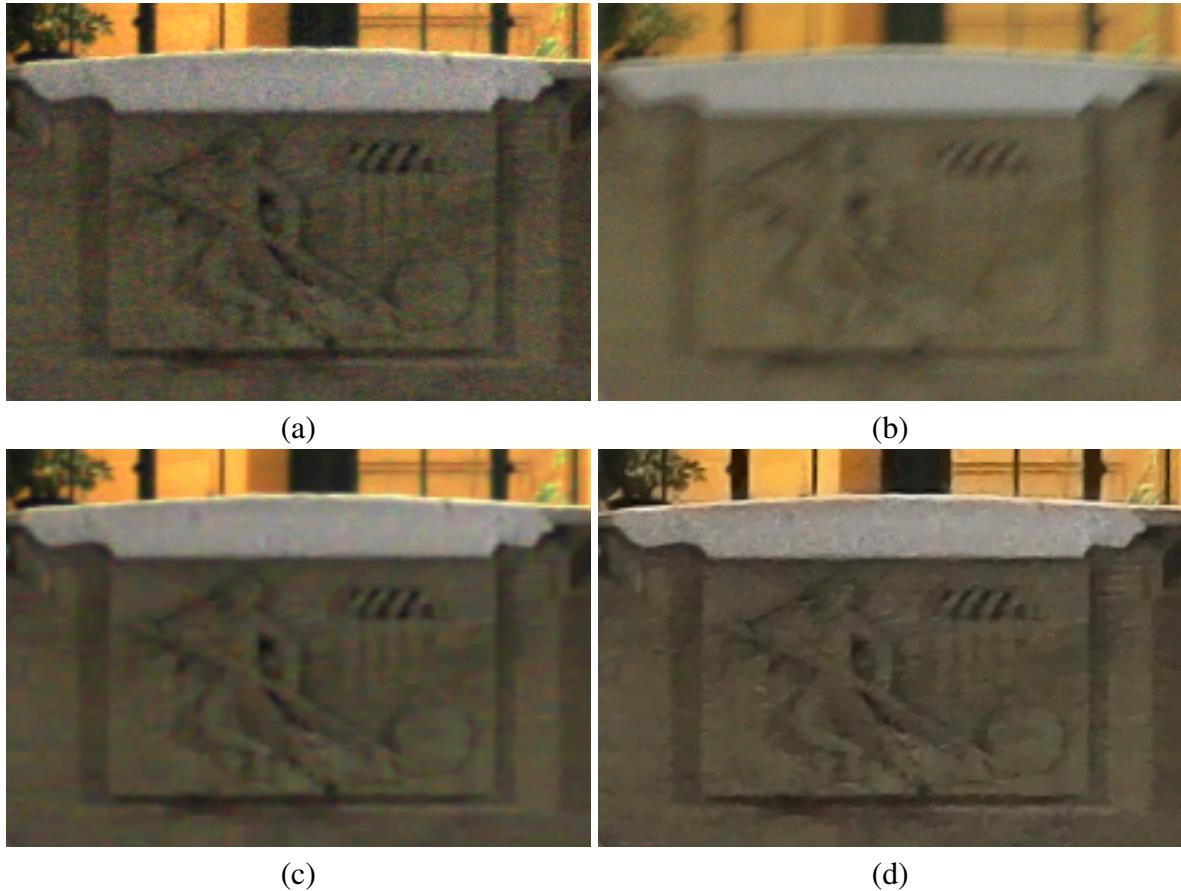


Figure 6.7. Center regions of the images shown in Figure 6.6. (a) Short exposure image, (b) long exposure image, (c) denoised short exposure image, (d) result of the proposed algorithm.

very flexible so that when some prior knowledge about the unknowns is available, it can easily be incorporated into the algorithm. The developed algorithm simultaneously estimates the distributions of the unknowns which allows for the computation of the estimation uncertainties and also incorporates these uncertainties into the restoration procedure. The algorithm does not rely on non-robust and input-dependent ad hoc methods (such as blur thresholding or blur denoising). Moreover, although the proposed method does not require user-intervention but instead provides a fully automated estimation of the algorithmic parameters, experimental results

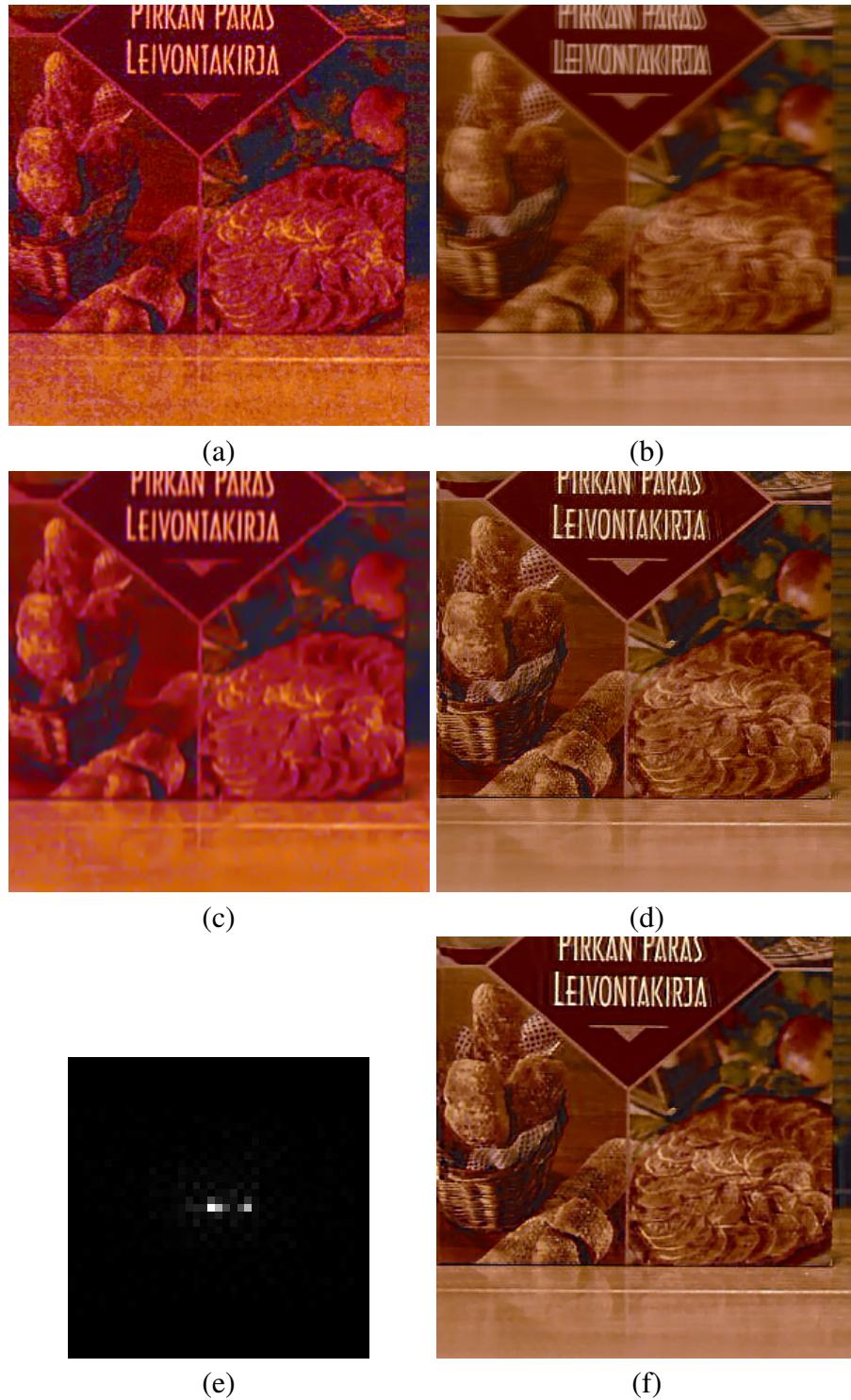


Figure 6.8. Real image example (courtesy of [224]). (a) Short exposure image, (b) long exposure image, (c) denoised short exposure image, (d) restored image using [224], (e) recovered PSF using the proposed algorithm (support :  $41 \times 41$ ), and (f) restored image using the proposed algorithm.

demonstrate that it results in very high quality restored images even with high degradations in both synthetic and real image cases, and compares favorably with existing methods.

Future work includes the incorporation of the geometric and photometric registration in the Bayesian framework. This will allow simultaneous estimation of the registration parameters along with the unknown image and the blur, which has the potential of providing more accurate registration estimates than preprocessing methods.

## CHAPTER 7

# Variational Bayesian Super Resolution

### 7.1. Introduction

Super resolution methods (see [128, 119, 176] and Sec 2.4 for reviews) generally consist of two parts: Registration, where the motion between the LR images are estimated; and image estimation, where the HR image is recovered from the LR images using information about the motion and blurring. Many conventional methods in the literature assume that the motion information is known *a priori*. However, this assumption does not hold in many practical systems since exact motion information is very hard to obtain and implement. Therefore, a registration step is needed to obtain the motion parameters from the LR images.

Super resolution is a highly ill-posed problem, especially when the motion parameters are estimated along with the HR image solely from the LR images. The registration parameters are generally very hard to estimate using only LR observations, which makes estimation errors unavoidable in many practical systems. The errors in estimating the registration parameters cause significant drawbacks in super resolution, causing instabilities in the recovery of the HR image and significantly affecting the robustness of the restoration procedures.

A number of approaches have been proposed to address this problem, which can be classified into two major categories based on the stage where the registration is performed. The

---

<sup>0</sup>This work has appeared in [17, 11].

first class of methods employ registration as a preprocessing stage [76, 257, 79, 234]. The motion parameters are estimated from the observed LR images, and then used in a separate image estimation process. Since the motion parameters estimated only using the LR images can be unreliable, a desired property is robustness to outliers and errors in motion estimates. A robust backprojection method is proposed in [257] based on median estimators. Farsiu *et. al.* [79] proposed to use an observation model based on  $l_1$ -norms and image priors based on bilateral total-variation (BTv) functions, whose combination makes the algorithm robust to motion outliers. Other methods employ regularization by modeling the registration errors as Gaussian noise [140, 108]. All methods in this category attempt to reduce the effect of estimation errors and noise by decreasing the weight of unreliable observations in the restoration process, but they do not attempt to correct the errors in the motion estimation process.

Another class of SR methods estimate both the HR image and the motion parameters simultaneously. The most common approach in this category is alternating minimization (AM), where at each iteration, the estimates of the HR image and the motion parameters are improved progressively in an alternating fashion [105, 177, 210, 218, 249, 118, 228, 191]. Some methods in this category also employ explicit models of the errors in motion estimates. In [228] and [191], the errors in motion and blur parameters are assumed to follow Gaussian distributions. In [228], the HR image is marginalized out from the joint distribution and the motion and blur parameters are estimated from this marginal distribution. A major disadvantage of this method is that the marginalization of the HR image requires the utilization of a Gaussian image prior, which overpenalizes strong image edges and therefore reduces the quality of the estimated HR image. In [191], this problem is overcome by marginalizing the motion and blur parameters, and employing a Huber prior to model the HR image. Recently, a joint identification method

is proposed in [109] where the optimization problem is solved simultaneously for both the HR image and motion parameters. Finally, methods which do not utilize explicit knowledge of the motion estimation parameters have been proposed in [221, 193].

A major drawback of most super resolution methods is that they employ a number of algorithmic parameters that need to be tuned. This tuning process can be cumbersome and time-consuming since the parameter values have to be chosen differently for each image and degradation condition. Moreover, the algorithm performance depends significantly on the appropriate choice of parameters, such that generally a long supervision process is needed to obtain useful results.

In this chapter, we develop novel Bayesian super resolution methods which address both of the above mentioned issues. We provide a systematic modeling of the unknown HR image and the motion parameters within a novel hierarchical Bayesian formulation, and develop SR algorithms which jointly estimate the HR image and the motion. Through the utilization of variational Bayesian analysis, the proposed framework provides uncertainties of the estimates during the restoration process, which helps in preventing error-propagation and improves robustness. All required algorithmic parameters are estimated along with the HR image and the motion parameters, and therefore algorithms do not require user super vision. Moreover, the algorithmic parameters are estimated optimally in a stochastic sense, which provides high reconstruction performance. We show that the proposed methods are very robust to errors in motion estimates due to adaptive parameter and motion estimation. We demonstrate with experimental results that the proposed methods provide HR images with high quality and accurate motion information, and compare favorably to existing SR methods.

The rest of this chapter is organized as follows. Section 7.2 provides the mathematical model for the LR image acquisition process. We provide the description of the hierarchical Bayesian framework modeling the unknowns in Section 7.3. The inference procedure to develop the proposed methods is presented in Section 7.4. We demonstrate the effectiveness of the proposed methods with experimental results in Section 7.5 and conclusions are drawn in Section 7.6.

## 7.2. Problem Formulation

The imaging process is assumed to have generated  $L$  LR images  $\mathbf{y}_k$ ,  $k = 1, \dots, L$ , from the HR image  $\mathbf{x}$ . The LR images  $\mathbf{y}_k$  and the HR image  $\mathbf{x}$  consist of  $N$  and  $PN$  pixels, respectively, where  $P$  is the factor of increase in resolution. We adopt the matrix-vector notation such that the images  $\mathbf{y}_k$  and  $\mathbf{x}$  are arranged as  $N \times 1$  and  $PN \times 1$  vectors, respectively. The imaging process introduces shifting, blurring and downsampling, which is modeled as a linear space-invariant system as

$$(7.1) \quad \mathbf{y}_k = \mathbf{A}\mathbf{H}_k\mathbf{C}(\mathbf{s}_k)\mathbf{x} + \mathbf{n}_k = \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} + \mathbf{n}_k,$$

where  $\mathbf{A}$  is the  $N \times PN$  downsampling matrix,  $\mathbf{H}_k$  is the  $PN \times PN$  blurring matrix,  $\mathbf{C}(\mathbf{s}_k)$  is the  $PN \times PN$  warping matrix generated by the motion vector  $\mathbf{s}_k$ , and  $\mathbf{n}_k$  is  $N \times 1$  the acquisition noise. Note that the matrices  $\mathbf{H}_k$  and  $\mathbf{C}(\mathbf{s}_k)$  and the noise  $\mathbf{n}_k$  can be different for each LR image  $\mathbf{y}_k$ . We assume that the blurring matrices  $\mathbf{H}_k$  are known.

In this work we assume that the motion vectors  $\mathbf{s}_k$  are not known, so they have to be estimated along with the HR image  $\mathbf{x}$ . We consider a motion model consisting of translational and rotational motion, so that  $\mathbf{s}_k = (\theta_k, c_k, d_k)^T$ , where  $\theta_k$  is the rotation angle, and  $c_k$  and  $d_k$  are horizontal and vertical translations of the  $k^{th}$  LR image, respectively. This motion model

is quite general as opposed to many existing SR methods assuming only translational motion. Additionally, as will be shown later, the proposed framework can easily be extended to more complex motion models such as affine and projective motion.

Finally, the effects of downsampling, warping, and blurring can be combined into a single  $N \times PN$  system matrix  $\mathbf{B}_k(\mathbf{s}_k)$ , such that each row in matrix  $\mathbf{B}_k(\mathbf{s}_k)$  maps the pixels in the HR image  $\mathbf{x}$  to one pixel in the LR image  $\mathbf{y}_k$ . Given (7.1), the super resolution problem is to find an estimate of the HR image  $\mathbf{x}$  from the set of LR images  $\{\mathbf{y}_k\}$  and using prior knowledge about  $\{\mathbf{C}(\mathbf{s}_k)\}$ ,  $\{\mathbf{n}_k\}$ , and  $\mathbf{x}$ .

### 7.3. Hierarchical Bayesian Model

In order to obtain high quality estimates of  $\mathbf{x}$  and  $\{\mathbf{s}_k\}$  from  $\{\mathbf{y}_k\}$ , properties of the unknowns and the acquisition process have to be taken into account. In Bayesian models, the incorporation of prior knowledge is achieved by treating all unknown parameters as stochastic quantities and by assigning probability distributions to them. These distributions are used to reflect prior knowledge onto the estimation process.

In this work, we adopt a hierarchical Bayesian framework consisting of two stages. The first stage is used to model the acquisition process, the unknown HR image  $\mathbf{x}$ , and the motion vectors  $\{\mathbf{s}_k\}$ . The unknowns  $\mathbf{x}$  and  $\mathbf{s}_k$  are assigned *prior* distributions  $p(\mathbf{x}|\alpha_{im})$  and  $p(\mathbf{s}_k)$ , respectively. The observation  $\mathbf{y} = \{\mathbf{y}_k\}$  is also a random process with the corresponding *conditional* distribution  $p(\mathbf{y}|\mathbf{x}, \{\mathbf{s}_k\}, \{\beta_k\})$ . These distributions depend on additional parameters  $\alpha_{im}$  and  $\{\beta_k\}$  (called *hyperparameters*), which are modeled by assigning *hyperprior* distributions in the second stage of the hierarchical model.

In the following subsections we provide the description of individual distributions used to model the unknowns.

### 7.3.1. Observation Model

Using the model in (7.1) and assuming that  $\mathbf{n}_k$  is zero-mean white Gaussian noise with the inverse variance (precision)  $\beta_k$ , the conditional distribution of the LR image  $\mathbf{y}_k$  is given by

$$(7.2) \quad p(\mathbf{y}_k | \mathbf{x}, \mathbf{s}_k, \beta_k) \propto \beta_k^{N/2} \exp \left[ -\frac{\beta_k}{2} \| \mathbf{y}_k - \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} \|^2 \right].$$

Assuming statistical independence of the noise between the LR image acquisitions, the conditional probability of the set of LR images  $\mathbf{y}$  given  $\mathbf{x}$  can be expressed as

$$(7.3) \quad \begin{aligned} p(\mathbf{y} | \mathbf{x}, \{\mathbf{s}_k\}, \{\beta_k\}) &= \prod_{k=1}^L p(\mathbf{y}_k | \mathbf{x}, \mathbf{s}_k, \beta_k) \\ &= \left[ \prod_{k=1}^L \beta_k^{N/2} \right] \exp \left( -\frac{1}{2} \sum_{k=1}^L \beta_k \| \mathbf{y}_k - \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} \|^2 \right) \end{aligned}$$

The independent Gaussian model in (7.3) is used in most of the existing super resolution methods [140, 228, 108, 191, 228, 56, 118]. Some methods utilized  $l_1$ -norm based observation models which take both acquisition and registration noise into account [257, 79]. In this chapter, we use (7.3) to model only the acquisition noise. We incorporate an explicit modeling of the registration errors separately and therefore they are not taken into account in (7.3).

Let us now explicitly state the form of the matrices  $\mathbf{C}(\mathbf{s}_k)$ . We denote the coordinates of the reference HR grid by  $(u, v)$  and the coordinates of the  $k^{th}$  warped HR grid, after applying  $\mathbf{C}(\mathbf{s}_k)$

to  $\mathbf{x}$ , by  $(u_k, v_k)$ . Let us also define

$$(7.4) \quad \Delta u_k = u_k - u = u \cos(\theta_k) - v \sin(\theta_k) + c_k - u$$

$$(7.5) \quad \Delta v_k = v_k - v = u \sin(\theta_k) + v \cos(\theta_k) + d_k - v$$

Note that the coordinates  $(u_k, v_k)$  generally correspond to fractional values, and therefore the HR image value at pixel  $(u_k, v_k)$  in the  $k^{th}$  HR grid has to be calculated using resampling. As in [109], we incorporate bilinear interpolation to approximate the HR image value at  $(u_k, v_k)$  using the four neighboring HR image values  $x_{tl(s_k)}$ ,  $x_{tr(s_k)}$ ,  $x_{bl(s_k)}$  and  $x_{br(s_k)}$ , which are the pixels at the top-left, top-right, bottom-left and bottom-right locations of the pixel at  $(u_k, v_k)$ , respectively.

Let us denote by  $(a_k(s_k), b_k(s_k))^T$  the vector difference between the pixel at  $(u_k, v_k)$  and the pixel at the top-left position in the reference HR grid, that is,

$$(7.6) \quad a_k(s_k) = \Delta u_k - \text{floor}(\Delta u_k)$$

$$(7.7) \quad b_k(s_k) = \Delta v_k - \text{floor}(\Delta v_k)$$

Using bilinear interpolation, the warped image  $\mathbf{C}(\mathbf{s}_k)\mathbf{x}$  can be approximated as (see [109] for details)

$$(7.8) \quad \begin{aligned} \mathbf{C}(\mathbf{s}_k)\mathbf{x} \approx & \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)}(\mathbf{I} - \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)})\mathbf{L}_{\mathbf{bl}(s_k)}\mathbf{x} + \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)}\mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)}\mathbf{L}_{\mathbf{br}(s_k)}\mathbf{x} \\ & + (\mathbf{I} - \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)})(\mathbf{I} - \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)})\mathbf{L}_{\mathbf{tl}(s_k)}\mathbf{x} + (\mathbf{I} - \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)})\mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)}\mathbf{L}_{\mathbf{tr}(s_k)}\mathbf{x} \end{aligned}$$

where  $\mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)}$  and  $\mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)}$  denote diagonal matrices with the vectors  $\mathbf{a}_k(\mathbf{s}_k)$  and  $\mathbf{b}_k(\mathbf{s}_k)$  in their diagonal, respectively. The matrices  $L_z$  with  $z \in \{\mathbf{bl}(s_k), \mathbf{br}(s_k), \mathbf{tl}(s_k), \mathbf{tr}(s_k)\}$  are constructed

in such a way that the product  $L_{\mathbf{z}} \mathbf{x}$  produces pixels at the top-left, top-right, bottom-left and bottom-right locations of  $(u_k, v_k)$ , respectively.

### 7.3.2. Image Model

The quality of the estimated HR image as well as the accuracy in the estimates of other unknowns depends on incorporating good image models. As we have demonstrated in the previous chapters, TV priors are very effective in preserving the edges while imposing smoothness. Therefore, as the HR image prior, we again utilize the quadratic approximation of the TV prior

$$(7.9) \quad p(\mathbf{x} | \alpha_{\text{im}}) \propto \alpha_{\text{im}}^{PN/2} \exp \left[ -\frac{1}{2} \alpha_{\text{im}} \text{TV}(\mathbf{x}) \right],$$

where

$$(7.10) \quad \text{TV}(\mathbf{x}) = \sum_{i=1}^{PN} \sqrt{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2}.$$

The operators  $\Delta_i^h(\mathbf{x})$  and  $\Delta_i^v(\mathbf{x})$  correspond to, respectively, horizontal and vertical first order differences at pixel  $i$ .

### 7.3.3. Modeling the uncertainties in the registration parameters

Let us denote by  $\bar{\mathbf{s}}_k^P$  the estimate of  $\mathbf{s}_k$  obtained from LR observations in a preprocessing step, using registration algorithms such as [235, 148]. As mentioned earlier, these estimates are generally inaccurate, which lowers the image restoration quality. Therefore, we model the motion parameters as stochastic variables following Gaussian distributions with *a priori* means

set equal to the preliminary motion parameters  $\bar{\mathbf{s}}_k^p$ , that is,

$$(7.11) \quad p(\mathbf{s}_k) = \mathcal{N}(\mathbf{s}_k | \bar{\mathbf{s}}_k^p, \boldsymbol{\Lambda}_k^p)$$

with  $\boldsymbol{\Lambda}_k^p$  the *a priori* covariance matrix. The parameters  $\bar{\mathbf{s}}_k^p$  and  $\boldsymbol{\Lambda}_k^p$  incorporate prior knowledge about the motion parameters into the estimation procedure. If such knowledge is not available,  $\bar{\mathbf{s}}_k^p$  and  $(\boldsymbol{\Lambda}_k^p)^{-1}$  can be set equal to zero, which makes the observations solely responsible for the estimation process. Similar models utilizing Gaussian distributions to model the uncertainty in preliminary motion parameters have also been used in some existing algorithms [140, 228, 191], but with different inference methods.

#### 7.3.4. Hyperpriors on the Hyperparameters

The hyperparameters  $\alpha_{im}$  and  $\{\beta_k\}$  are crucial in determining the performance of the SR algorithm. For their modeling, we employ Gamma distributions

$$(7.12) \quad p(\omega) = \Gamma(\omega | a_\omega^o, b_\omega^o) = \frac{(b_\omega^o)^{a_\omega^o}}{\Gamma(a_\omega^o)} \omega^{a_\omega^o - 1} \exp[-b_\omega^o \omega],$$

where  $\omega > 0$  denotes a hyperparameter, and  $a_\omega^o > 0$  and  $b_\omega^o > 0$  are the shape and scale parameters, respectively. The hyperpriors are chosen as Gamma distributions since they are the conjugate priors for the variance of the Gaussian distribution, that is, they have the same functional form with the product of the prior distributions and the observation model [23].

Finally, combining (7.3), (7.9), (7.11) and (7.12), we obtain the joint probability distribution of all unknowns as

$$(7.13) \quad p(\mathbf{y}, \mathbf{x}, \{\mathbf{s}_k\}, \alpha_{im}, \{\beta_k\}) = p(\mathbf{y}|\mathbf{x}, \{\mathbf{s}_k\}, \{\beta_k\}) p(\mathbf{x}|\alpha_{im}) p(\mathbf{s}_k) p(\alpha_{im}) \prod_{k=1}^L p(\beta_k).$$

#### 7.4. Variational Bayesian Inference

Let us first denote the set of all unknowns by  $\Theta = \{\mathbf{x}, \{\mathbf{s}_k\}, \alpha_{im}, \{\beta_k\}\}$  for clarity. The Bayesian inference is based on the posterior distribution

$$(7.14) \quad p(\Theta | \mathbf{y}) = \frac{p(\Theta, \mathbf{y})}{p(\mathbf{y})},$$

However, as in many applications, this distribution is intractable, since  $p(\mathbf{y})$  cannot be computed. Therefore, approximation methods are utilized, some of which are evidence analysis (type-II maximum likelihood) and sampling methods [28]. In this work, we resort to a variational Bayesian analysis due to its certain advantages, including accounting for the uncertainties in the estimation processes and computational efficiency compared to the sampling approaches, among others.

In the variational Bayesian analysis, the posterior distribution  $p(\Theta | \mathbf{y})$  is approximated by a tractable distribution  $q(\Theta)$ . This approximating distribution is found by minimizing the Kullback-Leibler (KL) distance between  $q(\Theta)$  and the posterior  $p(\Theta | \mathbf{y})$ , given by

$$(7.15) \quad C_{KL}(q(\Theta) \| p(\Theta | \mathbf{y})) = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta | \mathbf{y})} \right) d\Theta = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta, \mathbf{y})} \right) d\Theta + \text{const.}$$

Generally, the only assumption made in variational Bayesian analysis is that the distribution  $q(\Theta)$  factorable [28, 21, 162]. In this work, we use the following factorization

$$(7.16) \quad q(\Theta) = q(\mathbf{x}, \{\mathbf{s}_k\}, \alpha_{im}, \{\beta_k\}) = q(\alpha_{im})q(\mathbf{x}) \prod_{k=1}^L q(\mathbf{s}_k) \prod_{k=1}^L q(\beta_k)$$

Unfortunately, the TV image prior makes the calculation of the KL distance very hard. As in the previous chapters, this difficulty is overcome by resorting to majorization-minimization (MM) approaches. In the following we present an outline of the MM approach as applied to the super resolution problem. A lower bound of the distribution in (7.13) can be found as follows. Let us first consider again the following inequality, which states that for real numbers  $a \geq 0$  and  $b > 0$

$$(7.17) \quad \sqrt{ab} \leq \frac{a+b}{2} \Rightarrow \sqrt{a} \leq \frac{a+b}{2\sqrt{b}}.$$

Next we define the functional  $\mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{w})$  with a  $PN$ -dimensional vector  $\mathbf{w} \in (R^+)^{PN}$ , with components  $w_i$ ,  $i = 1, \dots, PN$ , as follows

$$(7.18) \quad \mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{w}) = \alpha_{im}^{PN/2} \exp \left[ -\frac{\alpha_{im}}{2} \sum_i \frac{(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 + w_i}{\sqrt{w_i}} \right].$$

As will be clear later, the auxiliary variable  $\mathbf{w}$  is a quantity that needs to be computed and it has an interpretation related to the unknown HR image  $\mathbf{x}$ . Using  $a = (\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2$  and  $b = w_i$  in the inequality (7.17) it is easy to show that the functional  $\mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{w})$  is a lower bound of the image prior  $p(\mathbf{x}|\alpha_{im})$ , that is,

$$(7.19) \quad p(\mathbf{x}|\alpha_{im}) \geq \mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{w}).$$

This lower bound can be used to find a lower bound for the joint distribution in (7.13)

$$\begin{aligned}
 p(\mathbf{y}, \Theta) &\geq p(\mathbf{y}|\Theta)\mathbf{M}(\alpha_{im}, \mathbf{x}, \mathbf{w})(\alpha_{im}, \mathbf{x}, \mathbf{w})p(\mathbf{s}_k)p(\alpha_{im}) \prod_{k=1}^L p(\beta_k) \\
 (7.20) \quad &= \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{w}, \mathbf{y}),
 \end{aligned}$$

which results in an upper bound of the KL distance in (7.15) as

$$(7.21) \quad C_{KL}(q(\Theta) \| p(\Theta, \mathbf{y})) \leq C_{KL}(q(\Theta) \| \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{w}, \mathbf{y})).$$

In previous chapters, we have shown that the minimization of (7.15) can be replaced by the minimization of its upper bound (7.21), as minimizing this bound with respect to the unknowns and the auxiliary variable  $\mathbf{w}$  in an alternating fashion results in closer bounds at each iteration. The bound in (7.21) is quadratic and therefore is easy to analyze analytically. Utilizing this bound, the standard solutions of the variational Bayesian methods [28] can be used to estimate the unknown distributions  $q(\xi)$  with  $\xi \in \Theta$  as follows

$$(7.22) \quad q(\xi) = \text{const} \times \exp \left( \langle \log \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{w}, \mathbf{y}) \rangle_{q(\Theta_\xi)} \right),$$

where  $\Theta_\xi$  denotes the set  $\Theta$  with  $\xi$  removed and  $E_{\Theta_\xi}[\cdot] = \langle \cdot \rangle_{q(\Theta_\xi)}$  denotes expected value with respect to the distribution  $q(\Theta_\xi)$ . In the following, the subscript of the expected value will be removed when it is clear from the context.

Let us now proceed with deriving the explicit forms of the solutions for each unknown using (7.22).

#### 7.4.1. Estimation of HR Image Distribution

From (7.22), the distribution  $q(\mathbf{x})$  can be found as

$$(7.23) \quad q(\mathbf{x}) \propto \exp \left( -\frac{1}{2} \langle \alpha_{im} \rangle \text{TV}(\mathbf{x}) - \frac{1}{2} \sum_k \langle \beta_k \rangle E_{\mathbf{s}_k} [\| \mathbf{y}_k - \mathbf{A}\mathbf{H}_k \mathbf{C}(\mathbf{s}_k) \mathbf{x} \|^2] \right)$$

The explicit form of this distribution depends on the expectation  $E_{\mathbf{s}_k} [\| \mathbf{y}_k - \mathbf{A}\mathbf{H}_k \mathbf{C}(\mathbf{s}_k) \mathbf{x} \|^2]$ . This calculation is not easy since  $\mathbf{C}(\mathbf{s}_k)\mathbf{x}$  is nonlinear with respect to  $\mathbf{s}_k$ . Therefore, we expand  $\mathbf{C}(\mathbf{s}_k)\mathbf{x}$  using its first-order Taylor series around the mean value  $\langle \mathbf{s}_k \rangle = \bar{\mathbf{s}}_k = (\bar{\theta}_k, \bar{c}_k, \bar{d}_k)^T$  of the distribution  $q(\mathbf{s}_k)$  ( $\bar{\mathbf{s}}_k$  denotes the estimate obtained at the previous iteration). Proceeding in this fashion, we obtain the following approximation of  $\mathbf{C}(\mathbf{s}_k)\mathbf{x}$

$$(7.24) \quad \mathbf{C}(\mathbf{s}_k)\mathbf{x} \approx \mathbf{C}(\bar{\mathbf{s}}_k)\mathbf{x} + \mathbf{L}_{\mathbf{x}}(\bar{\mathbf{s}}_k)\mathbf{P}(\bar{\mathbf{s}}_k)(\mathbf{s}_k - \bar{\mathbf{s}}_k)$$

where

$$(7.25) \quad \begin{aligned} \mathbf{L}_{\mathbf{x}}(\bar{\mathbf{s}}_k) &= [\text{diag} \{ (\mathbf{I} - \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)})(\mathbf{L}_{\text{tr}(s_k)} - \mathbf{L}_{\text{tl}(s_k)})\mathbf{x} + \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)}(\mathbf{L}_{\text{br}(s_k)} - \mathbf{L}_{\text{bl}(s_k)})\mathbf{x} \}, \\ &\quad \text{diag} \{ (\mathbf{I} - \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)})(\mathbf{L}_{\text{bl}(s_k)} - \mathbf{L}_{\text{tl}(s_k)})\mathbf{x} + \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)}(\mathbf{L}_{\text{br}(s_k)} - \mathbf{L}_{\text{tr}(s_k)})\mathbf{x} \}] \end{aligned}$$

and

$$(7.26) \quad \mathbf{P}(\bar{\mathbf{s}}_k) = \begin{pmatrix} -\mathbf{u} \sin(\bar{\theta}_k) - \mathbf{v} \cos(\bar{\theta}_k) & \mathbf{1} & \mathbf{0} \\ \mathbf{u} \cos(\bar{\theta}_k) - \mathbf{v} \sin(\bar{\theta}_k) & \mathbf{0} & \mathbf{1} \end{pmatrix}$$

We now rewrite  $\mathbf{L}_{\mathbf{x}}(\bar{s}_k)\mathbf{P}(\bar{s}_k)$  in a more convenient form. We first define the matrices

$$(7.27) \quad \mathbf{M}_1(\bar{\mathbf{s}}_k) = (\mathbf{I} - \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)})(\mathbf{L}_{\mathbf{tr}(s_k)} - \mathbf{L}_{\mathbf{tl}(s_k)}) + \mathbf{D}_{\mathbf{b}_k(\mathbf{s}_k)}(\mathbf{L}_{\mathbf{br}(s_k)} - \mathbf{L}_{\mathbf{bl}(s_k)})$$

$$(7.28) \quad \mathbf{M}_2(\bar{\mathbf{s}}_k) = (\mathbf{I} - \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)})(\mathbf{L}_{\mathbf{bl}(s_k)} - \mathbf{L}_{\mathbf{tl}(s_k)}) + \mathbf{D}_{\mathbf{a}_k(\mathbf{s}_k)}(\mathbf{L}_{\mathbf{br}(s_k)} - \mathbf{L}_{\mathbf{tr}(s_k)}),$$

and

$$(7.29) \quad \mathbf{P}_1(\bar{\mathbf{s}}_k) = [\text{diag}(-\mathbf{u}\sin(\bar{\theta}_k) - \mathbf{v}\cos(\bar{\theta}_k))]$$

$$(7.30) \quad \mathbf{P}_2(\bar{\mathbf{s}}_k) = [\text{diag}(\mathbf{u}\cos(\bar{\theta}_k) - \mathbf{v}\sin(\bar{\theta}_k))]$$

Then, (7.25) can rewritten as

$$(7.31) \quad \mathbf{L}_{\mathbf{x}}(\bar{s}_k)\mathbf{P}(\bar{s}_k) = [(\mathbf{P}_1(\bar{\mathbf{s}}_k)\mathbf{M}_1(\bar{\mathbf{s}}_k) + \mathbf{P}_2(\bar{\mathbf{s}}_k)\mathbf{M}_2(\bar{\mathbf{s}}_k))\mathbf{x}, \mathbf{M}_1(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{M}_2(\bar{\mathbf{s}}_k)\mathbf{x}]$$

$$(7.32) \quad = [\mathbf{N}_1(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{N}_2(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{N}_3(\bar{\mathbf{s}}_k)\mathbf{x}],$$

such that using (7.24) we obtain

$$(7.33) \quad \begin{aligned} \mathbf{B}(\mathbf{s}_k) &= \mathbf{A}\mathbf{H}_k\mathbf{C}(\mathbf{s}_k) \approx \mathbf{A}\mathbf{H}_k\mathbf{C}(\bar{\mathbf{s}}_k)\mathbf{x} + \mathbf{A}\mathbf{H}_k[\mathbf{N}_1(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{N}_2(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{N}_3(\bar{\mathbf{s}}_k)\mathbf{x}](\mathbf{s}_k - \bar{\mathbf{s}}_k) \\ &= \mathbf{A}\mathbf{H}_k\mathbf{C}(\bar{\mathbf{s}}_k)\mathbf{x} + [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k), \mathbf{O}_{k2}(\bar{\mathbf{s}}_k), \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)](\mathbf{s}_k - \bar{\mathbf{s}}_k), \end{aligned}$$

with

$$(7.34) \quad \mathbf{O}_{kr}(\bar{\mathbf{s}}_k) = \mathbf{A}\mathbf{H}_k\mathbf{N}_r(\bar{\mathbf{s}}_k), \quad r = 1, 2, 3$$

The quantity  $\sum_k \langle \beta_k \rangle E_{\mathbf{s}_k} [\| \mathbf{y}_k - \mathbf{AD}_k \mathbf{C}(\mathbf{s}_k) \mathbf{x} \|^2]$  can then be calculated using (7.33) as

$$\begin{aligned}
 & \sum_k \langle \beta_k \rangle E_{\mathbf{s}_k} [\| \mathbf{y}_k - \mathbf{AH}_k \mathbf{C}(\mathbf{s}_k) \mathbf{x} \|^2] \\
 & \approx \sum_k \langle \beta_k \rangle \| \mathbf{y}_k - \mathbf{AH}_k \mathbf{C}(\bar{\mathbf{s}}_k) \mathbf{x} \|^2 \\
 & \quad + \sum_k \langle \beta_k \rangle \text{trace} \left( [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k), \mathbf{O}_{k2}(\bar{\mathbf{s}}_k), \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)]^T [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k), \mathbf{O}_{k2}(\bar{\mathbf{s}}_k), \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)] \boldsymbol{\Lambda}_k \right) \\
 (7.35) \quad & = \sum_k \langle \beta_k \rangle \| \mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k) \mathbf{x} \|^2 + \sum_k \sum_{i=1}^3 \sum_{j=1}^3 \langle \beta_k \rangle \lambda_{kij} \mathbf{x}^T \mathbf{O}_{ki}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{kj}(\bar{\mathbf{s}}_k) \mathbf{x}
 \end{aligned}$$

where  $\boldsymbol{\Lambda}_k$  is the covariance matrix of the posterior distribution  $q(\mathbf{s}_k)$  constructed by  $\lambda_{kij}$ ,  $i = 1, 2, 3$ ,  $j = 1, 2, 3$ , as follows

$$(7.36) \quad \boldsymbol{\Lambda}_k = \begin{pmatrix} \lambda_{k11} & \lambda_{k12} & \lambda_{k13} \\ \lambda_{k21} & \lambda_{k22} & \lambda_{k23} \\ \lambda_{k31} & \lambda_{k32} & \lambda_{k33} \end{pmatrix}$$

Finally, using (7.35) in (7.23), the posterior distribution  $q(\mathbf{x})$  of the HR image  $\mathbf{x}$  is found to be a multivariate Gaussian distribution given by

$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}).$$

with parameters

$$(7.37) \quad \mu_{\mathbf{x}} = \Sigma_{\mathbf{x}} \left[ \sum_k \langle \beta_k \rangle \mathbf{B}_k (\bar{\mathbf{s}}_k)^T \mathbf{y}_k \right]$$

$$\Sigma_{\mathbf{x}}^{-1} = \sum_k \langle \beta_k \rangle \mathbf{B}_k (\bar{\mathbf{s}}_k)^T \mathbf{B}_k (\bar{\mathbf{s}}_k) + \sum_k \sum_{i=1}^3 \sum_{j=1}^3 \beta_k \lambda_{kij} \mathbf{O}_{ki} (\bar{\mathbf{s}}_k)^T \mathbf{O}_{kj} (\bar{\mathbf{s}}_k)$$

$$(7.38) \quad + \langle \alpha_{im} \rangle (\Delta^h)^T \mathbf{W} \Delta^h + \langle \alpha_{im} \rangle (\Delta^v)^T \mathbf{W} \Delta^v$$

where

$$(7.39) \quad \mathbf{W} = \text{diag} \left( \frac{1}{\sqrt{w_i}} \right), \quad i = 1, \dots, PN.$$

The elements  $w_i$  of the auxiliary vector  $\mathbf{w} = (w_1, w_2, \dots, w_{PN})$  are calculated as

$$(7.40) \quad w_i = E_{\mathbf{x}}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2]$$

$$= (\Delta_i^h \mu_{\mathbf{x}})^2 + (\Delta_i^v \mu_{\mathbf{x}})^2 + \text{trace} [(\Delta_i^h)^T (\Delta_i^h) \Sigma_{\mathbf{x}}] + \text{trace} [(\Delta_i^v)^T (\Delta_i^v) \Sigma_{\mathbf{x}}], \quad i = 1, \dots, PN,$$

It is clear from (7.40) that the vector  $\mathbf{w}$  represents the local spatial activity in the HR image  $\mathbf{x}$ . Therefore, the matrix  $\mathbf{W}$  introduces spatial adaptivity into the estimation process of the HR image in (7.37)-(7.38) by controlling the smoothing applied at different locations. Moreover, the uncertainty of the image estimate is also taken into account by the last two terms in (7.40) when calculating the spatial adaptivity vector  $\mathbf{w}$  using the distribution  $q(\mathbf{x})$ .

#### 7.4.2. Estimation of the registration parameter distributions

The posterior distribution approximation  $q(\mathbf{s}_k)$  is found from (7.22) as

$$(7.41) \quad q(\mathbf{s}_k) \propto \exp \left( -\frac{1}{2} < \beta_k > E_{\mathbf{x}} [\| \mathbf{y}_k - \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} \|^2] - \frac{1}{2} (\mathbf{s}_k - \bar{\mathbf{s}}_k^p)^T (\Lambda_k^p)^{-1} (\mathbf{s}_k - \bar{\mathbf{s}}_k^p) \right)$$

To obtain the explicit form of this distribution, the expectation  $E_{\mathbf{x}} [\| \mathbf{y}_k - \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} \|^2]$  needs to be calculated. We proceed as in the previous section by using its Taylor series expansion around  $\bar{\mathbf{s}}_k$ , the estimate of the registration parameters obtained in the previous iteration. By utilizing the approximation (7.33) to obtain

$$\begin{aligned} E_{\mathbf{x}} [\| \mathbf{y}_k - \mathbf{B}_k(\mathbf{s}_k)\mathbf{x} \|^2] &= E_{\mathbf{x}} [\| \mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\mathbf{x} - [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\mathbf{x}, \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\mathbf{x}](\mathbf{s}_k - \bar{\mathbf{s}}_k) \|^2] \\ &= \| \mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\mu_{\mathbf{x}} - [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\mu_{\mathbf{x}}, \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\mu_{\mathbf{x}}, \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\mu_{\mathbf{x}}](\mathbf{s}_k - \bar{\mathbf{s}}_k) \|^2 + \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{B}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \\ &\quad + 2 [\text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}], \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}], \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}]] (\mathbf{s}_k - \bar{\mathbf{s}}_k) \\ &\quad + (\mathbf{s}_k - \bar{\mathbf{s}}_k)^T \Psi_k (\mathbf{s}_k - \bar{\mathbf{s}}_k) \end{aligned} \quad (7.42)$$

with

(7.43)

$$\Psi_k = \begin{pmatrix} \text{trace} [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \\ \text{trace} [\mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \\ \text{trace} [\mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] & \text{trace} [\mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \end{pmatrix}$$

The distribution  $q(\mathbf{s}_k)$  can then be explicitly expressed from (7.41) and (7.42) as a Gaussian distribution

$$(7.44) \quad q(\mathbf{s}_k) = \mathcal{N}(\mathbf{s}_k | \langle \mathbf{s}_k \rangle, \boldsymbol{\Lambda}_k)$$

with parameters

$$(7.45) \quad \begin{aligned} \langle \mathbf{s}_k \rangle &= \boldsymbol{\Lambda}_k \left[ (\boldsymbol{\Lambda}_k^p)^{-1} \bar{\mathbf{s}}_k^p + \langle \beta_k \rangle \Phi_k \bar{\mathbf{s}}_k + \langle \beta_k \rangle \Psi_k \bar{\mathbf{s}}_k \right. \\ &\quad \left. + \langle \beta_k \rangle \left[ (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}})^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}}, (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}})^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}}, (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}})^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} \right]^T \right. \\ &\quad \left. - \langle \beta_k \rangle \left[ \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}], \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}], \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \right]^T \right] \end{aligned}$$

and

$$(7.46) \quad \boldsymbol{\Lambda}_k^{-1} = (\boldsymbol{\Lambda}_k^p)^{-1} + \langle \beta_k \rangle \Psi_k + \langle \beta_k \rangle \Phi_k$$

with

$$(7.47) \quad \Phi_k = \begin{pmatrix} \mu_{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} & \mu_{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}}^T & \mu_{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} \\ \mu_{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} & \mu_{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} & \mu_{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} \\ \mu_{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} & \mu_{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} & \mu_{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} \end{pmatrix}$$

Note, however, due to the approximation (7.33), the distribution (7.44) should be estimated by iterating through (7.45) and (7.46) and recalculating the matrices  $\mathbf{O}_{k1}(\bar{\mathbf{s}}_k)$ ,  $\mathbf{O}_{k2}(\bar{\mathbf{s}}_k)$ ,  $\mathbf{O}_{k3}(\bar{\mathbf{s}}_k)$  and  $\mathbf{B}(\bar{\mathbf{s}}_k)$ , where at each iteration  $\bar{\mathbf{s}}_k$  is set equal to  $\langle \mathbf{s}_k \rangle$  and  $\mathbf{B}(\bar{\mathbf{s}}_k)$ . Prior knowledge about

the motion vectors are incorporated into the estimation procedure by their prior means  $\bar{\mathbf{s}}_k^P$  and covariance matrices  $\mathbf{\Lambda}_k^P$ .

Note that the proposed registration method in (7.44) with (7.45) and (7.46) provides an estimate of the distribution of the registration parameters, where the mean (7.45) is utilized as their point estimate. An interesting observation is that this registration method is a generalized stochastic version of the Lucas-Kanade registration algorithm [148] as applied to the super resolution problem. The classical Lucas-Kanade method can be obtained as a special case of (7.45) by setting the matrix  $\Psi_k$  equal to zero. This matrix incorporates the uncertainty of the image estimate  $\mathbf{x}$  into the motion estimation procedure. As will be demonstrated with experiments, this incorporation significantly helps in the motion estimation process and results in more accurate estimates, especially when the observation noise is high.

In this work, we considered a motion model including translation and rotation. However, the proposed framework can easily be extended to more complex parametric motion models, such as affine (with 6 parameters) or projective (with 8 parameters) motion models. In these cases, the definition of the matrices  $\mathbf{P}(\bar{\mathbf{s}}_k)$  in (7.26) is slightly different, resulting in a modification of the approximation in (7.33). With some algebra, it can be shown that results similar to (7.45) and (7.46) are obtained. Additionally, the proposed framework can be extended to include the estimation of the blur PSF. As with the registration parameters, the parameters of the blur PSF can be assumed to follow Gaussian distributions with *a priori* means computed using the LR images. Using a linearization procedure similar to (7.33), the distributions of the PSF parameters can be estimated in a separate step of the algorithm. The only requirement is that the derivatives of the blur PSF with respect to the PSF parameters must be available. Similar approaches have been used in existing methods (see, for instance, [228, 191]). This extension

to develop blind SR algorithms is left as future work. In conclusion, the proposed framework is very flexible to be extended to obtain SR methods for more complex imaging applications.

#### 7.4.3. Estimation of the hyperparameter distributions

In the last step of the algorithm, the distributions of the hyperparameters  $q(\alpha_{im})$  and  $q(\beta_k)$  are found from (7.22) as Gamma distributions, expressed as

$$(7.48) \quad q(\alpha_{im}) \propto \alpha_{im}^{PN/2-1+a_{\alpha_{im}}^o} \exp \left[ -\alpha_{im}(b_{\alpha_{im}}^o + \sum_i \sqrt{w_i}) \right]$$

and

$$(7.49) \quad q(\beta_k) \propto \beta_k^{N/2-1+a_{\beta}^o} \exp \left[ -\beta_k(b_{\beta}^o + \frac{E_{\mathbf{x}, \mathbf{s}_k} [\|\mathbf{y}_k - \mathbf{B}(\mathbf{s}_k)\mathbf{x}\|^2]}{2}) \right]$$

The quantity  $E_{\mathbf{x}, \mathbf{s}_k} [\|\mathbf{y}_k - \mathbf{B}(\mathbf{s}_k)\mathbf{x}\|^2]$  can be calculated using (7.35) as

$$(7.50) \quad \begin{aligned} E_{\mathbf{x}, \mathbf{s}_k} [\|\mathbf{y}_k - \mathbf{B}(\mathbf{s}_k)\mathbf{x}\|^2] &= E_{\mathbf{x}} [E_{\mathbf{s}_k} [\|\mathbf{y}_k - \mathbf{B}(\mathbf{s}_k)\mathbf{x}\|^2]] \\ &= E_{\mathbf{x}} \left[ \|\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\mathbf{x}\|^2 + \sum_{i=1}^3 \sum_{j=1}^3 \lambda_{kij} \mathbf{x}^T \mathbf{O}_{ki}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{kj}(\bar{\mathbf{s}}_k) \mathbf{x} \right] \\ &= \|\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\mu_{\mathbf{x}}\|^2 + \text{trace} [\mathbf{B}(\bar{\mathbf{s}}_k)^T \mathbf{B}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}}] \\ &\quad + \sum_{i=1}^3 \sum_{j=1}^3 \lambda_{kij} \mu_{\mathbf{x}}^T \mathbf{O}_{ki}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{kj}(\bar{\mathbf{s}}_k) \mu_{\mathbf{x}} + \text{trace} \left[ \sum_{i=1}^3 \sum_{j=1}^3 \lambda_{kij} \mathbf{O}_{ki}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{kj}(\bar{\mathbf{s}}_k) \Sigma_{\mathbf{x}} \right] \end{aligned}$$

Table 7.1. Proposed Algorithm I

**Algorithm 8.** *Variational Bayesian Super Resolution*

*Calculate initial estimates of the HR image, registration parameters, and hyperparameters*

*While convergence criterion is not met*

- (1) *Estimate the HR image distribution using (7.37) and (7.38).*
- (2) *Compute spatial adaptivity vector  $\mathbf{w}$  using (7.40).*
- (3) *Estimate the distribution of the registration parameters using (7.45) and (7.46).*
- (4) *Estimate the distributions of the hyperparameters  $\alpha_{\text{im}}$ ,  $\{\beta_k\}$  using (7.48) and (7.49).*

The means of the distributions in (7.48) and (7.49), which are used as hyperparameter estimates, are given by

$$(7.51) \quad \langle \alpha_{\text{im}} \rangle = \frac{PN/2 + a_{\alpha_{\text{im}}}^o}{\sum_i \sqrt{w_i} + b_{\alpha_{\text{im}}}^o},$$

$$(7.52) \quad \langle \beta_k \rangle = \frac{N + 2a_{\beta}^o}{E_{\mathbf{s}_k, \mathbf{x}} [\|\mathbf{y}_k - \mathbf{B}(\mathbf{s}_k)\mathbf{x}\|^2] + 2b_{\beta}^o},$$

Note that the shape and scale parameters  $a_{\alpha_{\text{im}}}^o$ ,  $a_{\beta}^o$ ,  $b_{\alpha_{\text{im}}}^o$ ,  $b_{\beta}^o$  can be used to incorporate prior knowledge about the variances of the HR image and observation noise, in case such a knowledge is available. If they are set equal to zero, which corresponds to utilizing flat hyperprior distributions for the hyperparameters, the observed LR images are made solely responsible for the whole estimation process.

In summary, the algorithm iterates between estimating the HR image using (7.37) and (7.38), the spatial adaptivity vector  $\mathbf{w}$  using (7.40), the registration parameters using (7.45) and (7.46), and finally the hyperparameters using (7.51) and (7.52). The algorithm is summarized in Algorithm 8. A major computational difficulty in Algorithm 8 is the explicit construction of the

matrix  $\Sigma_{\mathbf{x}}$  in (7.38), which requires the inversion of an  $PN \times PN$  matrix. To avoid this computation, we solve (7.37) efficiently using the conjugate gradient method, and in equations where the explicit form of  $\Sigma_{\mathbf{x}}$  is needed, i.e., in (7.40), (7.42), (7.43) and (7.50),  $\Sigma_{\mathbf{x}}$  is approximated by a diagonal matrix obtained by inverting the diagonal elements of (7.38). We have conducted extensive experiments with small images which permit the explicit inversion of (7.38) to verify the validity of this approximation, and we found out empirically that this approximation results in very close estimates and has a minor effect in the estimation process. Similar approximations have also been utilized in other Bayesian recovery methods [228, 10, 118].

It is worth emphasizing here that we did not assume *a priori* that  $q(\mathbf{x})$  and  $q(\mathbf{s}_k)$  are Gaussian distributions. This result is derived due to the minimization of the KL divergence with respect to all possible distributions according to the factorization  $q(\alpha_{im})q(\mathbf{x})\prod_{k=1}^L q(\mathbf{s}_k)\prod_{k=1}^L q(\beta_k)$  [31]. We can, however, make an assumption that these distributions are *degenerate*, i.e., they take one value with probability one and the rest of the values with probability zero. Using this assumption, we obtain another algorithm very similar to the one presented above, with the only exception that the uncertainty terms arising from the covariance matrices are removed. The derivation of this algorithm is very similar to the first one, and therefore we omit its details and provide the iterative procedure in Algorithm 9.

It is clear that using degenerate distributions for  $\mathbf{x}$  and  $\mathbf{s}_k$  in Algorithm 2 removes the uncertainty terms of the image and motion estimates. As demonstrated in the experimental results section, incorporation of this uncertainty through the covariances of  $\mathbf{x}$  and  $\mathbf{s}_k$  improve the restoration performance, especially in cases when the observation noise is high. This is mainly due to the fact that poor estimations of one variable (due to noise or outliers) can influence the estimation of other unknowns, and the overall performance can significantly be affected as

Table 7.2. Proposed Algorithm II

**Algorithm 9.** *Variational Bayesian Super Resolution with Degenerate Distributions Calculate initial estimates of the initial HR image, registration parameters, and hyperparameters While convergence criterion is not met*

(1) *Calculate the HR image estimate  $\hat{\mathbf{x}}$  using*

$$(7.53) \quad \hat{\mathbf{x}} = \left[ \sum_k \hat{\beta}_k \mathbf{B}_k(\bar{\mathbf{s}}_k)^T \mathbf{B}_k(\bar{\mathbf{s}}_k) + \hat{\alpha}_{\text{im}}(\Delta^h)^T \mathbf{W} \Delta^h + \hat{\alpha}_{\text{im}}(\Delta^v)^T \mathbf{W} \Delta^v \right]^{-1} \left[ \sum_k \hat{\beta}_k \mathbf{B}_k(\bar{\mathbf{s}}_k)^T \mathbf{y}_k \right]$$

(2) *Compute spatial adaptivity vector  $\mathbf{w}$  using*

$$(7.54) \quad w_i = (\Delta_i^h(\hat{\mathbf{x}}))^2 + (\Delta_i^v(\hat{\mathbf{x}}))^2$$

(3) *Estimate registration parameters using*

$$(7.55)$$

$$(\bar{\mathbf{s}}_k)^{\text{new}} = \left[ (\mathbf{\Lambda}_k^p)^{-1} + \hat{\beta}_k \Phi_k \right]^{-1} \left[ (\mathbf{\Lambda}_k^p)^{-1} \bar{\mathbf{s}}_k^p + \hat{\beta}_k \Phi_k \bar{\mathbf{s}}_k + \hat{\beta}_k \begin{pmatrix} (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}})^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \\ (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}})^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \\ (\mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}})^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \end{pmatrix} \right]$$

*with*

$$(7.56) \quad \Phi_k = \begin{pmatrix} \hat{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} & \hat{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}}^T & \hat{\mathbf{x}}^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \\ \hat{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} & \hat{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} & \hat{\mathbf{x}}^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \\ \hat{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k1}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} & \hat{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k2}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} & \hat{\mathbf{x}}^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)^T \mathbf{O}_{k3}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \end{pmatrix}$$

(4) *Compute hyperparameter estimates  $\hat{\alpha}_{\text{im}}, \hat{\beta}_k$  using*

$$(7.57) \quad \hat{\alpha}_{\text{im}} = \frac{PN/2 + a_{\alpha_{\text{im}}}^o}{\sum_i \sqrt{w_i} + b_{\alpha_{\text{im}}}^o},$$

$$(7.58) \quad \hat{\beta}_k = \frac{N + 2a_{\beta}^o}{\| \mathbf{y}_k - \mathbf{B}(\bar{\mathbf{s}}_k)\hat{\mathbf{x}} \|^2 + 2b_{\beta}^o},$$

a result. By estimating the full posterior distribution of the unknowns instead of point estimates corresponding to the maximum probability (such as MAP estimates), the uncertainty of the estimates are incorporated into the estimation procedure to ameliorate the propagation of estimation errors between unknowns.

We conclude this section by commenting on the computational complexity of the algorithms. Algorithms 8 and 9 have similar complexities, with Algorithm 8 requiring more computations per iteration due to the incorporation of the covariance matrices. The majority of computations in both algorithms is performed for estimating the HR image and the registration vectors. The HR image is calculated efficiently using the conjugate gradient method, and the registration parameters are calculated by inverting a  $3 \times 3$  matrix for each observed LR image. Therefore, the algorithms have computational demands very similar to most existing SR algorithms in the literature (for instance, the AM methods [105, 177, 228, 191]).

## 7.5. Experimental Results

In this section, we analyze the performance of the proposed algorithms on both synthetic and real images under various conditions. In synthetic experiments, the quality of the restored HR image is measured quantitatively by the peak signal-to-noise ratio (PSNR), which is defined as

$$(7.59) \quad \text{PSNR} = 10 \log_{10} \frac{NP}{\|\hat{\mathbf{x}} - \mathbf{x}\|^2}$$

where  $\hat{\mathbf{x}}$  and  $\mathbf{x}$  are the estimated and original HR images, respectively, and pixel values in both images are normalized to lie in the interval  $[0, 1]$ . We also provide examples of estimated HR images to assess the visual quality. To evaluate the estimated motion parameters we use the mean squared error (MSE), given by

$$(7.60) \quad \text{MSE} = \sum_{k=1}^L \|\hat{\mathbf{s}}_k - \mathbf{s}_k\|^2$$

In the following, Algorithm 8 will be abbreviated as *ALG1*, and Algorithm 9 as *ALG2*. In all experiments reported below, the initial values of the algorithms *ALG1* and *ALG2* are chosen as follows: The initial registration parameters are estimated using the standard Lucas-Kanade method [148], although other registration algorithms such as [235] can also be used. The HR image estimate is then initialized using the *average image* [191], which is an oversmooth estimate of the HR image obtained using the LR images as

$$(7.61) \quad \mathbf{x}_a = \mathbf{S}^{-1} \sum_{k=1}^L \mathbf{B}_k(\bar{\mathbf{s}}_k)^T \mathbf{y}_k,$$

where  $\mathbf{S}$  is a diagonal matrix with the column sums of  $\mathbf{B}_k(\bar{\mathbf{s}}_k)$  as its elements. Note that this initial estimate is calculated very efficiently, and it generally increases the robustness of the algorithm to the noise. On the other hand, other initializations (such as bicubic interpolation) generally resulted in similar restorations.

The inverse covariance matrices  $(\mathbf{\Lambda}_k^p)^{-1}$  are set equal to zero matrices, that is, no prior information is utilized about the uncertainty of motion vectors. The covariance matrices in *ALG1* are also initially set equal to zero. The rest of the algorithm parameters are automatically calculated from the HR image using the algorithm steps provided in Algorithm (8) and Algorithm (9). As the convergence criterion we used  $\|\mathbf{x}^n - \mathbf{x}^{n-1}\|^2 / \|\mathbf{x}^{n-1}\|^2 < 10^{-5}$ , where  $\mathbf{x}^n$  and  $\mathbf{x}^{n-1}$  are the image estimates at the  $n^{th}$  and  $(n-1)^{th}$  iterations, respectively.

In the following subsections, we present experimental results demonstrating a number of properties of the proposed methods in comparison with existing approaches.

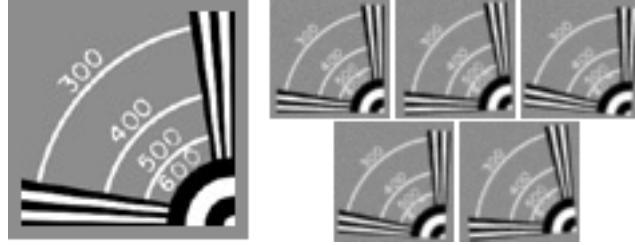


Figure 7.1. (*Left*) Original HR image, (*right*) Five synthetically generated LR images.

### 7.5.1. Synthetic Experiments with Exact Motion Information

In this section, we evaluate the performance of a number of SR methods in comparison with the proposed algorithms in cases where exact motion information is available. This study presents a comparison of the best possible performances achieved by the algorithms, and in addition it provides reference information to evaluate their behavior when the motion information is inaccurate, which will be studied in the next subsection.

We used the following methods for comparison: 1) Bicubic interpolation, 2) the robust SR method in [257] (denoted by *ZMT*), which is based on backprojection with median filtering, and 3) the robust SR method in [79] (denoted by *RSR*), which is based on bilateral TV priors. We also experimented with other SR methods contained in the EPFL SR software [236], but they provided inferior results compared to *ZMT* and *RSR*, and therefore they are not reported here.

We generated 5 synthetic LR images from the HR image shown on the left in Fig. (7.1) through warping, blurring and downsampling by a factor of 2. The warping consists of both translation and rotation, where the translations are chosen as

$$(7.62) \quad \begin{pmatrix} 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix}$$

pixels, and the rotation angles are chosen as  $(0^\circ, 3^\circ, -3^\circ, 5^\circ, -5^\circ)$ . As the blur we used a  $3 \times 3$  uniform PSF. The LR images obtained after the warping, blurring and downsampling operations are further degraded by additive white Gaussian noise at SNR levels of 5dB, 15dB, 25dB, 35dB and 45dB. Example LR images corresponding to the 25dB SNR case are shown in Fig. (7.1). Note that this resolution chart image is chosen for better illustration of the performance in resolution enhancement; similar results were obtained in experiments with other images.

We conducted simulations with 20 different noise realizations at each SNR level, and the average and variance of these experiments are reported. Since the algorithms *ZMT* and *RSR* contain algorithmic parameters, we exhaustively searched for the parameters resulting in the maximum PSNR value to report their best performance. Moreover, we reported the maximum PSNR result obtained during their iterations rather than the PSNR result at the convergence, and initialized the algorithms with both the bicubic interpolation result and the average image in (7.61), and chose the best resulting image among them. Note, however, that the parameters of the proposed methods are estimated automatically so there is no need for parameter tuning.

Mean PSNR values with the standard deviations provided by the algorithms are shown in Table 7.3, and the mean PSNR values are plotted in Fig. 7.2(a). As expected, all SR algorithms result in better reconstructions than bicubic interpolation. It is also clear that the proposed methods provide the best performance among all methods across all noise levels. It should be emphasized that the PSNR values of the methods *ZMT* and *RSR* are obtained by exhaustively tuning their parameters, which requires multiple runs, whereas the proposed methods provided their results in an fully-automated fashion in a single run. Therefore, even in the cases where the PSNR values are close, algorithms *ALG1* and *ALG2* should be preferred as the method of choice.

Table 7.3. Mean PSNRs with standard deviations in 20 experiments provided by the SR algorithms in different SNR levels for the case where motion information is exact.

SNR	5dB	15dB	25dB	35dB	45dB
Bicubic	$15.96 \pm 0.077$	$17.02 \pm 0.027$	$17.14 \pm 0.008$	$17.16 \pm 0.003$	$17.16 \pm 0.001$
ZMT	$18.71 \pm 0.082$	$20.48 \pm 0.242$	$20.69 \pm 0.378$	$20.55 \pm 0.275$	$20.53 \pm 0.002$
RSR	$19.07 \pm 0.075$	$22.25 \pm 0.049$	$26.4 \pm 0.080$	$31.22 \pm 0.059$	$33.67 \pm 0.070$
ALG1	$20.34 \pm 0.049$	$24.93 \pm 0.130$	$28.66 \pm 0.119$	$32.67 \pm 0.110$	$36.85 \pm 0.142$
ALG2	$17.48 \pm 0.049$	$24.93 \pm 0.123$	$28.48 \pm 0.112$	$32.15 \pm 0.092$	$36.05 \pm 0.155$

In general, *ALG1* provides restored HR images with slightly higher quality than *ALG2*. This is especially evident in high-noise cases (e.g., SNR = 5dB), where the incorporation of the uncertainty prevents the algorithm from overfitting due to high noise.

Example HR restorations are shown in Fig. (7.3) for the SNR = 25dB case, and in Fig. (7.4) for the SNR = 45dB case. It is clear that proposed methods provide the most visually enhanced restorations with significantly reduced ringing artifacts and much sharper edges compared to other methods. Restorations of *ALG1* and *ALG2* are very similar, with *ALG1* providing slightly sharper edges with less ringing artifacts.

### 7.5.2. Synthetic Experiments with Inaccurate Motion Information

In this section, we compare the performance of the SR methods when the motion parameters are inaccurate. We utilized the same setup as in the previous section, and used the same datasets to measure the decrease in performance due to the errors in registration parameters. In order to simulate the errors in motion estimation, we corrupted the original translation parameters by white Gaussian noise with standard deviation of 1, and the rotation parameters with noise uniformly distributed in  $[-2^\circ, 2^\circ]$ .

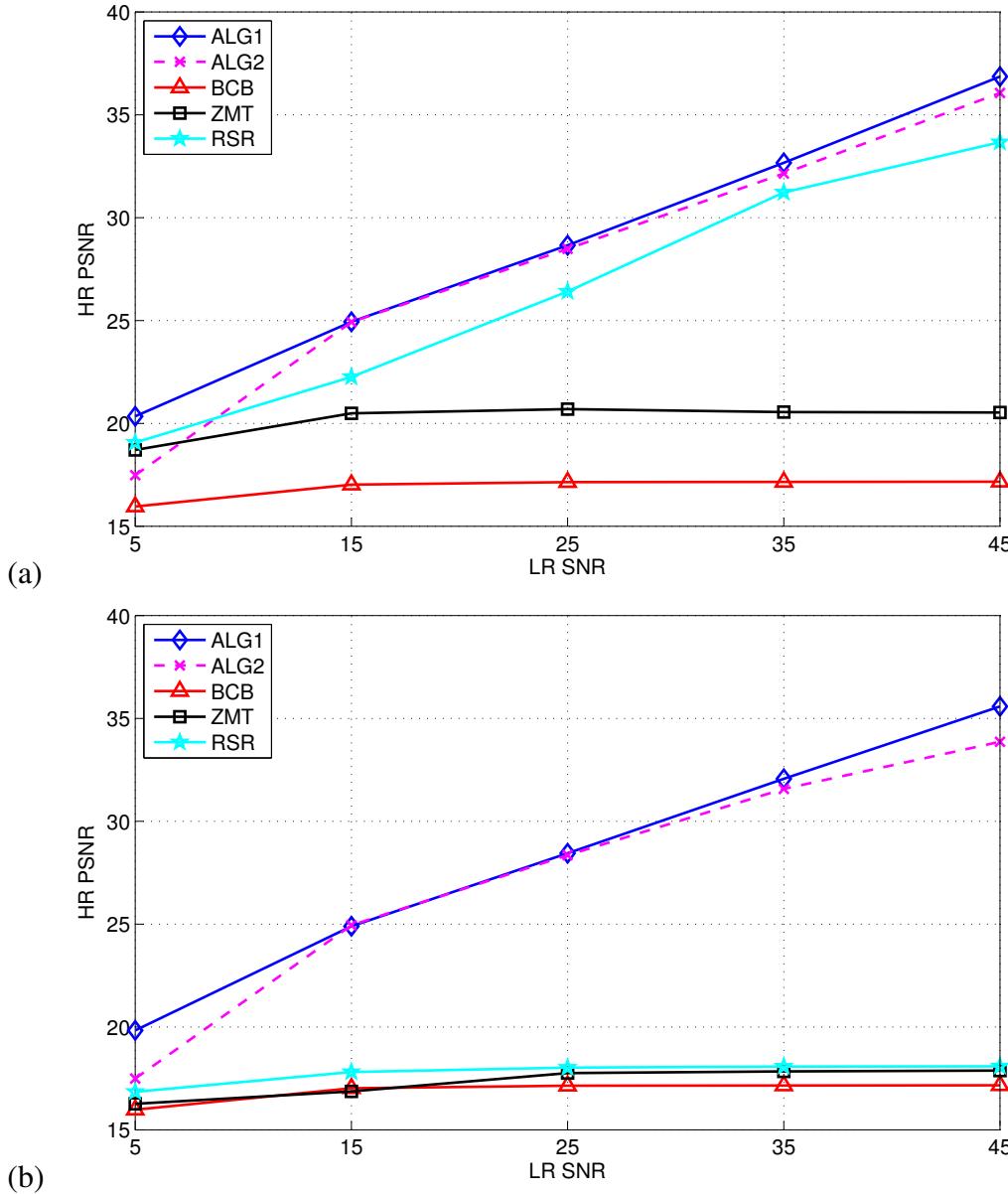


Figure 7.2. Mean PSNR values of SR algorithms for different input SNR levels when (a) exact motion information is available, and (b) motion information is inaccurate.

Mean PSNR values with standard deviations in 20 experiments are reported in Table 7.4, and the mean PSNR values are plotted in Fig. 7.2(b). Comparing Table 7.3 and Table 7.4, it

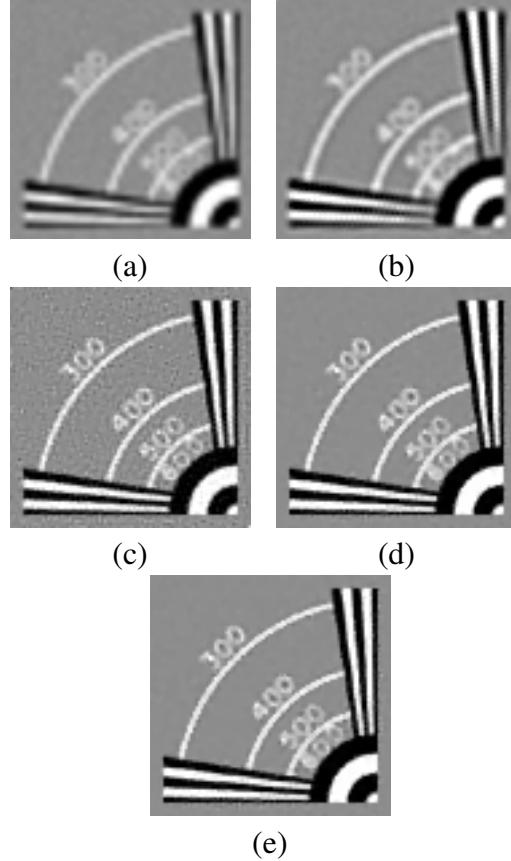


Figure 7.3. Example estimated HR images from different SR methods in the case when SNR=25dB and motion information is exact. Results of (a) Bicubic interpolation (PSNR = 17.14dB), (b) *ZMT* (PSNR = 20.55dB), (c) *RSR* (PSNR = 26.41dB), and the proposed methods (d) *ALG1* (PSNR = 28.75dB), and (e) *ALG2* (PSNR = 28.58dB).

can be seen that the performances of all algorithms decrease due to the motion errors, as expected. However, the performance degradation is severe with algorithms *ZMT* and *RSR*, mainly due to the fact that they do not incorporate motion estimation, but try to compensate for the motion errors using robust observation models. On the other hand, it is clear from Table 7.4 and Fig. 7.2(b) that the performance degradation with algorithms *ALG1* and *ALG2* is minor, and they resulted in almost the same PSNR values as in the case where motion information is exact.

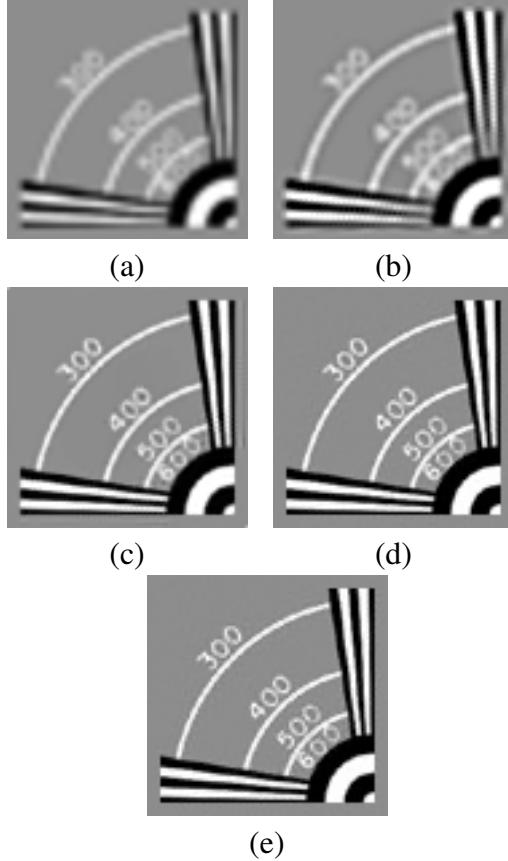


Figure 7.4. Example estimated HR images from different SR methods in the case when SNR=45dB and motion information is exact. Results of (a) Bicubic interpolation (PSNR = 17.16dB), (b) ZMT (PSNR = 20.53dB), (c) *RSR* (PSNR = 33.56dB), and the proposed methods (d) *ALG1* (PSNR = 36.81dB), and (e) *ALG2* (PSNR = 35.85dB).

This indicates that the restoration quality is significantly improved when the motion is accurately estimated. The corresponding MSE values of the motion parameters estimated by *ALG1* and *ALG2* are shown in Table 7.5. Note that *ALG1* and *ALG2* estimate the motion parameters very accurately in all noise levels. Another observation is that the variances among the resulting PSNR values obtained by *ALG1* and *ALG2* are much smaller than the PSNR variances obtained

Table 7.4. Mean PSNRs with standard deviations in 20 experiments provided by the SR algorithms in different SNR levels for the case motion information is inaccurate (see text).

SNR	5dB	15dB	25dB	35dB	45dB
Bicubic	$15.97 \pm 0.077$	$17.02 \pm 0.027$	$17.14 \pm 0.008$	$17.16 \pm 0.003$	$17.16 \pm 0.001$
ZMT	$16.26 \pm 0.978$	$16.86 \pm 1.035$	$17.76 \pm 0.503$	$17.84 \pm 0.552$	$17.88 \pm 0.600$
RSR	$16.84 \pm 0.412$	$17.81 \pm 0.567$	$18.02 \pm 0.655$	$18.07 \pm 0.688$	$18.08 \pm 0.692$
ALG1	$19.83 \pm 0.050$	$24.89 \pm 0.132$	$28.44 \pm 0.154$	$32.06 \pm 0.151$	$35.58 \pm 0.330$
ALG2	$17.49 \pm 0.050$	$24.94 \pm 0.127$	$28.34 \pm 0.115$	$31.57 \pm 0.107$	$33.86 \pm 0.416$

by *ZMT* and *RSR*, indicating the robustness of the proposed methods to inaccurate initialization of motion parameters.

Examples of estimated motion parameters are shown in Fig. (7.7) for the SNR = 25dB case, and in Fig. (7.8) for the SNR = 45dB case. Note that in both cases the algorithms provide very accurate estimates of the motion vectors, even though initial vectors contain high amounts of noise. For demonstration purposes, the contours of the covariance matrices at 10 standard deviations are plotted in Fig. 7.7(a), as estimated by *ALG1*. It is clear that the posterior distribution of the estimated vectors is very sharply peaked around the true vectors.

Examples of HR images estimated by the algorithms are shown in Fig. (7.5) for the SNR = 25dB case, and in Fig. (7.6) for the SNR = 45dB case. The degradation of visual quality in the methods *ZMT* and *RSR* is clear, especially comparing Figs. 7.3(b) and 7.3(c) to Figs. 7.5(b) and 7.5(c), respectively. However, *ALG1* and *ALG2* provided very high quality restorations, and there is almost no quality degradation when the initial motion parameters are inaccurate (compare Figs. 7.3(d) and 7.3(e) to Figs. 7.5(d) and 7.5(e), respectively, for *ALG1* and *ALG2*).

Table 7.5. Mean MSEs with standard deviations of the motion parameters in 20 experiments provided by the proposed methods in different SNR levels.

Methods	SNR				
	5dB	15dB	25dB	35dB	45dB
ALG1	$0.1436 \pm 0.0182$	$0.009526 \pm 0.0026$	$0.002675 \pm 0.0008$	$0.000341 \pm 0.0001$	$3.17 \times 10^{-5} \pm 10^{-6}$
ALG2	$0.1451 \pm 0.0315$	$0.001659 \pm 0.0006$	$0.000332 \pm 0.0001$	$0.001526 \pm 0.0005$	$0.002617 \pm 0.0011$

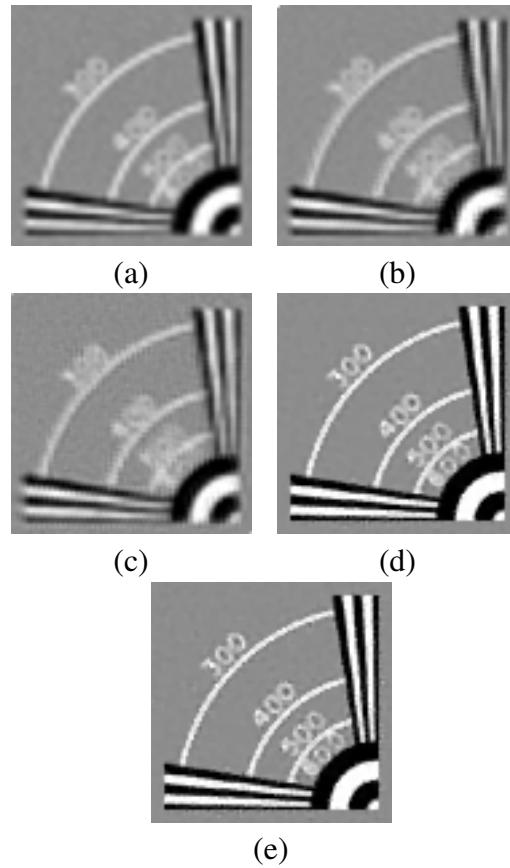


Figure 7.5. Example estimated HR images from different SR methods in the case when  $\text{SNR}=25\text{dB}$  and motion information is inaccurate (see text). Results of (a) Bicubic interpolation ( $\text{PSNR} = 17.14\text{dB}$ ), (b) *ZMT* ( $\text{PSNR} = 17.47\text{dB}$ ), (c) *RSR* ( $\text{PSNR} = 17.41\text{dB}$ ), and the proposed methods (d) *ALG1* ( $\text{PSNR} = 28.47\text{dB}$ ), and (e) *ALG2* ( $\text{PSNR} = 28.34\text{dB}$ ).

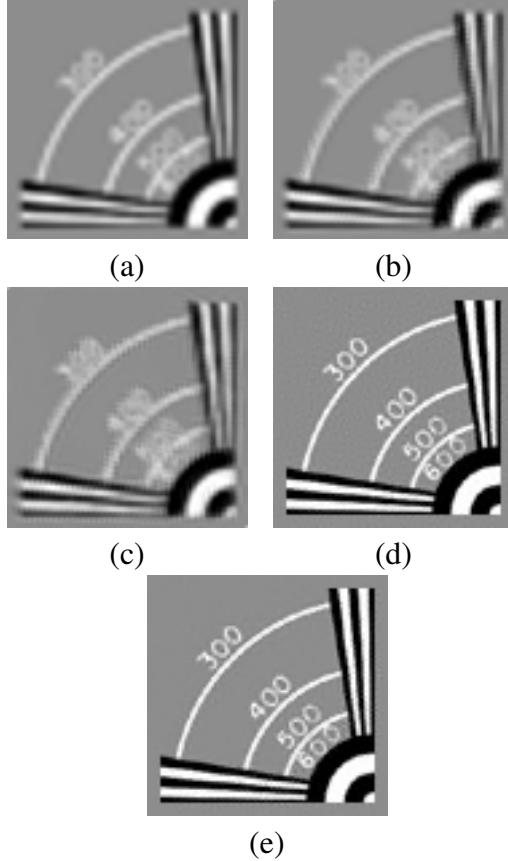


Figure 7.6. Example estimated HR images from different SR methods in the case when SNR=45 dB and motion information is inaccurate (see text). Results of (a) Bicubic interpolation (PSNR = 17.16dB), (b) *ZMT* (PSNR = 17.49dB), (c) *RSR* (PSNR = 17.44dB), and the proposed methods (d) *ALG1* (PSNR = 35.11dB), and (e) *ALG2* (PSNR = 33.40dB).

### 7.5.3. Experiments with Real Images

We conducted a number of experiments with the proposed algorithms on real SR applications, some of which are presented in this section. We report real image experiments performed on the datasets provided by the UCSC [3]. The algorithms *ZMT* and *RSR* are used again for comparing the performance of the algorithms, and we used the MDSP software [78] to obtain their results. We also provide results from the algorithm in [76], denoted by *EF*. The motion

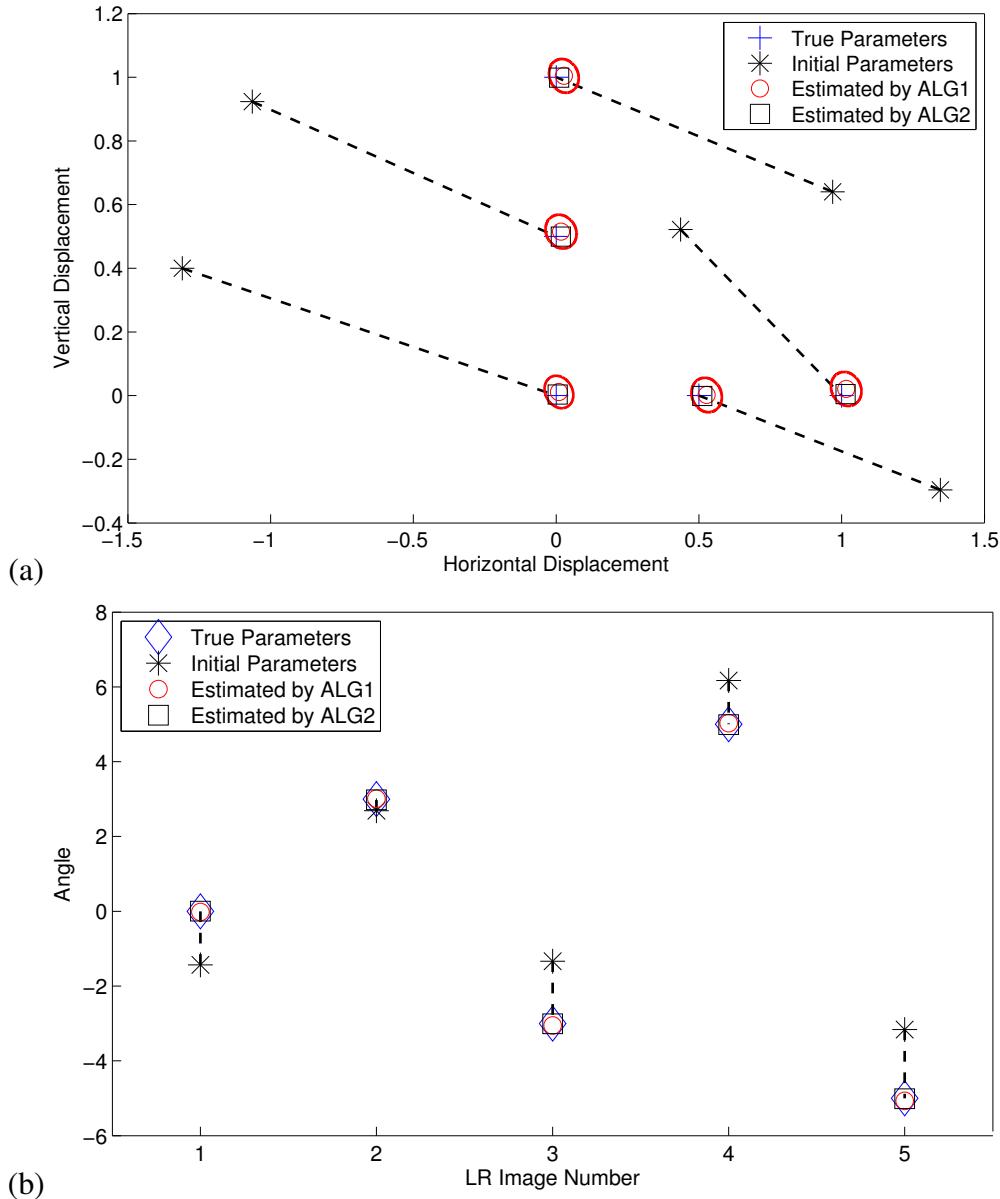


Figure 7.7. Estimated motion parameters by the algorithms *ALG1* and *ALG2* when SNR = 25dB. (a) Estimated translation parameters, (b) Estimated rotation angles. The resulting MSEs of the estimated parameters are 0.0029 for *ALG1* and  $7.91 \times 10^{-4}$  for *ALG2*.

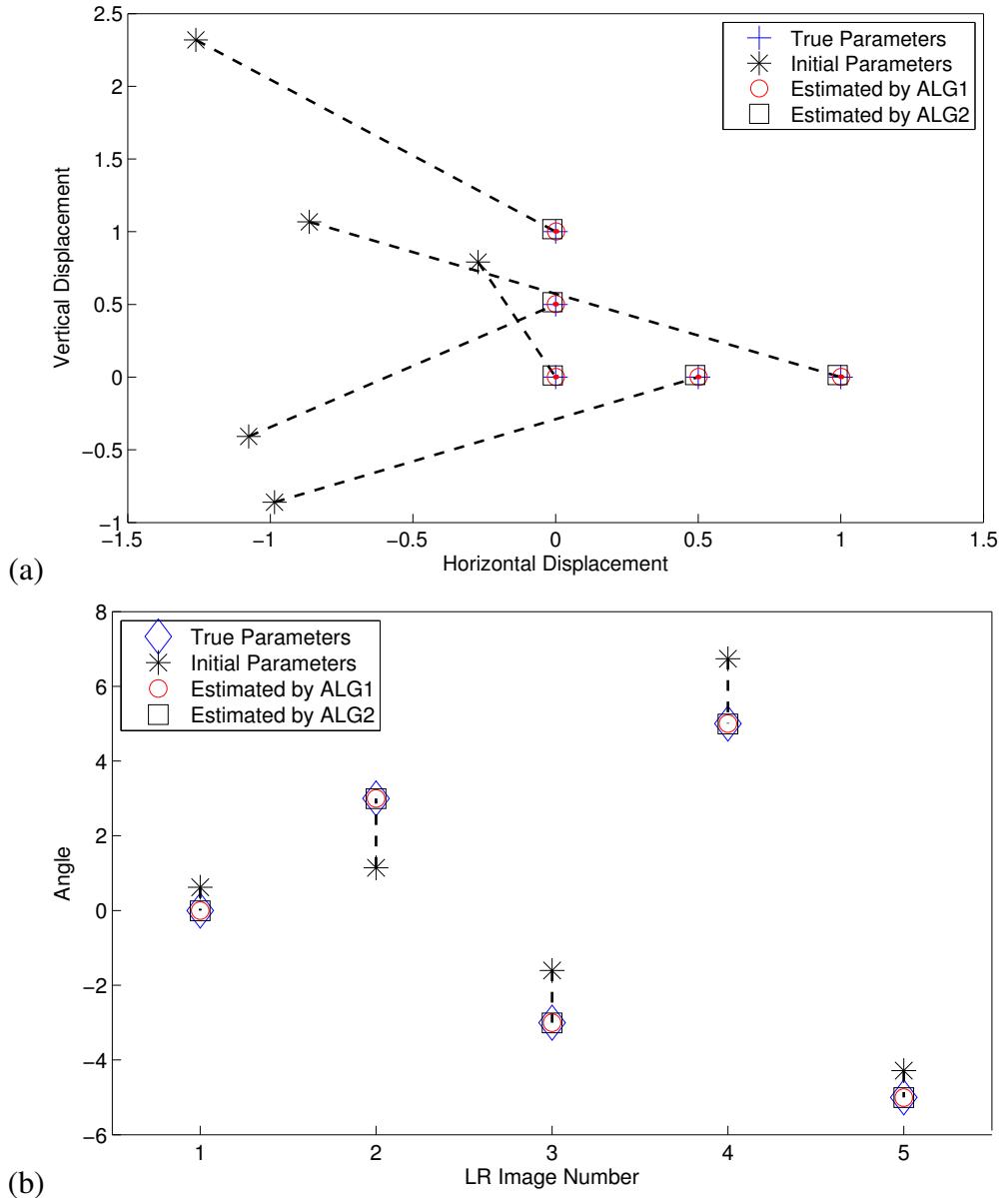


Figure 7.8. Estimated motion parameters by the algorithms *ALG1* and *ALG2* when SNR = 45dB. (a) Estimated translation parameters, (b) Estimated rotation angles. The resulting MSEs of the estimated parameters are  $2.48 \times 10^{-5}$  for *ALG1* and 0.0017 for *ALG2*.

parameters are estimated again using the MDSP software, and provided to *ALG1* and *ALG2* as initial parameters.

As with the synthetic experiments, we manually tuned all required parameters of the algorithms *ZMT*, *RSR* and *EF* to obtain the most visually appealing results. On the other hand, no prior knowledge is assumed in the proposed methods except for the initial motion parameters. The inverse covariance matrices  $(\Lambda_k^p)^{-1}$  are set equal to zero matrices so that the estimation process only depends on the LR images.

In the first experiment, 20 LR images taken from the disk dataset from [79]. The blur PSF is assumed to be a 5x5 Gaussian with variance 1. The reconstructed HR images by a factor of four resolution enhancement obtained by bicubic interpolation and SR algorithms are shown in Fig. (7.9).

Second dataset consists of 15 LR images taken from the adyoron dataset from [3]. The blur PSF is assumed to be a 5x5 Gaussian with variance 1. The reconstructed HR images by a factor of three resolution enhancement obtained by bicubic interpolation and SR algorithms are shown in Fig. (7.10).

The final dataset consists of 53 LR images taken from the *Emily* dataset from [3]. The blur PSF is assumed to be a 5x5 Gaussian with variance 1. The reconstructed HR images by a factor of five resolution enhancement obtained by bicubic interpolation and SR algorithms are shown in Fig. (7.11).

It is clear from Figs. (7.9), (7.10) and (7.11) that the proposed methods provide HR image estimates with sharper edges and less ringing artifacts than other methods. Note that estimation errors were present in the motion parameters estimated from the LR images, and the proposed methods provided high restoration quality by incorporating these estimation errors and improving the motion errors simultaneously with the HR image estimates. The algorithms *ZMT*, *RSR* and *EF* do not provide images as sharp as *ALG1* and *ALG2*, since robustness to estimation errors

is the sole purpose in these methods. On the other hand, it is clear that HR image estimates with higher quality can be obtained by the proposed methods with improved estimation of the motion. *ALG1* and *ALG2* provide very similar results, but *ALG1* results in slightly sharper images where the edges are slightly smoother than the results of *ALG2*.

In summary, experimental results with both synthetic and real image sets demonstrate that the proposed algorithms are very effective in providing high quality super resolution results, and they compare favorably to some of the state-of-the-art super resolution methods.

## 7.6. Conclusions

In this chapter, we presented a novel Bayesian formulation for joint image registration and super resolution. The unknown high resolution image, motion parameters and algorithm parameters, including the noise variances, are modeled within a hierarchical Bayesian framework. Using this model, we develop two algorithms with variational Bayesian analysis, both of which estimate all unknowns and algorithmic parameters solely from the observed low resolution images without prior knowledge or user intervention. We have shown that the proposed motion estimation method generalizes the classical Lucas-Kanade registration method in a stochastic sense. The proposed methods have the following advantages: First, the proposed framework allows for estimation of distributions of unknowns, which prevent the propagation of estimation errors within the estimation procedure. This is especially useful when the acquisition noise is heavy. Second, through the incorporation of motion estimation and adaptive estimation of the algorithm parameters, the algorithms are very robust to errors in motion estimates. We have demonstrated that the algorithms are virtually unaffected by inaccuracies in motion estimates.

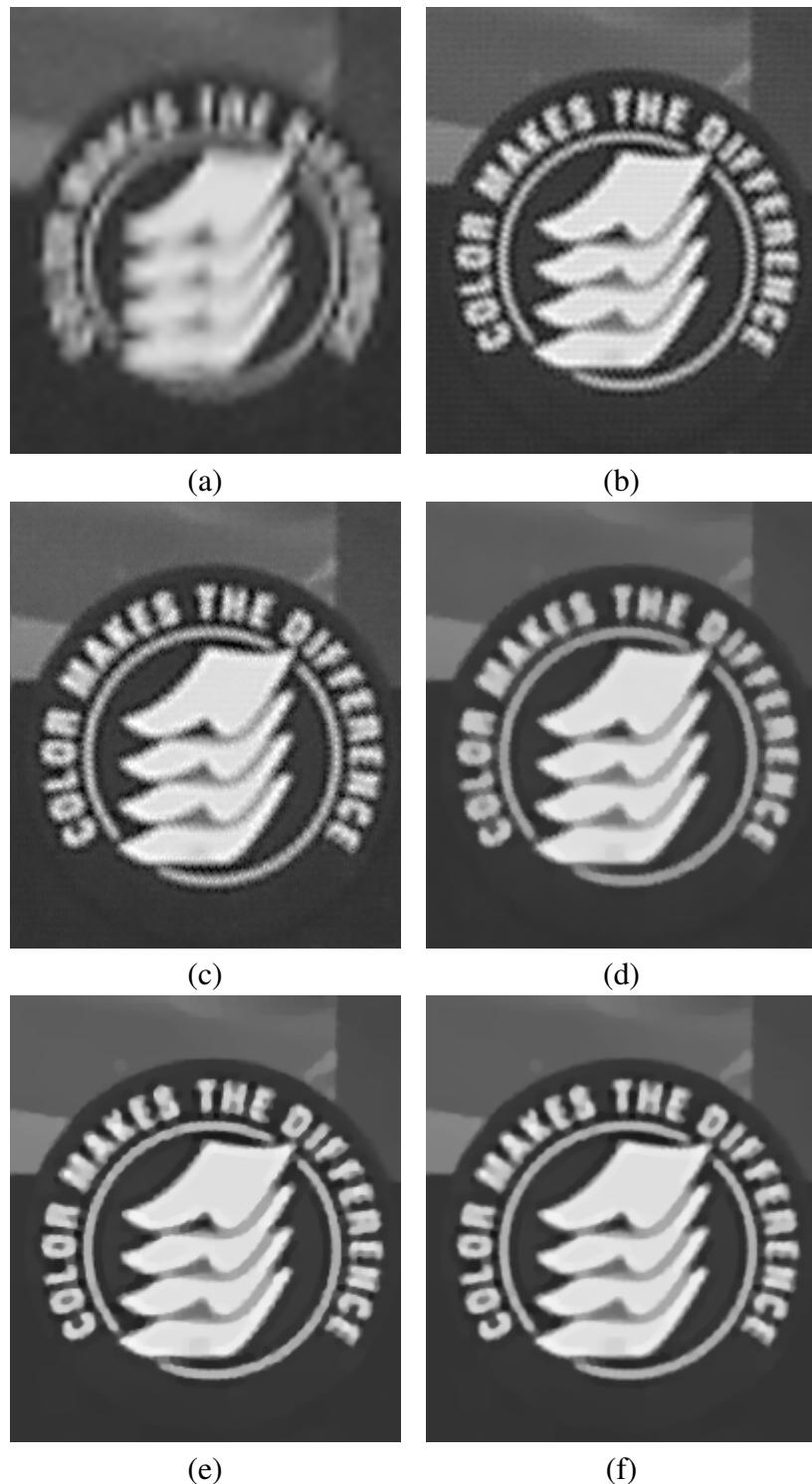


Figure 7.9. Super resolution results (4x resolution increase) by (a) bicubic interpolation, (b) *EF*, (c) *ZMT*, (d) *RSR*, (e) *ALG1* and (f) *ALG2*.

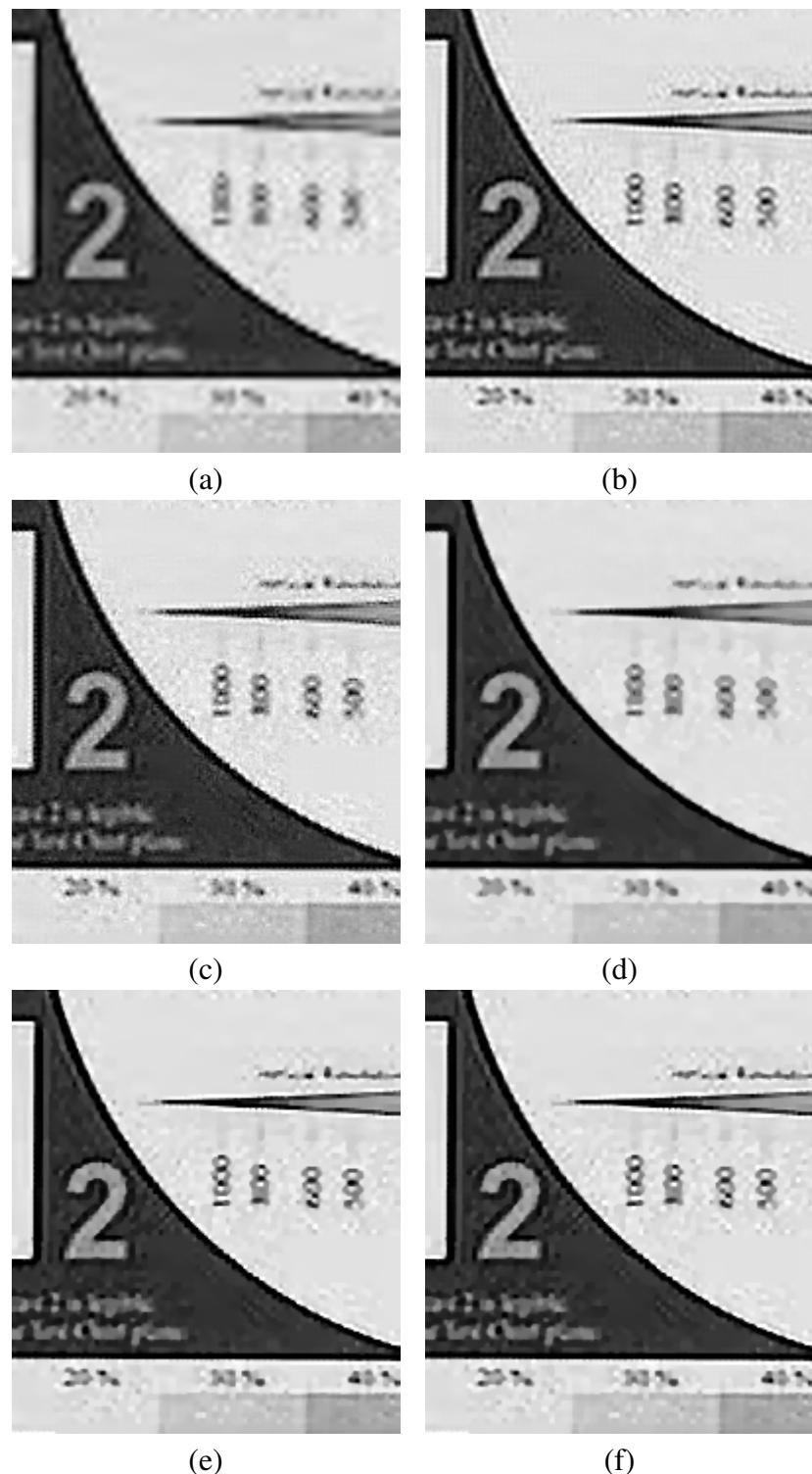


Figure 7.10. Super resolution results (3x resolution increase) by (a) bicubic interpolation, (b) EF, (c) ZMT, (d) RSR, (e) ALG1 and (f) ALG2.



Figure 7.11. Super resolution results (5x resolution increase) by (a) bicubic interpolation, (b) *EF*, (c) *ZMT*, (d) *RSR*, (e) *ALG1* and (f) *ALG2*.

Third, all required parameters of the algorithms are calculated automatically so they do not require user supervision unlike most existing super resolution methods. Experimental results with both synthetic and real images demonstrate that despite the lack of manual parameter tuning,

the proposed methods provide super resolution results superior to existing algorithms. Finally, we have shown that the proposed framework can be extended to handle more complex super resolution applications, such as blur estimation and more complex parametric motion models.

## CHAPTER 8

### **Bayesian Compressive Sensing using Laplace Priors**

#### **8.1. Introduction**

Compressive sensing (or sampling) (CS) has become a very active research area in recent years due to its interesting theoretical nature and its practical utility in a wide range of applications. Let  $\mathbf{f}$  represent the  $N \times 1$  unknown signal, which is compressible in a linear basis  $\Psi$  (such as a wavelet basis). In other words,  $\mathbf{f} = \Psi\mathbf{w}$ , where  $\mathbf{w}$  is an  $N \times 1$  sparse signal, i.e., most of its coefficients are zero. Consider the following acquisition system

$$(8.1) \quad \mathbf{y} = \Phi'\mathbf{f} + \mathbf{n},$$

where  $M \times 1$  linear measurements  $\mathbf{y}$  of the original unknown signal  $\mathbf{f}$  are taken with an  $M \times N$  measurement matrix  $\Phi' = [\phi'_1, \phi'_2, \dots, \phi'_N]$  and  $\mathbf{n}$  represents the acquisition noise. We can also write (8.1) in terms of the sparse signal coefficients as

$$(8.2) \quad \mathbf{y} = \Phi\mathbf{w} + \mathbf{n},$$

where  $\Phi = \Phi'\Psi$ , which is the commonly used notation in the CS literature and will be adapted in the rest of this chapter.

---

<sup>0</sup>This work has appeared in [9, 14].

According to the theory of compressive sensing when the number of measurements is small compared to the number of signal coefficients ( $M \ll N$ ), under certain conditions the original signal  $\mathbf{f}$  can be reconstructed very accurately by utilizing appropriate reconstruction algorithms [41, 72]. Compressive sensing can be seen as the combination of the conventional acquisition and compression processes: Traditionally, the signal  $\mathbf{f}$  is acquired in a lossless manner followed by compression where only the most important features are kept, such as the largest wavelet coefficients. In [41, 72] it has been shown that since the signal is compressible, it is possible to merge the acquisition and compression processes by performing a reduced number of measurements and recovering the most important features by utilizing an incoherent sampling mechanism, i.e., the sensing basis  $\Phi'$  and the representation basis  $\Psi$  have low coherence. Recent theoretical results show that random sampling matrices exhibit such low-coherence with the representation bases. Deterministic designs have also been proposed with slightly reduced performance [71, 103].

There are many applications of compressive sensing, including medical imaging [150] where reducing the number of measurements results in reduced image acquisition time, imaging processes where the cost of taking measurements is high, and sensor networks, where the number of sensors may be limited [107].

Since the number of measurements  $M$  is much smaller than the number of unknown coefficients  $\mathbf{w}$ , the original signal cannot be obtained directly from the measurements. The inversion of (8.1) or (8.2) is required, which is an ill-posed problem. Therefore, compressive sensing incorporates a reconstruction mechanism to obtain the original signal. By exploiting the sparsity of  $\mathbf{w}$ , the inverse problem is regularized constraining the  $l_0$  norm of  $\mathbf{w}$ ,  $\|\mathbf{w}\|_0$ , which is equal to the number of nonzero terms in  $\mathbf{w}$ . An approximation to the original signal is then obtained by

solving the following optimization problem

$$(8.3) \quad \hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \left\{ \| \mathbf{y} - \Phi \mathbf{w} \|_2^2 + \tau \| \mathbf{w} \|_0 \right\}.$$

This optimization problem is NP-hard, therefore some simplifications are used. The most common one is to use the  $l_1$ -norm instead of the  $l_0$ -norm, so that the optimization problem becomes

$$(8.4) \quad \hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \left\{ \| \mathbf{y} - \Phi \mathbf{w} \|_2^2 + \tau \| \mathbf{w} \|_1 \right\},$$

where  $\| \cdot \|_1$  denotes the  $l_1$ -norm.

A number of methods have been proposed to solve the CS reconstruction problems defined in (8.3) and (8.4) or their extensions (for example, formulations utilizing  $l_p$  norms for  $\mathbf{w}$  with  $0 < p \leq 1$ ). Most of the proposed methods are examples of energy minimization methods, including linear programming algorithms [64, 83] and constructive (greedy) algorithms [232, 73, 34]. Additionally, sparse signal representation is a very close topic to CS, and many algorithms proposed there can also be applied to the CS reconstruction problem (see [247] and references therein).

The compressive sensing formulation in (8.3) and (8.4) can be considered as the application of a deterministic regularization approach to signal reconstruction. However, the problem can also be formulated in a Bayesian framework, which provides certain distinct advantages over other formulations. These include providing probabilistic predictions, automatic incorporation and estimation of model parameters, and estimation of the uncertainty of reconstruction. The latter advantage also facilitates the estimation of the quality of the measurements which can be used to design adaptive measurements [116, 209]. The Bayesian framework was utilized for the

compressive sensing problem in [116, 209]. In [116], the *relevance vector machine* (RVM) proposed in [225] is adapted to the CS problem. Independent Laplace priors are utilized for each coefficient in an expectation-propagation framework in [209], and both signal reconstruction and measurement design problems are considered. However, the resulting algorithm is complicated to implement, and all required parameters are not estimated, but rather left as parameters to be tuned.

In this work, we also formulate the CS reconstruction problem from a Bayesian perspective. We utilize a Bayesian model for the CS problem and propose the use of Laplace priors on the basis coefficients in a hierarchical manner. As will be shown, our formulation includes the RVM formulation [225] as a special case, but results in smaller reconstruction errors while imposing sparsity to a higher extent. Moreover, we provide an alternative Bayesian inference procedure which results in an efficient greedy constructive algorithm. Our formulation naturally incorporates the advantages of the Bayesian framework, such as providing posterior distributions rather than point estimates, and therefore, providing an estimate of the uncertainty in the reconstructions, which, for instance, can be used as a feedback mechanism for adapting the data acquisition process. Furthermore, the resulting algorithms are fully automated since all required model parameters are estimated along with the unknown signal coefficients  $\mathbf{w}$ . This is in contrast to most of the proposed methods in the literature which include a number of parameters to be tuned specifically to the data, which is a cumbersome process. We will demonstrate with experimental results that despite being fully automated, the proposed algorithm provides competitive and even higher reconstruction performance than state-of-the-art methods.

The rest of this chapter is organized as follows: In Section 8.2, we present the hierarchical Bayesian modeling of the CS problem, the observation model and the prior model on the signal

coefficients. In this section, we review existing prior models for sparse learning and show that some of them are special cases of the proposed model. In Section 8.3 we apply the evidence procedure to the CS problem and propose two reconstruction algorithms. We present experimental results in Section 8.4 and conclusions are drawn in Section 8.5.

## 8.2. Bayesian Modeling

In Bayesian modeling, all unknowns are treated as stochastic quantities with assigned probability distributions. The unknown signal  $\mathbf{w}$  is assigned a *prior* distribution  $p(\mathbf{w}|\boldsymbol{\gamma})$ , which models our knowledge on the nature of  $\mathbf{w}$ . The observation  $\mathbf{y}$  is also a random process with *conditional* distribution  $p(\mathbf{y}|\mathbf{w}, \beta)$ , where  $\beta = 1/\sigma^2$  is the inverse noise variance. These distributions depend on the model parameters  $\boldsymbol{\gamma}$  and  $\beta$ , which are called *hyperparameters*, and additional prior distributions, called *hyperpriors*, are assigned to them.

The Bayesian modeling of the CS reconstruction problem requires the definition of a joint distribution  $p(\mathbf{w}, \boldsymbol{\gamma}, \beta, \mathbf{y})$  of all unknown and observed quantities. In this work we use the following factorization

$$(8.5) \quad p(\mathbf{w}, \boldsymbol{\gamma}, \beta, \mathbf{y}) = p(\mathbf{y}|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) p(\beta).$$

### 8.2.1. Observation (Noise) Model

The observation noise is independent and Gaussian with zero mean and variance equal to  $\beta^{-1}$ , that is, with (8.2),

$$(8.6) \quad p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \beta^{-1}),$$

with a Gamma prior placed on  $\beta$  as follows

$$(8.7) \quad p(\beta | a^\beta, b^\beta) = \Gamma(\beta | a^\beta, b^\beta).$$

The Gamma distribution is defined as

$$(8.8) \quad \Gamma(\xi | a^\xi, b^\xi) = \frac{(b^\xi)^{a^\xi}}{\Gamma(a^\xi)} \xi^{a^\xi - 1} \exp[-b^\xi \xi],$$

where  $\xi > 0$  denotes a hyperparameter,  $b^\xi > 0$  is the scale parameter, and  $a^\xi > 0$  is the shape parameter. The mean and variance of  $\xi$  are given respectively by

$$(8.9) \quad \text{Mean}[\xi] = \langle \xi \rangle = \frac{a^\xi}{b^\xi}, \quad \text{Var}[\xi] = \frac{a^\xi}{(b^\xi)^2}.$$

The Gamma distribution is generally chosen as the prior for the inverse variance (precision) of a Gaussian distribution because it is its conjugate prior which greatly simplifies the analysis and also includes the uniform distribution as a limiting case.

### 8.2.2. Signal Model

The  $l_1$  regularization formulation in (8.4) is equivalent to using a Laplace prior on the coefficients  $\mathbf{w}$ , that is,

$$(8.10) \quad p(\mathbf{w} | \lambda) = \frac{\lambda}{2} \exp(-\frac{\lambda}{2} |\mathbf{w}|)$$

and using a *maximum a posteriori* (MAP) formulation with (8.6) and (8.10) for  $\tau = \lambda / \beta$ . However, this formulation of the Laplace prior does not allow for a tractable Bayesian analysis, since it is not conjugate to the conditional distribution in (8.6). To alleviate this, hierarchical priors

are employed. In the following, we review the models so far utilized in the literature to model  $\mathbf{w}$  and introduce the prior structure utilized in this work.

In [84], as the first stage of a hierarchical model, the following prior is employed on  $\mathbf{w}$

$$(8.11) \quad p(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{i=1}^N \mathcal{N}(w_i|0, \gamma_i),$$

where  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_N)$ . In the second stage of the hierarchy, a Jeffrey's hyperprior is utilized independently on each  $\gamma_i$ , that is,

$$(8.12) \quad p(\gamma_i) \propto \frac{1}{\gamma_i}.$$

Observe that since

$$(8.13) \quad p(\gamma_i) = \lim_{\zeta \rightarrow 0} \Gamma(\gamma_i|\zeta, 0),$$

we can obtain a sample from the prior distribution of each  $w_i$  independently by first obtaining a sample  $\gamma_i$  from a  $\Gamma(\zeta, 0)$  distribution when  $\zeta \rightarrow 0$  and then sampling a  $\mathcal{N}(0, \gamma_i)$ .

Alternatively, in [225], the prior model on  $\mathbf{w}$  is formulated conditioned on the precision variables  $\alpha_i = \gamma_i^{-1}$ . A Gamma hyperprior is utilized on the precision variables, that is,

$$(8.14) \quad p(\alpha_i|a_i^\alpha, b_i^\alpha) = \Gamma(\alpha_i|a_i^\alpha, b_i^\alpha).$$

This formulation with the hierarchical prior in (8.11) and (8.14) is commonly referred to as the *relevance vector machine* (RVM), or *sparse Bayesian learning* (SBL) [225, 247]. Note, however, that both in the original work [225] and its adaptation to the compressive sensing

problem [116], the shape and scale parameters are set respectively equal to  $a_i^\alpha = 1, b_i^\alpha = 0$ , thus obtaining *uniform* or *noninformative* distributions for these parameters.

When using non-informative priors on  $\alpha_i$ ,  $p(\alpha_i|a_i^\alpha, b_i^\alpha)$  becomes

$$(8.15) \quad p(\alpha_i|1,0) = \lim_{\zeta \rightarrow 1} \Gamma(\alpha_i|\zeta,0).$$

It is important to mention that when changing variables from  $\gamma_i$  to  $\alpha_i$  their corresponding maximum a posteriori estimations are not related by  $(\alpha_i)_{MAP} = 1/(\gamma_i)_{MAP}$ .

Other values of  $a_i^\alpha$  and  $b_i^\alpha$  than  $a_i^\alpha = 1, b_i^\alpha = 0$  will result in Student's t distributions for the marginal distribution  $p(\mathbf{w})$ . It is argued in [245] that Student's t priors will lead to less sparse solutions than RVM.

As explained in [209], compared to the separate Gaussian priors employed on the entries of  $\mathbf{w}$  in the RVM framework, Laplace priors enforce the sparsity constraint more heavily by distributing the posterior mass more on the axes so that signal coefficients close to zero are preferred. Furthermore, the Laplace prior is also the prior that promotes sparsity to the largest extent while being log-concave [209]. The log-concavity provides the very useful advantage of eliminating local-minima since it leads to unimodal posterior distributions [209, 186, 246].

Based on the above, in this work, we propose to use Laplace priors on the signal coefficients  $\mathbf{w}$ . In order to overcome the fact that the Laplace distribution is not conjugate to the observation model in (8.6), we model it in a hierarchical way by using the following hyperpriors on  $\gamma_i$  [84],

$$(8.16) \quad p(\gamma_i|\lambda) = \Gamma(\gamma_i|1,\lambda/2) = \frac{\lambda}{2} \exp\left(-\frac{\lambda \gamma_i}{2}\right), \quad \gamma_i \geq 0, \lambda \geq 0$$

and then using the Gaussian model in (8.11) to model  $p(\mathbf{w}|\boldsymbol{\gamma})$ . In other words we have

$$(8.17) \quad p(\mathbf{w}|\lambda) = \int p(\mathbf{w}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\lambda)d\boldsymbol{\gamma} = \prod_i \int p(w_i|\gamma_i)p(\gamma_i|\lambda)d\gamma_i = \frac{\lambda^{N/2}}{2^N} \exp\left(-\sqrt{\lambda} \sum_i |w_i|\right).$$

Finally, we model  $\lambda$  as the realization of the following Gamma hyperprior

$$(8.18) \quad p(\lambda|v) = \Gamma(\lambda|v/2, v/2).$$

The proposed modeling constitutes a three-stage hierarchical form. The first two stages of this hierarchical prior (8.11) and (8.16) result in a Laplace distribution  $p(\mathbf{w}|\lambda)$  [84], and the last stage (8.18) is embedded to calculate  $\lambda$ . This formulation can be shown to be very closely related to the convex variational formulation in [97], and the total-variation priors used in image restoration [13].

The prior distribution on  $\lambda$  is flexible enough so as to provide a range of restrictions on  $\lambda$ ; from very vague information on  $\lambda$

$$(8.19) \quad p(\lambda) \propto \frac{1}{\lambda}$$

which would be obtained when  $v \rightarrow 0$ , to very precise information

$$(8.20) \quad p(\lambda) = \begin{cases} 1 & \text{if } \lambda = 1 \\ 0 & \text{elsewhere} \end{cases}$$

which is obtained when  $v \rightarrow \infty$ . Note that by using a  $\Gamma(av/2, v/2)$  we have more flexibility regarding the hyperprior of  $\lambda$  but at the cost of having to estimate an additional parameter.

Observe that we can obtain a sample from the prior distribution of  $\mathbf{w}$  by first sampling a  $\Gamma(v/2, v/2)$  distribution to obtain  $\lambda$ , then sample a  $\Gamma(1, \lambda/2)$  distribution  $N$  times to obtain  $\gamma_i$ ,  $i = 1, \dots, N$  and finally sample  $\mathcal{N}(0, \gamma_i)$  to obtain  $w_i$ .

We can now see a clear difference between how a realization from each of the prior distributions is obtained. While the  $\{\gamma_i\}$  in (8.12) and  $\{\alpha_i\}$  in (8.14) are obtained as realizations of independent distributions, the  $\{\gamma_i\}$  values in (8.16) all come from a common distribution (see Fig. (8.1)). The advantage of using the model in (8.12) and the model in (8.14) with  $a_i^\alpha = 1, b_i^\alpha = 0$  is that there is no need to estimate the parameter  $\lambda$  [84]. However, as we will see in the next section, the inference based on the model in (8.12) is a particular case of the one based on the model in (8.16) which leads to the Laplace prior model. We will also show with experimental results that the performance of our model is superior to the alternative prior models of the signal coefficients.

By combining the stages of the hierarchical Bayesian model, the joint distribution can finally be defined as  $p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta, \mathbf{y}) = p(\mathbf{y}|\mathbf{w}, \beta)p(\beta)p(\mathbf{w}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\lambda)p(\lambda)$ , where  $p(\mathbf{y}|\mathbf{w}, \beta)$ ,  $p(\beta)$ ,  $p(\mathbf{w}|\boldsymbol{\gamma})$ ,  $p(\boldsymbol{\gamma}|\lambda)$  and  $p(\lambda)$  are defined in (8.6), (8.7), (8.11), (8.16), and (8.18) respectively. The dependencies in this joint probability model are shown in graphical form in Fig. (8.1), where the arrows are used to denote the generative model. Note that the hierarchical structure can also be seen from the first four blocks from the left, which correspond to the variables  $v, \lambda, \gamma_1, \dots, \gamma_N$  and  $w_1, \dots, w_N$ .

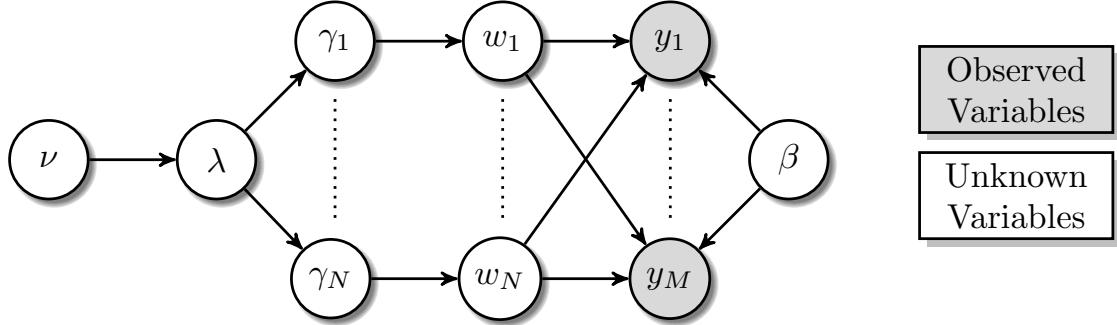


Figure 8.1. Directed acyclic graph representing the Bayesian model.

### 8.3. Bayesian Inference

As widely known, Bayesian inference is based on the posterior distribution

$$(8.21) \quad p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta | \mathbf{y}) = \frac{p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta, \mathbf{y})}{p(\mathbf{y})}.$$

However, the posterior  $p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta | \mathbf{y})$  is intractable, since

$$(8.22) \quad p(\mathbf{y}) = \int \int \int \int p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta, \mathbf{y}) d\mathbf{w} d\boldsymbol{\gamma} d\lambda d\beta$$

can not be calculated analytically. Therefore, approximation methods are utilized. We utilize the evidence procedure (type-II maximum likelihood approach) to perform Bayesian inference.

#### 8.3.1. Evidence Procedure

We will now derive the Bayesian inference using an evidence procedure with the conditional distribution in (8.6) and the priors in (8.11), (8.16) and (8.18). Our inference procedure is based

on the following decomposition

$$(8.23) \quad p(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta | \mathbf{y}) = p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda) p(\boldsymbol{\gamma}, \beta, \lambda | \mathbf{y}),$$

where the dependency on  $\nu$  is dropped for clarity. Since  $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda) \propto p(\mathbf{w}, \mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda)$ , then  $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda)$  is found to be a multivariate Gaussian distribution  $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with parameters

$$(8.24) \quad \boldsymbol{\mu} = \boldsymbol{\Sigma} \boldsymbol{\beta} \boldsymbol{\Phi}^T \mathbf{y},$$

$$(8.25) \quad \boldsymbol{\Sigma} = [\boldsymbol{\beta} \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{\Lambda}]^{-1},$$

with

$$(8.26) \quad \boldsymbol{\Lambda} = \text{diag}(1/\gamma_i).$$

We now utilize  $p(\boldsymbol{\gamma}, \beta, \lambda | \mathbf{y})$  in (8.23) to estimate the hyperparameters. In the type-II maximum likelihood procedure we represent  $p(\boldsymbol{\gamma}, \beta, \lambda | \mathbf{y})$  by a degenerate distribution where the distribution is replaced by a delta function at its mode, where we assume that this posterior distribution is sharply peaked around its mode [31]. Then, using  $p(\boldsymbol{\gamma}, \beta, \lambda | \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda)}{p(\mathbf{y})} \propto p(\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda)$ , we estimate the hyperparameters by the maxima of the joint distribution  $p(\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda)$  which is

obtained from  $p(\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}, \beta, \lambda)$  by integrating out  $\mathbf{w}$ . Consequently we have

$$\begin{aligned}
p(\mathbf{y}, \boldsymbol{\gamma}, \beta, \lambda) &= \int p(\mathbf{y}|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\lambda) p(\lambda) p(\beta) d\mathbf{w} \\
&= \left( \frac{1}{2\pi} \right)^{N/2} |\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^t|^{-1/2} \exp \left[ -\frac{1}{2} \mathbf{y}^t (\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^t)^{-1} \mathbf{y} \right] p(\boldsymbol{\gamma}|\lambda)p(\lambda)p(\beta) \\
(8.27) \quad &= \left( \frac{1}{2\pi} \right)^{N/2} |\mathbf{C}|^{-1/2} \exp \left[ -\frac{1}{2} \mathbf{y}^t \mathbf{C}^{-1} \mathbf{y} \right] p(\boldsymbol{\gamma}|\lambda)p(\lambda)p(\beta),
\end{aligned}$$

with  $\mathbf{C} = (\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^t)$ .

Instead of maximizing this distribution, we maximize equivalently its logarithm, which results in the following functional to be maximized

$$\begin{aligned}
\mathcal{L} &= -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{y}^t \mathbf{C}^{-1} \mathbf{y} + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_i \gamma_i \\
(8.28) \quad &+ \frac{\nu}{2} \log \frac{\nu}{2} - \log \Gamma(\nu/2) + \left( \frac{\nu}{2} - 1 \right) \log \lambda - \frac{\nu}{2} \lambda + (a^\beta - 1) \log \beta - b^\beta \beta.
\end{aligned}$$

Let us now state some equivalences that will help us in solving this maximization problem.

First we obtain

$$(8.29) \quad |\mathbf{C}| = |\Lambda|^{-1} |\beta^{-1}\mathbf{I}| |\Lambda + \beta\Phi^t\Phi| = |\Lambda|^{-1} |\beta^{-1}\mathbf{I}| |\Sigma^{-1}|$$

using the determinant identity [158] and thus

$$(8.30) \quad \log |\mathbf{C}| = -\log |\Lambda| - N \log \beta - \log |\Sigma|.$$

Furthermore using the Woodbury identity [98] we have

$$\begin{aligned}
 \mathbf{C}^{-1} &= (\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^t)^{-1} = \beta\mathbf{I} - \beta\Phi(\Lambda + \beta\Phi^t\Phi)^{-1}\Phi^t\beta \\
 (8.31) \quad &= \beta\mathbf{I} - \beta\Phi\Sigma\Phi^t\beta
 \end{aligned}$$

and therefore

$$\begin{aligned}
 \mathbf{y}^t\mathbf{C}^{-1}\mathbf{y} &= \beta\mathbf{y}^t\mathbf{y} - \beta\mathbf{y}^t\Phi\Sigma\Phi^t\beta\mathbf{y} \\
 &= \beta\mathbf{y}^t(\mathbf{y} - \Phi\mu) \\
 &= \beta\|\mathbf{y} - \Phi\mu\|^2 + \beta\mu^t\Phi^t(\mathbf{y} - \Phi\mu) \\
 (8.32) \quad &= \beta\|\mathbf{y} - \Phi\mu\|^2 + \mu^t\Lambda\mu.
 \end{aligned}$$

Using these identities, the derivative of  $\mathcal{L}$  with respect to  $\gamma_i$  is given by

$$(8.33) \quad \frac{d\mathcal{L}}{d\gamma_i} = \frac{1}{2} \left[ -\frac{1}{\gamma_i} + \frac{\langle w_i^2 \rangle}{\gamma_i^2} - \lambda \right]$$

where  $\langle w_i^2 \rangle = \mu_i^2 + \Sigma_{ii}$  with  $\Sigma_{ii}$  the  $i^{\text{th}}$  diagonal element of  $\Sigma$ . Setting this equal to zero results in

$$(8.34) \quad \gamma_i = -\frac{1}{2\lambda} + \sqrt{\frac{1}{4\lambda^2} + \frac{\langle \omega_i^2 \rangle}{\lambda}}.$$

The updates of the other hyperparameters are found similarly by taking the derivative of (8.28) with respect to each hyperparameter and setting it equal to zero. The updates found in

this manner are given by

$$(8.35) \quad \lambda = \frac{N + v/2 - 1}{\sum_i \gamma_i/2 + v/2}$$

$$(8.36) \quad \beta = \frac{N/2 + a^\beta}{< \| \mathbf{y} - \Phi \mathbf{w} \| ^2 >/2 + b^\beta}$$

where the expected value is calculated with respect to the conditional distribution of  $\mathbf{w}$ .

Finally, we can also estimate  $v$  by maximizing (8.28) with respect to  $v$ . This results in solving the following equation

$$(8.37) \quad \log \frac{v}{2} + 1 - \psi\left(\frac{v}{2}\right) + \log \lambda - \lambda = 0$$

This equation does not have a closed-form solution so it is solved numerically.

In summary, at each iteration of the algorithm, given an estimate of  $\gamma$ ,  $\beta$ , and  $\lambda$ , the estimate of the distribution of  $\mathbf{w}$  is calculated using (8.25) and (8.24), followed by the estimation of the variances  $\gamma_i$  from (8.34), the hyperparameter  $\lambda$  from (8.35), the noise inverse variance (precision)  $\beta$  from (8.36) and  $v$  from (8.37), where the expected values needed in these equations are calculated using the current distribution of  $\mathbf{w}$ .

Note that the same update equations can be obtained by applying an expectation-maximization (EM) procedure instead of the direct maximization method employed in this section. Fixed point iterations [152] can also be applied to find  $\gamma$ ,  $\beta$ , and  $\lambda$ . Note also that a similar optimization procedure is used in [168] for a different modeling of the signal.

### 8.3.2. Fast Suboptimal Solutions

There is a major disadvantage of the method presented in the previous section, namely, it requires the solution of a linear system of  $N$  equations in (8.24), which requires  $\mathbf{O}(N^3)$  computations. Moreover, since the system in (8.2) is overdetermined with  $M \ll N$ , numerical errors create major practical difficulties in solving this system. Although the matrix  $\Sigma$  can be written using the Woodbury matrix identity as follows

$$\begin{aligned} \Sigma &= \Lambda^{-1} - \Lambda^{-1}\Phi^t(\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^t)^{-1}\Phi\Lambda^{-1} \\ (8.38) \quad &= \Lambda^{-1} - \Lambda^{-1}\Phi^t\mathbf{C}^{-1}\Phi\Lambda^{-1} \end{aligned}$$

which requires the solution of only  $M$  linear equations, therefore  $\mathbf{O}(M^3)$  time, this is in practice more problematic due to numerical errors and it still does not scale up well for large-scale problems. Therefore, the algorithm presented in the previous section cannot be easily applied to practical problems, but it will serve us as the starting point in developing a practical algorithm as follows.

To promote sparsity and to decrease the computational requirements, only a single  $\gamma_i$  will be updated at each iteration of the algorithm instead of updating the whole vector  $\boldsymbol{\gamma}$ . As will be shown later, updating a single hyperparameter leads to very efficient updates of the matrix  $\Sigma$  and the mean  $\mu$ . A fundamental observation is that if a single hyperparameter  $\gamma_i$  is set equal to zero,  $\mu_i$  must be equal to 0, and so the corresponding entry is pruned out from the model. Since it is assumed that the vector  $\mathbf{w}$  is sparse, many of its components are zero, therefore most  $\gamma_i$ 's are set equal to zero, and matrix  $\Sigma$  can be represented using fewer dimensions than  $N \times N$ . Exploiting these properties, one can obtain a much more efficient procedure than the algorithm

presented in the previous section, by starting with an “empty” model ( $\boldsymbol{\gamma} = 0$ ) and iteratively adding components to the model. In the following we will present such a procedure.

A fundamental observation to obtain the fast suboptimal solution is that the matrix  $\mathbf{C}$  in (8.28) can be written as follows:

$$\begin{aligned}
 \mathbf{C} &= \beta^{-1} \mathbf{I} + \sum_i \gamma_i \phi_i \phi_i^t \\
 &= \beta^{-1} \mathbf{I} + \sum_{j \neq i} \gamma_j \phi_j \phi_j^t + \gamma_i \phi_i \phi_i^t \\
 (8.39) \quad &= \mathbf{C}_{-i} + \gamma_i \phi_i \phi_i^t,
 \end{aligned}$$

where  $\mathbf{C}_{-i}$  denotes that the contribution of the  $i^{th}$  basis is not included. Using the Woodbury identity in (8.39) we obtain

$$(8.40) \quad \mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1} \phi_i \phi_i^t \mathbf{C}_{-i}^{-1}}{1/\gamma_i + \phi_i^t \mathbf{C}_{-i}^{-1} \phi_i}$$

and using the determinant identity we obtain

$$(8.41) \quad |\mathbf{C}| = |\mathbf{C}_{-i}| |1 + \gamma_i \phi_i^t \mathbf{C}_{-i}^{-1} \phi_i|.$$

Substituting the last two equations in (8.28) and treating  $\mathcal{L}$  as a function of  $\boldsymbol{\gamma}$  only, we obtain

$$(8.42) \quad \mathcal{L}(\boldsymbol{\gamma}) = -\frac{1}{2} \left[ \log |\mathbf{C}_{-i}| + \mathbf{y}^t \mathbf{C}_{-i}^{-1} \mathbf{y} + \frac{\lambda}{2} \sum_{j \neq i} \gamma_j \right] + \frac{1}{2} \left[ \log \frac{1}{1 + \gamma_i s_i} + \frac{q_i^2 \gamma_i}{1 + \gamma_i s_i} - \lambda \gamma_i \right]$$

$$(8.43) \quad = \mathcal{L}(\boldsymbol{\gamma}_{-i}) + l(\gamma_i)$$

where  $l(\gamma) = \frac{1}{2} \left[ \log \frac{1}{1+\gamma s_i} + \frac{q_i^2 \gamma}{1+\gamma s_i} - \lambda \gamma \right]$  and  $q_i$  and  $s_i$  are defined as

$$(8.44) \quad s_i = \phi_i^t \mathbf{C}_{-i}^{-1} \phi_i,$$

$$(8.45) \quad q_i = \phi_i^t \mathbf{C}_{-i}^{-1} \mathbf{y}$$

Note that the quantities  $q_i$  and  $s_i$  do not depend on  $\gamma_i$  (since  $\mathbf{C}_{-i}^{-1}$  is independent of  $\gamma_i$ ). Therefore, the terms related to a single hyperparameter  $\gamma_i$  are now separated from others. Let us now examine if the  $i^{th}$  basis should be included. A closed form solution of the maximum of  $\mathcal{L}(\boldsymbol{\gamma})$ , when only its  $i^{th}$  component is changed, can be found by holding other hyperparameters fixed, taking its derivative with respect to  $\gamma_i$  and setting it equal to zero. The derivative of  $\mathcal{L}(\boldsymbol{\gamma})$  with respect to  $\gamma_i$  can be expressed as

$$(8.46) \quad \begin{aligned} \frac{d\mathcal{L}(\boldsymbol{\gamma})}{d\gamma_i} &= \frac{dl(\gamma_i)}{d\gamma_i} = \frac{1}{2} \left[ -\frac{s_i}{1+\gamma_i s_i} + \frac{q_i^2}{(1+\gamma_i s_i)^2} - \lambda \right], \\ &= -\frac{1}{2} \left[ \frac{\gamma_i^2 (\lambda s_i^2) + \gamma_i (s_i^2 + 2\lambda s_i) + (\lambda + s_i - q_i^2)}{(1+\gamma_i s_i)^2} \right] \end{aligned}$$

Note that the numerator has a quadratic form while the denominator is always positive, and therefore  $\frac{dl(\gamma_i)}{d\gamma_i} = 0$  is satisfied at

$$(8.47) \quad \gamma_i = \frac{-s_i(s_i + 2\lambda) \pm s_i \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i^2}$$

$$(8.48) \quad = \frac{-s_i(s_i + 2\lambda) \pm s_i \sqrt{\Delta}}{2\lambda s_i^2}$$

where  $\Delta = (s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda) > 0$ . Observe that if  $q_i^2 - s_i < \lambda$ , then  $\Delta^2 < s_i + 2\lambda$  and both solutions in (8.48) are negative, and since  $\frac{dl(\gamma_i)}{d\gamma_i}|_{\gamma_i=0} < 0$ , the maximum occurs at  $\gamma_i = 0$ .

On the other hand if  $q_i^2 - s_i > \lambda$ , there are two real solutions, one negative and the variance estimate

$$(8.49) \quad \gamma_i = \frac{-s_i(s_i + 2\lambda) + s_i \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i^2}.$$

Since when  $q_i^2 - s_i > \lambda$  we have  $\frac{dl(\gamma_i)}{d\gamma_i}|_{\gamma_i=0} > 0$  and  $\frac{dl(\gamma_i)}{d\gamma_i}|_{\gamma_i=\infty} < 0$ , the obtained variance estimate in (8.49) maximizes  $l(\gamma_i)$  and therefore  $\mathcal{L}(\boldsymbol{\gamma})$ .

In summary, the maximum of  $\mathcal{L}(\boldsymbol{\gamma})$ , when all components of  $\boldsymbol{\gamma}$  except  $\gamma_i$  are kept fixed, is achieved at

$$(8.50) \quad \gamma_i = \begin{cases} \frac{-s_i(s_i + 2\lambda) + s_i \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i^2} & \text{if } q_i^2 - s_i > \lambda \\ 0 & \text{otherwise} \end{cases}$$

Note that in the case of  $\gamma_i = 0$ , the corresponding basis  $\phi_i$  is pruned out from the model and  $\mu_i$  is set equal to zero. Therefore (8.50) provides a systematic method of deciding which basis vectors should be included in the model and which should be excluded. Note that as in the previous section, the estimate of  $\lambda$  is provided by (8.35).

It is crucial for computational efficiency that once a hyperparameter  $\gamma_i$  is updated using (8.50), the quantities  $s_i$ ,  $q_i$ ,  $\mu$  and  $\Sigma$  are efficiently updated. Similarly to [227], the parameters

$q_i$  and  $s_i$  can be calculated for all basis vectors  $\phi_i$  efficiently using the following identities

$$(8.51) \quad S_i = \beta \phi_i^t \phi_i - \beta^2 \phi_i^t \Phi \Sigma \Phi^t \phi_i$$

$$(8.52) \quad Q_i = \beta \phi_i^t \mathbf{y} - \beta^2 \phi_i^t \Phi \Sigma \Phi^t \mathbf{y}$$

$$(8.53) \quad s_i = \frac{S_i}{1 - \gamma_i S_i}$$

$$(8.54) \quad q_i = \frac{Q_i}{1 - \gamma_i S_i},$$

where  $\Sigma$  and  $\Phi$  include only the columns  $i$  that are included in the model ( $\gamma_i \neq 0$ ). Moreover,  $\Sigma$  and  $\mu$  can be updated very efficiently when only a single coefficient  $\gamma_i$  is considered, as in [227]. Utilizing these equations, we can obtain an iterative procedure by updating one hyperparameter  $\gamma_i$  at each iteration, and updating  $s_i$ ,  $q_i$ ,  $\mu$  and  $\Sigma$  accordingly. The procedure is summarized below in Algorithm 10.

At step 5 of the algorithm, a candidate  $\gamma_i$  must be selected for updating. This can be done by randomly choosing a basis vector  $\phi_i$ , or by calculating each  $\gamma_i$  and choosing the one that results in the greatest increase in  $\mathcal{L}(\boldsymbol{\gamma})$  in (8.42), which results in a faster convergence. The latter is the method implemented in this work. Finally, the updates of  $\Sigma$ ,  $\mu$ ,  $s_i$ , and  $q_i$  in the add, delete, and re-estimate operations are the same as those in the RVM formulation (see Appendix A in [227] for details).

An important step in the algorithm is the estimation of the noise precision  $\beta$ , which is done in the previous section using (8.36). Unfortunately, this method cannot be used in practice in this fast algorithm since the proposed algorithm is constructive and the reconstruction and, therefore, the estimate in (8.36) are unreliable at early iterations. Due to the under-determined nature of the compressive sensing problem, once the estimate of  $\beta$  is very far from its true

Table 8.1. Proposed Algorithm

**Algorithm 10.** *Fast Laplace**INPUTS:*  $\Phi, \mathbf{y}$ *OUTPUTS:*  $\mathbf{w}, \Sigma, \gamma$ *Initialize all  $\gamma_i = 0, \lambda = 0$* *Convergence criterion not met*

- (1) *Choose a  $\gamma_i$  (or equivalently choose a basis vector  $\phi_i$ )*
- (2) *if  $q_i^2 - s_i > \lambda$  AND  $\gamma_i = 0$*   
*Add  $\gamma_i$  to the model*  
*else if  $q_i^2 - s_i > \lambda$  AND  $\gamma_i > 0$*   
*Re-estimate  $\gamma_i$*   
*else if  $q_i^2 - s_i < \lambda$*   
*Prune  $i$  from the model (set  $\gamma_i = 0$ )*
- (3) *Update  $\Sigma$  and  $\mu$*
- (4) *Update  $s_i, q_i$*
- (5) *Update  $\lambda$  using (8.35)*
- (6) *Update  $\nu$  using (8.37)*

value, the reconstruction quality is also significantly affected. Therefore, we fix the estimate of this parameter in the beginning of the algorithm using  $\beta = 0.01 \|\mathbf{y}\|_2^2$  inspired by [116, 83]. Alternatively, this parameter can be integrated out from the model as in [115].

Note that unlike other constructive (or greedy) methods such as OMP [232], StOMP [74], and Gradient Pursuit methods [34], included basis vectors can also be deleted once they are determined to be irrelevant. This is a powerful feature of the algorithm, since errors in the beginning of the reconstruction process can be fixed in later stages by effectively pruning out irrelevant basis vectors which can drive the algorithm away from the optimal result.

Let us complete this section by comparing the variance estimates provided by the relevance vector machine (where  $\lambda = 0$ ) with the ones provided by the proposed method in terms of

sparsity. The estimate  $\gamma_i^{\text{RVM}}$  in the RVM framework is given by [227]

$$(8.55) \quad \gamma_i^{\text{RVM}} = \operatorname{argmax}_{\gamma_i} \frac{1}{2} \left[ \log \frac{1}{1 + \gamma_i s_i} + \frac{q_i^2 \gamma_i}{1 + \gamma_i s_i} \right]$$

while as we have seen the estimate provided by the modeling using the Laplace distribution is given by

$$(8.56) \quad \begin{aligned} \gamma_i^{\text{L}} &= \operatorname{argmax}_{\gamma_i} l(\gamma_i) \\ &= \operatorname{argmax}_{\gamma_i} \frac{1}{2} \left[ \log \frac{1}{1 + \gamma_i s_i} + \frac{q_i^2 \gamma_i}{1 + \gamma_i s_i} - \lambda \gamma_i \right] \end{aligned}$$

Clearly the RVM model corresponds to the particular case of  $\lambda = 0$  in our model. The solution of (8.55) is given by

$$(8.57) \quad \gamma_i^{\text{RVM}} = \begin{cases} \frac{q_i^2 - s_i}{s_i^2} & \text{if } q_i^2 - s_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Let us now examine the difference  $\gamma_i^{\text{RVM}} - \gamma_i^{\text{L}}$ . When  $q_i^2 - s_i < \lambda$  we have

$$(8.58) \quad \gamma_i^{\text{RVM}} - \gamma_i^{\text{L}} = \begin{cases} 0 & \text{if } q_i^2 - s_i < 0 \\ \gamma_i^{\text{RVM}} & \text{if } 0 \leq q_i^2 - s_i < \lambda \end{cases}$$

When  $q_i^2 - s_i \geq \lambda$ , the derivative of the function  $l(\gamma_i)$  at  $\gamma_i = \gamma_i^{\text{RVM}}$  is  $-\lambda < 0$ . Since  $\frac{dl(\gamma_i)}{d\gamma_i}|_{\gamma_i=0} > 0$ , the maximum of  $l(\gamma_i)$  occurs at a smaller value  $\gamma_i^{\text{L}}$  than  $\gamma_i^{\text{RVM}}$ . Consequently we always have

$$(8.59) \quad \gamma_i^{\text{RVM}} \geq \gamma_i^{\text{L}}$$

Therefore, the estimates  $\gamma_i^L$  using the Laplace prior are always smaller than the estimates  $\gamma_i^{\text{RVM}}$  of the relevance vector machine. Note also that compared to RVM more components will possibly be pruned out from the model when  $\lambda > 0$ , since the cardinality of the set  $\{w_i\}$  for which  $q_i^2 - s_i > \lambda$  is smaller than of that of the set  $\{w_i\}$  for which  $q_i^2 - s_i > 0$ . These observations imply that the solution obtained by the proposed method is at least as sparse as the one provided by the RVM. This will also be shown empirically in Section 8.4.

## 8.4. Experiments

In this section we present experimental results with both one-dimensional (1D) synthetic signals and 2D images to demonstrate the performance of the proposed method. We considered experimental setups used widely in the literature. In the experiments reported below, we concentrated on the fast algorithm presented in Section 8.3.2 due to its wider applicability in practical settings. Although it is suboptimal in theory, it provides better reconstruction results than the algorithm in Section 8.3.1 since the computational cost and increased numerical errors render the optimal algorithm impractical. This is especially evident when applying compressive sensing reconstruction algorithms to large-scale problems, such as images. Note that this is also observed when applying RVM to machine learning problems [225, 227] and to CS [116].

### 8.4.1. 1D Synthetic Signals

We use the following default setup in the experimental results reported in this section. Four different types of signals of length  $N$  are generated, where  $T$  coefficients at random locations of the signals are drawn from five different probability distributions, and the rest ( $N - T$ ) of the coefficients are set equal to zero. The nonzero coefficients of the sparse signals are realizations

of the following five distributions: 1) Uniform  $\pm 1$  random spikes, 2) zero-mean unit variance Gaussian, 3) unit variance Laplace, and 4) Student's t with 3 degrees of freedom.

As the measurement matrix  $\Phi$  we chose a uniform spherical ensemble, where the columns  $\phi_i$  are uniformly distributed on the sphere  $R^N$ . Other measurement matrices such as partial Fourier and uniform random projection (URP) ensembles gave similar results and therefore they are not reported here.

In the experiments we fix  $N = 512$  and  $T = 20$  and vary the number of measurements  $M$  from 40 to 120 in steps of 5. Moreover, we present results with noiseless and noisy acquisitions, where for the noisy observations we added zero mean white Gaussian noise with standard deviation 0.03. We repeated each experiment 100 times and report the average of all experiments.

In the first set of experiments, we compare the effect of different choices of the parameter  $\lambda$  on the reconstruction performance. We ran the algorithm presented in Section 8.3.2 with  $\lambda = 0$ ,  $\lambda = 1$ ,  $\lambda = 10$ , and  $\lambda$  estimated using (8.35). As mentioned in Section 8.2,  $\lambda = 0$  corresponds to the RVM formulation [227] which will be denoted by BCS following [116]. Moreover, we show results when the parameter  $v$  is set equal to zero and when it is also estimated automatically using (8.37).

The reconstruction error is calculated as  $\| \hat{\mathbf{w}} - \mathbf{w} \|_2^2 / \| \mathbf{w} \|_2^2$ , where  $\hat{\mathbf{w}}$  and  $\mathbf{w}$  are the estimated and true coefficient vectors, respectively. The criterion  $\| \mathcal{L}(\boldsymbol{\gamma}^k) - \mathcal{L}(\boldsymbol{\gamma}^{k-1}) \| < 10^{-4}$  is used to terminate the iterative procedure.

Average reconstruction errors in 100 runs are shown for the noise-free case in Fig. (8.2) for all types of signals. It is clear that using nonzero values for  $\lambda$  results in lower reconstruction errors with all types of signals, and the  $\lambda = 0$  case (BCS) gives the worst reconstruction error. Even arbitrarily selected nonzero values of  $\lambda$  (see cases with  $\lambda = 1$  and  $\lambda = 10$ ) result in

better error rates. Automatically estimating  $\lambda$  using (8.35) results in the best reconstruction performance.

It is interesting to note that estimating the parameter  $v$  automatically using (8.37) results in slightly worse performance than setting it equal to zero. This suggests that the elements of  $\gamma$  can be used to estimate  $\lambda$  in combination with the improper prior  $p(\lambda) \propto 1/\lambda$ . In other words, not much more knowledge than the estimated  $\gamma$  is needed to estimate  $\lambda$ . Therefore  $v$  is fixed equal to zero in the remaining experiments.

The results of the same experimental setup with additive observation noise (zero mean white Gaussian noise with standard deviation 0.03) are shown in Fig. (8.3). Similar performance increases by using nonzero values of  $\lambda$  can be observed, with again estimating  $\lambda$  using (8.35) resulting in the best performance. Note that although the algorithms result in higher reconstruction errors than in the noise-free case and perfect reconstruction is not attained, good reconstruction performances are still obtained.

In summary, the experimental results suggest that the proposed framework clearly provides improved reconstruction performance over the RVM framework with only a slight difference in computations due to the calculation of (8.35).

In the second set of experiments, we repeat the same experiment and compare the proposed method with the algorithms BCS [116], BP [64], OMP [232], StOMP with CFAR thresholding (denoted by FAR) [74], and GPSR [83]. For all algorithms, their MATLAB implementations in the corresponding websites are used. The required algorithm parameters are set according to their default setups and in some cases adjusted for improved performance. The algorithms BCS, OMP, and FAR are greedy constructive algorithms like the proposed method, and the algorithms BP and GPSR are global optimization algorithms. We ran the GPSR method both with and

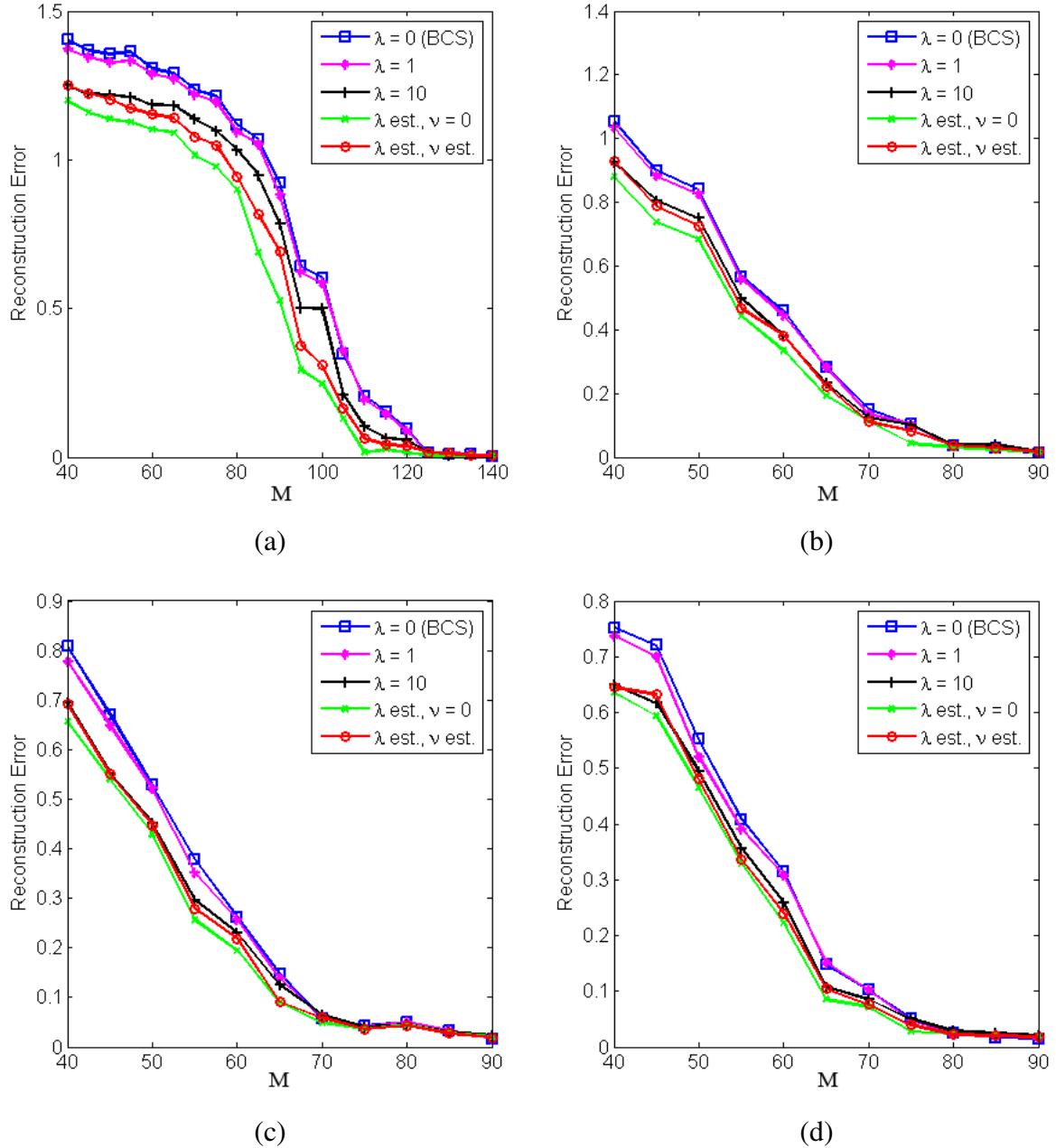


Figure 8.2. Number of measurements  $M$  vs reconstruction error for the noise-free case resulting from different values of  $\lambda$ . (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown as the error rates are negligible.

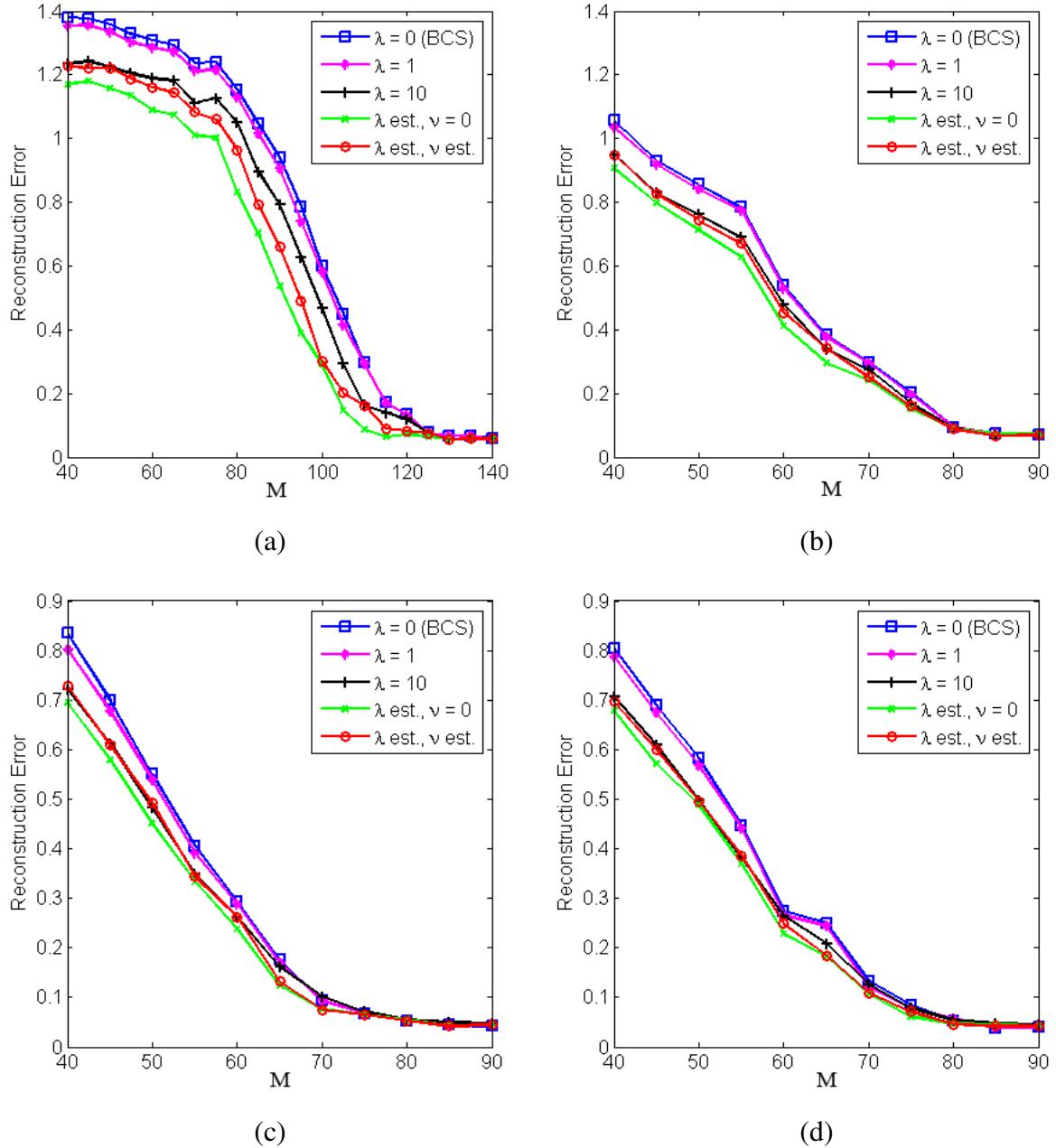


Figure 8.3. Number of measurements  $M$  vs reconstruction error with noisy observations with different values of  $\lambda$ . (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown as the error rates converged.

without the “debiasing” option, and reported the best result. In the results reported below, Laplace denotes the proposed method where  $\lambda$  is estimated using (8.35) and the parameter  $v$  is set equal to 0.

Average reconstruction errors of 100 runs are shown for the noise-free case in Fig. (8.4) for all types of signals. It is clear that the proposed algorithm outperforms all other methods in terms of reconstruction error except for the first signal, for which it provides the best performance after BP and GPSR. However, BP and GPSR result in worse performance than other methods for the rest of the signals. Note that with both algorithms we tuned the algorithm parameters by trial-and-error to achieve their best performance. On the other hand, both BCS and the proposed method do not require parameter tuning. Despite this fact, note that the proposed method provides the best overall performance among all methods.

Examples of reconstructions of the uniform spikes signal are shown in Fig. (8.5). An important property of the Bayesian methods BCS and Laplace is that they provide the posterior distribution of the unknown signal rather than point estimates. This distribution estimate can be used to provide uncertainty estimates of the coefficients using the covariance matrix  $\Sigma$ , which are shown as error-bars in Fig. (8.5). These error-bars are variance estimates of the coefficients corresponding to the diagonal elements  $\Sigma_{ii}$  of matrix  $\Sigma$ . Besides being a measure of the uncertainty of the reconstruction, the covariance matrix  $\Sigma$  can also be used to adaptively design the measurement matrix  $\Phi$  [116, 209].

We repeat the same experiment this time with additive observation noise (zero mean white Gaussian noise with standard deviation 0.03). Average reconstruction errors of 100 runs are shown in Fig. (8.6). The reconstruction errors obtained by the algorithms are slightly higher than in the noise-free case, and even with a high number of measurements exact reconstructions

are not obtained. However, the algorithms still provide accurate reconstructions with a low error rate. Note that the proposed method Laplace again provides the best overall performance for a reasonable number of observations.

For the 1D experiments reported in this section, the average running times are around 0.1 s for BCS and Laplace, around 0.15 s for BP, and around 0.01 s for the other methods. Therefore, the proposed method and BCS are computationally slightly more demanding than other methods except BP, but such differences are small and they are considered justified considering the improvement in error rates obtained by the proposed method. As will be shown in the next section, the proposed method is computationally very competitive when applied to larger scale problems, such as images.

#### 8.4.2. Images

In this section we present a comparison between the proposed method and a number of existing methods on a widely used experimental setup, namely the multiscale CS reconstruction [233] of the Mondrian image. We adapted the same test parameters as in the SparseLab package [1], where the number of samples are  $N = 4096$ , the number of measurements are  $M = 2073$ , and the measurement matrices are drawn from a uniform spherical distribution. The multiscale CS scheme is applied on the wavelet transform of the image with a “symmlet8” wavelet with the coarsest scale 4 and finest scale 6. We compared the performance of the algorithms BP, BCS, and StOMP with CFAR and CFDR thresholding with the proposed method. The parameters of the algorithms BP, CFAR and CFDR are chosen as in the SparseLab package. As in the previous section, the parameters of BCS and the proposed method are solely estimated from the measurements.

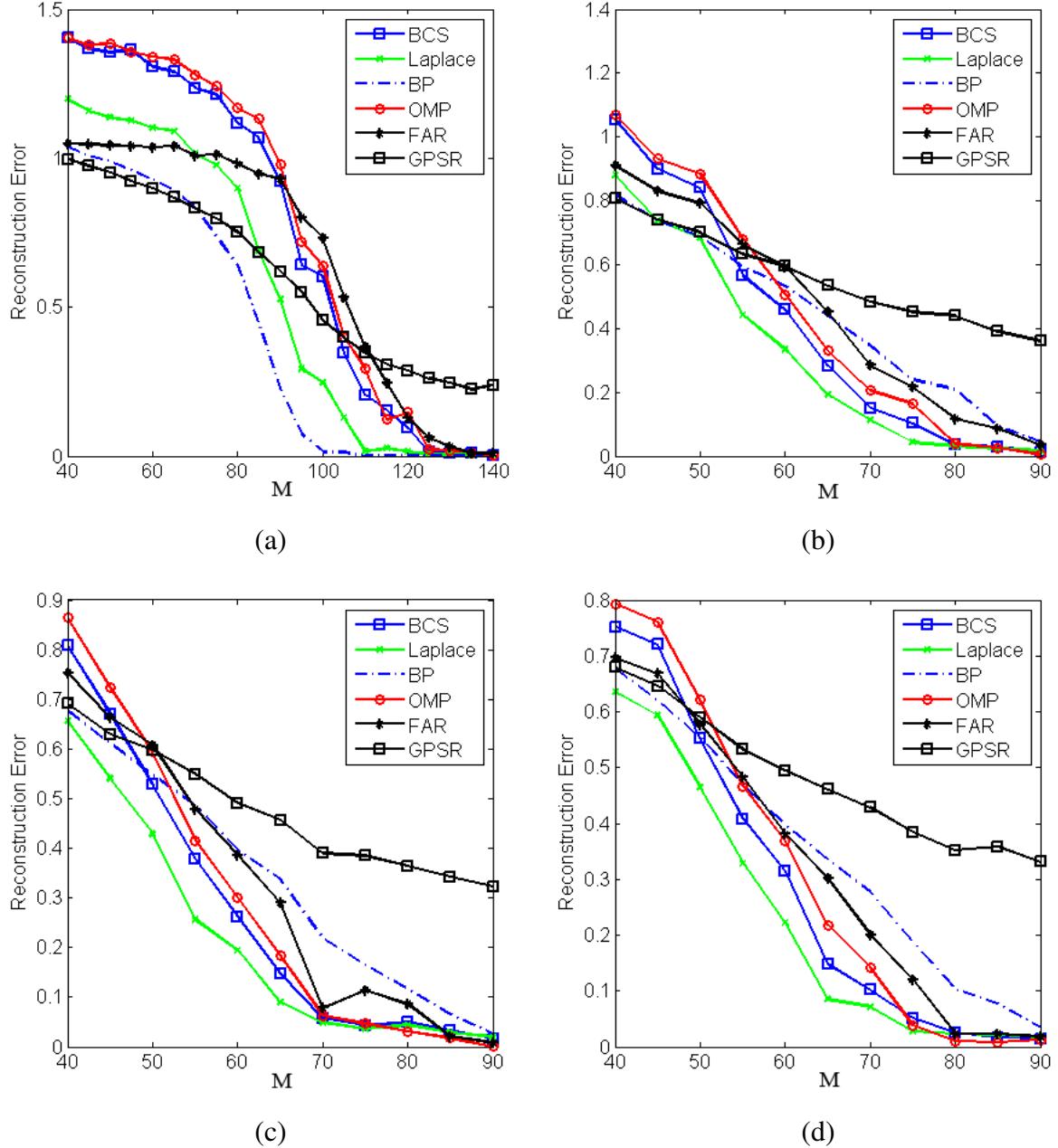


Figure 8.4. Number of measurements  $M$  vs reconstruction error for the noise-free case for different algorithms. (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown as the error rates converged.

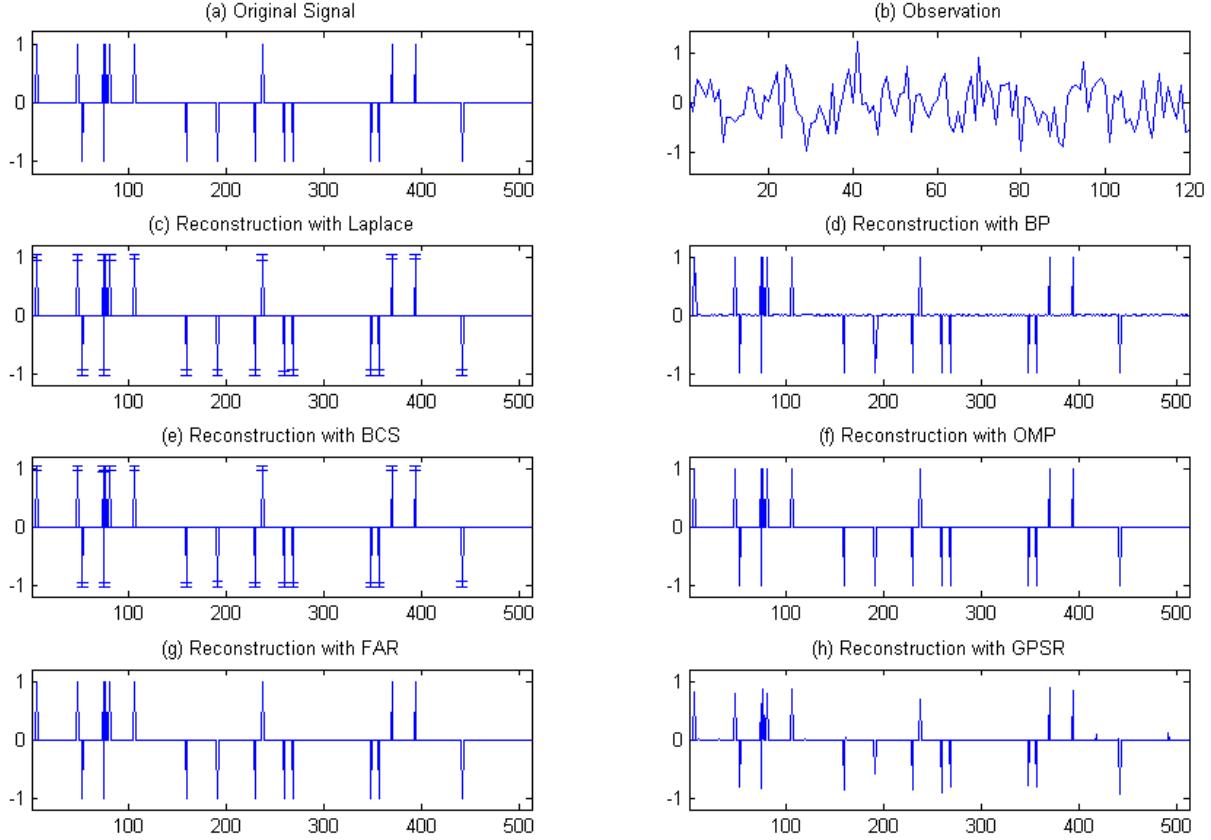


Figure 8.5. Reconstruction of uniform spikes signal with  $N=512$ ,  $M=120$ , and  $T=20$ . (a) Original Signal, (b) Observation, Reconstructions with (c) Laplace, (d) BP, (e) BCS, (f) OMP, (g) FAR, and (h) GPSR. All reconstructions have negligible errors except GPSR with reconstruction error = 0.2186. The error bars in (c) and (e) correspond to the estimated variances of the coefficients.

The reconstruction error is calculated as  $\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2 / \|\mathbf{f}\|_2^2$ , where  $\hat{\mathbf{f}}$  and  $\mathbf{f}$  are the estimated and true images, respectively. Since the measurement matrices are random, the experiment is repeated 100 times and their average is reported. Average reconstruction errors, running times and the number of nonzero components in the reconstructed images are shown in Table 8.2, where “Linear” denotes linear reconstruction with  $M = 4096$  measurements and represents the

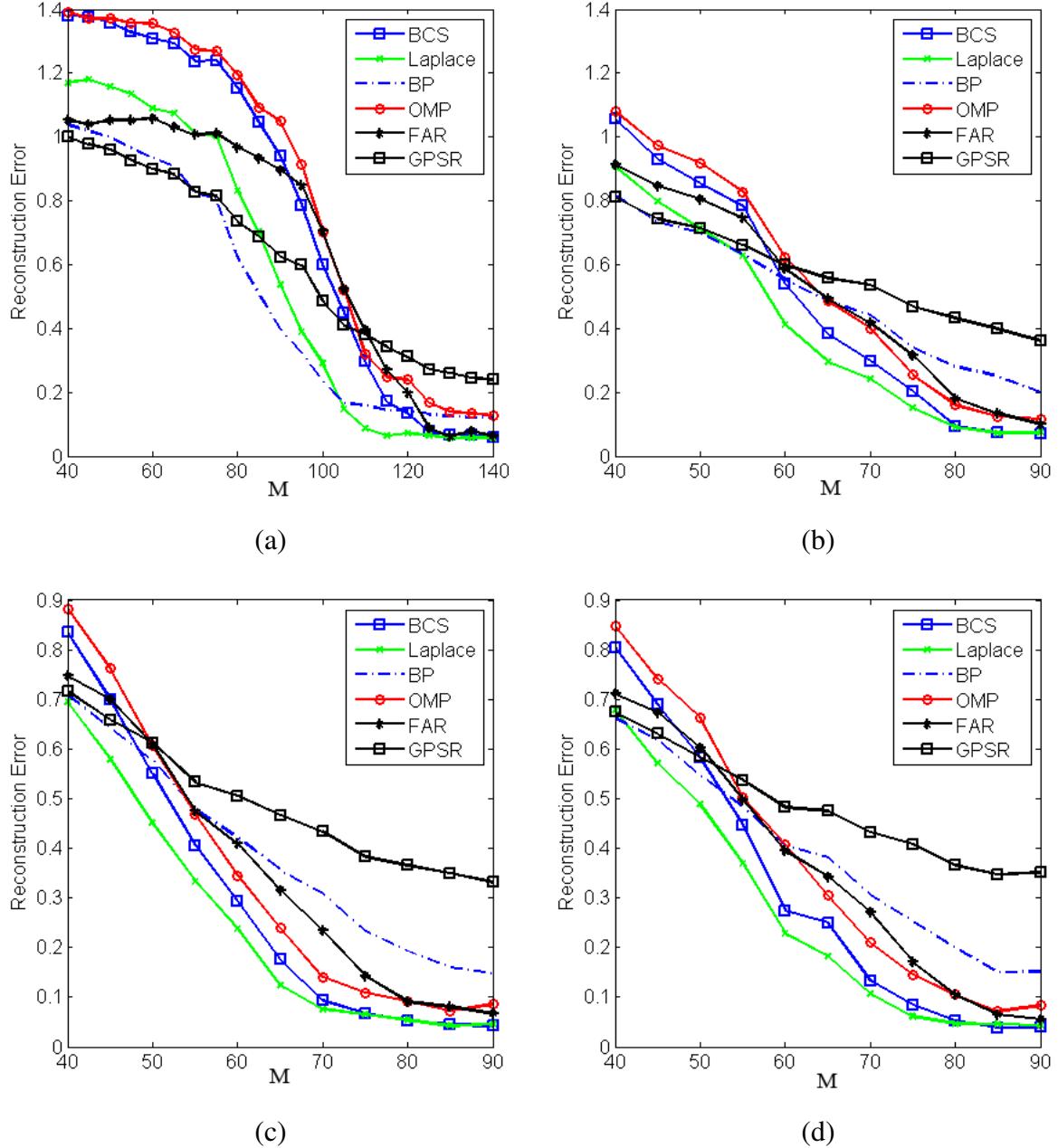


Figure 8.6. Number of measurements  $M$  vs reconstruction error with noisy observations for different algorithms. (a) Uniform spikes  $\pm 1$ ; Nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) Student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to  $M > 90$  are not shown for clarity as the error rates converged.

Table 8.2. Average reconstruction errors, running times and number of nonzero components for multi-scale CS reconstruction of the *Mondrian* image.

	Mondrian		
	# Nonzeros	Time (s)	Error
Linear	4096	-	0.13325
BP	4096	78.254	0.13933
CFAR	1139.2	13.88	0.14971
CFDR	2177.3	7.86	0.20867
BCS	1174.2	18.343	0.1443
Laplace	1078.7	15.372	0.1451

best reconstruction performance that can be achieved. It is clear that although BCS and Laplace have nearly the same error rate, Laplace is faster and the reconstructed image is sparser. In fact, Laplace provides the sparsest reconstructed image among all methods. The CFDR method, although it is the fastest, has the worst reconstruction error, and the BP method, although it has the best reconstruction error, has the largest computation time. Laplace and CFAR are clearly the methods that should be preferred, having near-best reconstruction errors and smallest computation times, where CFAR being slightly faster and Laplace having slightly lower reconstruction error. Examples of reconstructed images are shown in Fig. (8.7) where it can be observed that these methods provide fairly good reconstructions.

In summary, experimental results with both 1D synthetic signals and 2D images show that the proposed method provides improved performance in reconstruction quality with competitive performance in computational resources.

## 8.5. Conclusions

In this chapter we formulated the compressing sensing problem from a Bayesian perspective, and presented a framework to simultaneously model and estimate the sparse signal coefficients. Using this framework, we compared different sparsity priors, and proposed the use of

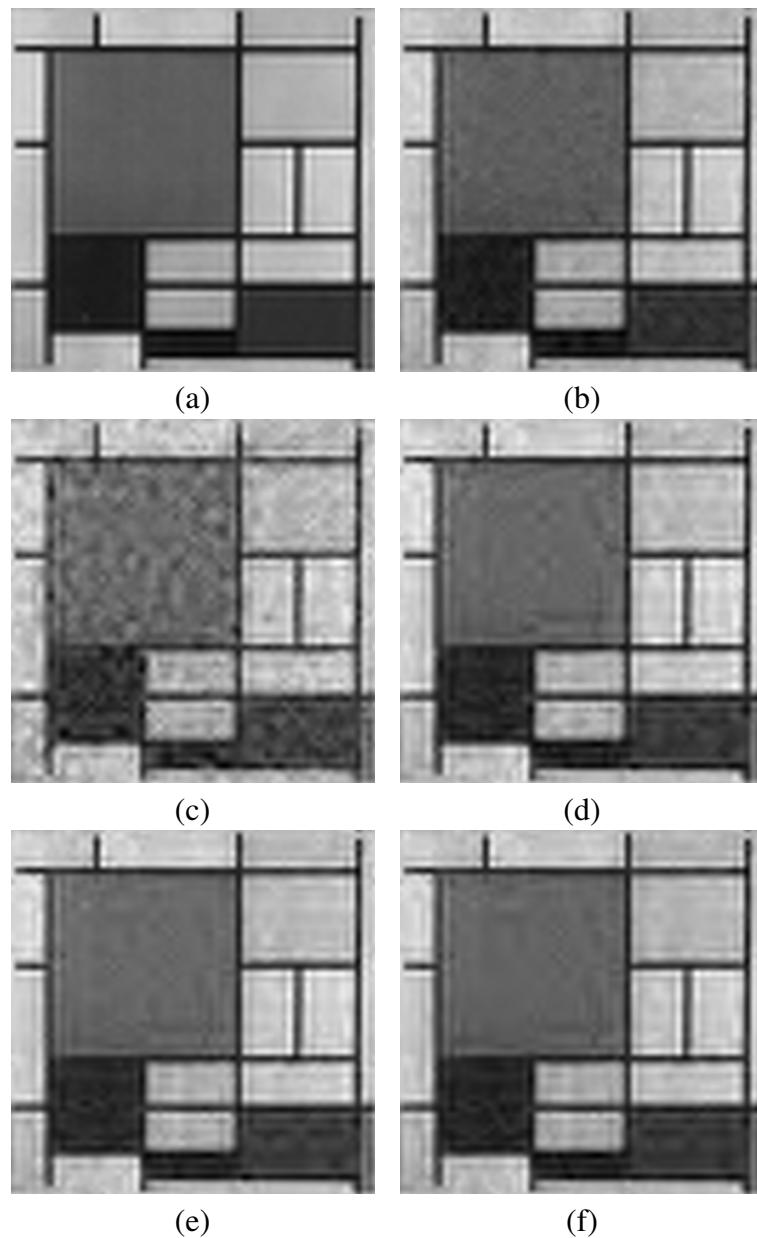


Figure 8.7. Examples of reconstructed Mondrian images using a multi-scale compressive sensing scheme by (a) linear reconstruction (error = 0.13325), (b) BP (error = 0.13874, time = 76.555 s, no. of nonzero components = 4096), (c) StOMP with FDR thresholding (error = 0.1747, time = 6.48 s, no. of nonzero components = 2032), (d) StOMP with FAR thresholding (error = 0.14673, time = 19.759 s, no. of nonzero components = 1196), (e) BCS (error = 0.14233, time = 16.086 s, no. of nonzero components = 1145) and (f) Laplace (error = 0.14234, time = 15.982, no. of nonzero components = 1125).

a hierarchical form of Laplace priors on signal coefficients. We have shown that the relevance vector machine is a special case of our formulation, and that our hierarchical prior modeling provides solutions with a higher degree of sparsity and lower reconstruction errors. We proposed a constructive (greedy) algorithm resulting from this formulation, which updates the signal coefficients sequentially in order to achieve low computation times and efficiency in practical problems. The proposed algorithm estimates the unknown signal coefficients simultaneously along with the unknown model parameters. The model parameters are estimated solely from the observation and therefore the proposed algorithm does not require user intervention unlike most existing methods. We demonstrated that overall, the proposed algorithm results in higher performance than most state-of-the-art algorithms. Moreover, the proposed method provides estimates to the distributions of the unknown signal which can be used to measure their uncertainty. The theoretical framework and the proposed algorithm are easy to implement and generalize to investigate further uses of the Bayesian framework in compressive sensing.

## CHAPTER 9

### Bayesian Compressive Sensing Using Non-Convex Priors

#### 9.1. Introduction

In Chapter 8, we developed a Bayesian compressive sensing reconstruction method with Laplace priors, which is equivalent to  $l_1$  regularized optimization problem given by

$$(9.1) \quad \hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \{ \| \mathbf{y} - \Phi \mathbf{x} \|_2^2 + \tau \| \mathbf{x} \|_1 \},$$

However, a more general form of the optimization problem can be obtained using  $l_p$ -norms, that is,

$$(9.2) \quad \hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \{ \| \mathbf{y} - \Phi \mathbf{x} \|_2^2 + \tau \| \mathbf{x} \|_p^p \},$$

where  $p$  is generally chosen to be within the interval  $[0, 2]$ .

The focus of this chapter is to formulate this problem in a Bayesian framework. A class of algorithms utilize  $l_p$ -norms with  $0 \leq p < 1$  with potentially higher accuracy in recovery than  $l_1$  norms [59]. The reconstruction performance of these algorithms is demonstrated with both theoretically and experimentally. This class of algorithms utilizing such non-convex sparsity constraints are known as *iteratively re-weighted least squares* (IRLS) methods. Early work on IRLS methods utilized  $l_p$  norms with  $p > 1$  [138, 22], and these algorithms have recently been

---

<sup>0</sup>This work has appeared in [8].

extended to non-convex optimization frameworks with  $0 \leq p \leq 1$  in [196, 60, 67]. A similar re-weighting approach is utilized for  $l_1$ -norms in [43].

In this chapter, we propose a novel Bayesian framework for compressive sensing recovery using non-convex  $l_p$ -norms. By utilizing a majorization-minimization approach, we demonstrate that Bayesian inference can be performed without resorting to sampling approaches. Specifically, we employ a variational Bayesian analysis for inference which provides distribution estimates of the unknowns, and therefore allows the calculation of the uncertainties of the estimates. The proposed algorithm resulting from this framework simultaneously estimates the unknown signal  $\mathbf{x}$  along with all needed algorithm parameters, so that no user-intervention is needed and the parameters are chosen optimally. We show that existing IRLS methods are special cases of the proposed method. Finally, we demonstrate with experimental results that the proposed algorithm provides high recovery performance compared to some of the most commonly used recovery algorithms in compressive sensing.

The rest of this chapter is organized as follows: We present the proposed hierarchical Bayesian model in Section 9.2 and provide the details of the observation model and the sparsity prior. The inference procedure utilizing a majorization-minimization approach and variational Bayesian analysis is presented in Section 9.3. We demonstrate the performance of the proposed method with experimental results in Sec. 9.4. Finally, conclusions are drawn in Section 9.5.

## 9.2. Bayesian modeling

We utilize a hierarchical Bayesian framework to model the components of the compressive acquisition system in (8.2). The Bayesian modeling of the CS reconstruction problem requires the definition of the joint distribution  $p(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y})$  of all unknown and observed quantities. In

this work we use the following factorization of the joint distribution

$$(9.3) \quad p(\mathbf{x}, \alpha, \beta, \mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \beta)p(\mathbf{x}|\alpha)p(\alpha)p(\beta).$$

In the first stage of the hierarchical model, the observation noise is modeled using the *conditional* distribution  $p(\mathbf{y}|\mathbf{x}, \beta)$  and the unknown signal  $\mathbf{x}$  is modeled by a sparsity prior  $p(\mathbf{x}|\alpha)$ . These distributions depend on model parameters, also called *hyperparameters*,  $\beta$  and  $\alpha$ , whose meanings will be made clear in the following sections. To model these parameters, we assign *hyperpriors*  $p(\alpha)$  and  $p(\beta)$  on them in the second stage.

In the following subsections, we provide specific forms of the distributions utilized in this work.

### 9.2.1. Observation Model

We assume that the observation noise is independent, Gaussian, zero-mean and with variance equal to  $\beta^{-1}$ . The distribution  $p(\mathbf{y}|\mathbf{x}, \beta)$  can then be expressed from (8.2) as

$$(9.4) \quad p(\mathbf{y}|\mathbf{x}, \beta) \propto \beta^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2\right].$$

### 9.2.2. Signal Model

The sparsity of the signal is modeled by the following signal prior

$$(9.5) \quad p(\mathbf{x}|\alpha) \propto \frac{1}{Z_x(\alpha)} \exp\left[-\alpha \sum_i |x_i|^p\right],$$

with  $Z_x(\alpha)$  the partition function which normalizes the distribution. This prior is also called the Generalized Gaussian prior. Note that a *maximum a posteriori* (MAP) formulation with the distributions in (9.4) and (9.5) results in the same inverse problem shown in (9.2), using  $\tau = \frac{\alpha}{\beta}$ .

The partition function  $Z_x(\alpha)$  can be calculated using

$$\int_0^\infty \exp[-\alpha u^p] du = \frac{1}{p} \int_0^\infty \exp[-\alpha v] v^{\frac{1-p}{p}} dv \propto \alpha^{-\frac{1}{p}},$$

with  $u^p = v$ , which results in  $Z_x(\alpha) = c \alpha^{-\frac{N}{p}}$ , with  $c$  a constant. The final form of the sparsity prior is therefore given by

$$(9.6) \quad p(\mathbf{x}|\alpha) = c \alpha^{\frac{N}{p}} \exp \left[ -\alpha \sum_i |x_i|^p \right].$$

Note that by using the sparsity prior in (9.6), the signal coefficients are modeled by a probability distribution with a single hyperparameter  $\alpha$ . On the other hand, existing Bayesian methods generally employ independent distributions on each signal coefficient [247, 116, 115, 14], where each distribution is modeled using separate hyperparameters. As will be shown in Sec. 9.3, although the proposed formulation uses a single hyperparameter for all signal coefficients, it introduces an additional variable which will separately enforce adaptivity for each coefficient.

### 9.2.3. Model for hyperparameters

As mentioned above, in the second stage of the hierarchical model, we model the hyperparameters  $\alpha$  and  $\beta$  by hyperprior distributions  $p(\alpha)$  and  $p(\beta)$ . In Bayesian models, hyperprior distributions are generally chosen to be conjugate distributions, i.e., they have the same form

as the product of the conditional distribution and the priors. This choice of hyperpriors simplifies the analytical derivation of the inference procedure. We utilize Gamma hyperpriors on both hyperparameters  $\alpha$  and  $\beta$ , since the Gamma distribution is the conjugate distribution for the precision of Gaussian distributions. Another advantage of the Gamma distribution is that it includes uniform distribution as a limiting case, which makes the estimation of the hyperparameters solely depend on the observations.

The distributions  $p(\alpha)$  and  $p(\beta)$  are therefore expressed as

$$(9.7) \quad p(\alpha) = \Gamma(\alpha|a_\alpha^0, b_\alpha^0) = \frac{(b_\alpha^0)^{a_\alpha^0}}{\Gamma(a_\alpha^0)} \alpha^{a_\alpha^0-1} \exp[-\alpha b_\alpha^0],$$

$$(9.8) \quad p(\beta) = \Gamma(\beta|a_\beta^0, b_\beta^0) = \frac{(b_\beta^0)^{a_\beta^0}}{\Gamma(a_\beta^0)} \beta^{a_\beta^0-1} \exp[-\beta b_\beta^0],$$

with  $a_\alpha^0, a_\beta^0$  the shape parameters and  $b_\alpha^0, b_\beta^0$  the scale parameters, respectively. The means and variances of  $\alpha$  and  $\beta$  are given respectively by

$$(9.9) \quad \text{Mean}[\alpha] = \langle \alpha \rangle = \frac{a_\alpha}{b_\alpha}, \quad \text{Var}[\alpha] = \frac{a_\alpha}{(b_\alpha)^2}.$$

$$(9.10) \quad \text{Mean}[\beta] = \langle \beta \rangle = \frac{a_\beta}{b_\beta}, \quad \text{Var}[\beta] = \frac{a_\beta}{(b_\beta)^2}.$$

In this work, we use small values (such as  $10^{-3}$ ) for the *a priori* shape and scale parameters  $a_\alpha^0, a_\beta^0, b_\alpha^0, b_\beta^0$  to make the reconstruction process rely more on the observations than prior knowledge. Note, however, that the parameters of the posterior distributions of  $\alpha$  and  $\beta$  will be estimated as explained in Sec. 9.3.

Combining the distributions at both stages of the hierarchical model defined in (9.4), (9.5) and (9.7)-(9.8), we obtain the joint distribution in (9.3). The dependencies within the proposed

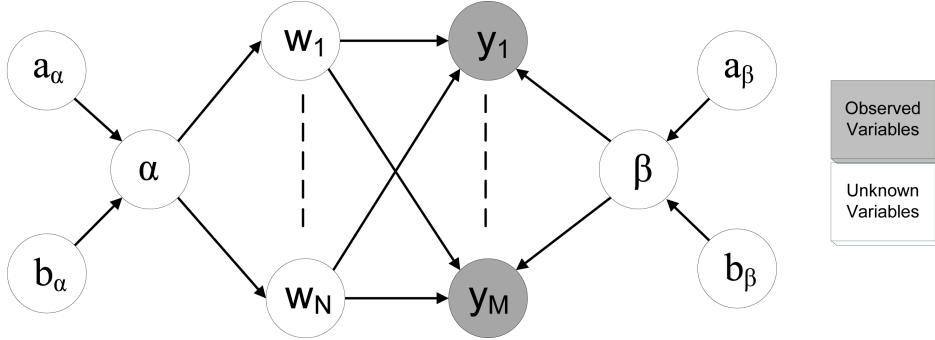


Figure 9.1. Graphical models representing the dependencies within the proposed Bayesian model.

model are shown graphically in Fig. (9.1), where the observed variables are presented by gray circles, and the unknown variables are shown with white circles. It should be noted that each observation coefficient  $y_i$  depends on all signal coefficients  $x_j$ ,  $j = 1, \dots, N$ .

### 9.3. Inference procedure

The Bayesian inference is based on the posterior distribution

$$(9.11) \quad p(\alpha, \beta, \mathbf{x} | \mathbf{y}) = \frac{p(\alpha, \beta, \mathbf{x}, \mathbf{y})}{p(\mathbf{y})},$$

which is analytically intractable, since  $p(\mathbf{y})$  cannot be computed. Therefore, approximation methods are utilized, including the evidence procedure (type-II maximum likelihood), maximum *a posteriori* solutions, and sampling methods [31]. In this work, we incorporate a variational Bayesian approach for the inference, which recently was found to be effective in many inference problems [31, 13, 162, 32]. In variational Bayesian approaches, the unknown posterior distribution  $p(\alpha, \beta, \mathbf{x} | \mathbf{y})$  is approximated by a simpler, analytically tractable distribution  $q(\alpha, \beta, \mathbf{x})$ . This approximating distribution is found by minimizing the Kullback-Leibler (KL)

distance between the posterior distribution and its approximation, that is,

$$\begin{aligned}
 C_{KL}(q(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y})) &= \int \int \int q(\alpha, \beta, \mathbf{x}) \log \left( \frac{q(\alpha, \beta, \mathbf{x})}{p(\alpha, \beta, \mathbf{x}|\mathbf{y})} \right) d\alpha d\beta d\mathbf{x} \\
 (9.12) \quad &= \int \int \int q(\alpha, \beta, \mathbf{x}) \log \left( \frac{q(\alpha, \beta, \mathbf{x})}{p(\alpha, \beta, \mathbf{x}, \mathbf{y})} \right) d\alpha d\beta d\mathbf{x} + \text{const.}
 \end{aligned}$$

The KL divergence is always non negative and equal to zero only when  $q(\alpha, \beta, \mathbf{x}) = p(\alpha, \beta, \mathbf{x}|\mathbf{y})$ .

Generally, the only assumption made in variational Bayesian analysis is that the distribution  $q(\alpha, \beta, \mathbf{x})$  can be factorized. In this work we use the following factorization

$$(9.13) \quad q(\alpha, \beta, \mathbf{x}) = q(\alpha)q(\beta)q(\mathbf{x})$$

Unfortunately, the form of the sparsity prior in (9.6) does not allow for the direct application of the variational Bayesian analysis, since the KL distance in (9.12) cannot be calculated. Therefore, we resort to a majorization-minimization approach, where the goal is to find a bound of the non-convex prior in (9.6) which can be utilized for further Bayesian analysis. To find such a bound, let us consider the weighted arithmetic and geometric mean inequality given by

$$(9.14) \quad a^{\frac{p}{2}} b^{1-\frac{p}{2}} \leq \frac{p}{2}a + (1 - \frac{p}{2})b,$$

with  $0 < p \leq 2$ , and nonnegative numbers  $a$  and  $b$ . Assuming  $b > 0$  and dividing both sides by  $b^{1-\frac{p}{2}}$  we obtain

$$(9.15) \quad a^{\frac{p}{2}} \leq \frac{p}{2} \frac{a + \frac{2-p}{p}b}{b^{1-p/2}}.$$

Next, let us define the following functional

$$(9.16) \quad \mathbf{M}(\alpha, \mathbf{x}, \mathbf{v}) = c \alpha^{\frac{N}{p}} \exp \left[ -\frac{\alpha p}{2} \sum_i \left( \frac{(x_i)^2 + \frac{2-p}{p} v_i}{(v_i)^{1-p/2}} \right) \right],$$

where  $\mathbf{v} \in (R^+)^N$  is a vector with components  $v_i$ , and  $c$  is the constant same as in (9.6). Using the inequality in (9.15) with  $a = (x_i)^2$  and  $b = v_i$  in  $\mathbf{M}(\alpha, \mathbf{x}, \mathbf{v})$ , and comparing with  $p(\mathbf{x}|\alpha)$  in (9.6), it is clear that  $\mathbf{M}(\alpha, \mathbf{x}, \mathbf{v})$  is a lower bound of the prior  $p(\mathbf{x}|\alpha)$ , that is,

$$(9.17) \quad p(\mathbf{x}|\alpha) \geq \mathbf{M}(\alpha, \mathbf{x}, \mathbf{v}).$$

Since the bounding functional  $\mathbf{M}(\alpha, \mathbf{x}, \mathbf{v})$  has a quadratic form, Bayesian inference can analytically be carried out by majorizing the prior  $p(\mathbf{x}|\alpha)$  by this quadratic functional. Using the lower bound in (9.17), a lower bound of the joint probability distribution in (9.3) can be found, that is

$$(9.18) \quad \begin{aligned} p(\alpha, \beta, \mathbf{x}, \mathbf{y}) &\geq p(\alpha)p(\beta)\mathbf{M}(\alpha, \mathbf{x}, \mathbf{v})p(\mathbf{y}|\mathbf{x}, \beta) \\ &= \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v}), \end{aligned}$$

which leads to the following upper bound for the KL divergence in (9.12)

$$(9.19) \quad C_{KL}(q(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y})) \leq C_{KL}(q(\alpha, \beta, \mathbf{x}) \| \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})) + \text{const.}$$

Note that this upper bound can be made tighter by minimizing it with respect to  $\mathbf{v}$ , since

$$(9.20) \quad C_{KL}(q(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y})) \leq \min_{\mathbf{v}} C_{KL}(q(\alpha, \beta, \mathbf{x}) \| \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})) + \text{const.}$$

Thus, minimizing the upper bound  $C_{KL}(q(\alpha, \beta, \mathbf{x}) \| \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v}))$  with respect to both  $q(\alpha, \beta, \mathbf{x})$  and  $\mathbf{v}$  will result in an ever decreasing sequence of upper bounds. Therefore,  $C_{KL}(q(\alpha, \beta, \mathbf{x}) \| p(\alpha, \beta, \mathbf{x}|\mathbf{y}))$  can be approximated by its upper bound, and this approximation can be made tighter by minimizing it iteratively with respect to both  $q(\alpha, \beta, \mathbf{x})$  and  $\mathbf{v}$ . Note that tightening the upper bound of the KL divergence also results in closer approximations of the signal prior  $p(\mathbf{x}|\alpha)$  by the bounding functional  $\mathbf{M}(\alpha, \mathbf{x}, \mathbf{v})$ .

Based on the above, we replace the minimization of the KL divergence in (9.12) by its upper bound given in (9.19), and therefore obtain a majorization-minimization procedure for the variational Bayesian inference. Note, however, that (9.19) cannot be analytically minimized with respect to all  $q(\cdot)$  distributions and the vector  $\mathbf{v}$  at the same time, and an alternating minimization procedure has to be employed as follows. Let us denote by  $\Theta = \{\mathbf{x}, \alpha, \beta\}$  the set of all unknowns, and by  $\Theta_\theta$  the set  $\Theta$  with  $\theta$  removed. Then, for each unknown  $\theta \in \Theta$ , the posterior  $q(\theta)$  can be computed by holding  $q(\Theta_\theta)$  constant and solving

$$(9.21) \quad q(\theta) = \operatorname{argmin}_{q(\theta)} C_{KL}(q(\Theta_\theta)q(\theta) \| \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})).$$

The standard solution of variational Bayesian analysis [162, 31] can then be used for (9.21), which results in

$$(9.22) \quad q(\theta) = \text{const} \times \exp \left( \mathbf{E}_{q(\Theta_\theta)} [\log \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})] \right),$$

where

$$\mathbf{E}_{q(\Theta_\theta)} [\log \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})] = \int \log \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v}) q(\Theta_\theta) d\Theta_\theta.$$

Applying this general solution to each unknown in an alternating fashion results in an iterative procedure, which converges to the best approximation of the true posterior distribution  $p(\mathbf{x}, \alpha, \beta | \mathbf{y})$  by distributions of the form  $q(\alpha, \beta, \mathbf{x}) = q(\alpha)q(\beta)q(\mathbf{x})$  when the majorization-minimization approach is used.

Note that majorization-minimization approaches are also utilized in image restoration methods based on wavelets [82] and total-variation priors [13]. Non-quadratic bounds can also be found for  $0 < p < 1$  [82], but they cannot be directly utilized within a Bayesian framework.

We next proceed to give the explicit forms of each  $q(\cdot)$  distribution. In what follows, the means of the distributions will be denoted by  $\langle \cdot \rangle = \mathbf{E}_{q(\theta)}(\cdot)$ , when the corresponding distribution is clear from the context.

The distribution  $q(\mathbf{x})$  is calculated from (9.22) as

$$(9.23) \quad q(\mathbf{x}) \propto \exp \left( \mathbf{E}_{q(\alpha, \beta)} [\log \mathbf{F}(\alpha, \beta, \mathbf{x}, \mathbf{v})] \right),$$

which corresponds to an  $N$ -dimensional multivariate Gaussian distribution  $\mathcal{N}(\mathbf{x} | \langle \mathbf{x} \rangle, \Sigma_{\mathbf{x}})$ .

The mean and covariance of this distribution are given by

$$(9.24) \quad \langle \mathbf{x} \rangle = \Sigma_{\mathbf{x}} \langle \beta \rangle \Phi^t \mathbf{y},$$

$$(9.25) \quad \Sigma_{\mathbf{x}} = (\langle \beta \rangle \Phi^t \Phi + p \langle \alpha \rangle \mathbf{W})^{-1}$$

where

$$(9.26) \quad \mathbf{W} = \text{diag} \left( v_i^{p/2-1} \right), i = 1, \dots, N.$$

The components  $v_i$  of the vector  $\mathbf{v}$  can be calculated using

$$v_i = \underset{v_i}{\operatorname{argmin}} \frac{\langle x_i^2 \rangle + \frac{2-p}{p} v_i}{v_i^{1-p/2}},$$

which results in the following update

$$(9.27) \quad v_i = \langle x_i^2 \rangle, \quad i = 1, \dots, N.$$

It is clear now that the matrix  $\mathbf{W}$  in (9.26) is a weighting matrix, which in combination with  $x_i^2$  provides an estimate of  $\| \mathbf{x} \|_p^p$ . A similar weighting matrix is utilized in IRLS algorithms [60]. Note, however, that in IRLS algorithms the elements of  $\mathbf{W}$  are chosen as  $(\langle x_i \rangle^2)^{p/2-1}$ , whereas in this work they are set equal to  $(\langle x_i^2 \rangle)^{p/2-1}$ , which is calculated from

$$(9.28) \quad \begin{aligned} \langle x_i^2 \rangle &= (\mathbf{E}_{q(\mathbf{x})}[x_i])^2 + \mathbf{E}_{q(\mathbf{x})}[(x_i - \mathbf{E}_{q(\mathbf{x})}[x_i])^2] \\ &= \langle x_i \rangle^2 + (\Sigma_{\mathbf{x}})_{ii}, \end{aligned}$$

where  $(\Sigma_{\mathbf{x}})_{ii}$  denotes the  $i^{\text{th}}$  diagonal element of the matrix  $\Sigma_{\mathbf{x}}$ , and it is the variance of the coefficient  $x_i$ . The first term is equivalent to the one used in IRLS algorithms, and the second term (the variance) incorporates the uncertainty of the estimate  $\mathbf{x}$  in the reweighting procedure. It will be shown in the experimental results section that utilizing this information results in significant improvement in the reconstruction performance compared to the IRLS methods. Additionally, the estimated variances can be utilized for designing adaptive measurement systems as in [116, 209].

Finally, we utilize (9.22) to calculate the hyperprior distributions. The distributions  $q(\alpha)$  and  $q(\beta)$  are found to be Gamma distributions, expressed as

$$(9.29) \quad q(\alpha) \propto \alpha^{N/p+a_\alpha^0-1} \exp \left[ -\alpha \left( \sum_i v_i^{p/2} + b_\alpha^0 \right) \right],$$

$$(9.30) \quad q(\beta) \propto \beta^{N/2+a_\beta^0-1} \exp \left[ -\beta \left( \frac{\mathbf{E}_{q(\mathbf{x})}(\|\mathbf{y} - \Phi\mathbf{x}\|_2^2)}{2} + b_\beta^0 \right) \right].$$

The means of these distributions are given by

$$(9.31) \quad \langle \alpha \rangle = \mathbf{E}_{q(\alpha)}[\alpha] = \frac{N/p + a_\alpha^0}{\sum_i v_i^{p/2} + b_\alpha^0},$$

and

$$(9.32) \quad \langle \beta \rangle = \mathbf{E}_{q(\beta)}[\beta] = \frac{N/2 + a_\beta^0}{\mathbf{E}_{q(\mathbf{x})}(\|\mathbf{y} - \Phi\mathbf{x}\|_2^2)/2 + b_\beta^0},$$

The denominator in (9.32) is calculated using

$$(9.33) \quad \mathbf{E}_{q(\mathbf{x})}[\|\mathbf{y} - \Phi\mathbf{x}\|_2^2] = \|\mathbf{y} - \Phi\langle \mathbf{x} \rangle\|_2^2 + \text{trace}(\Sigma_{\mathbf{x}}\Phi^t\Phi).$$

In summary, the algorithm iterates between (9.24), (9.26), (9.31) and (9.32) until convergence. Note that the estimate  $\langle \mathbf{x} \rangle$  in (9.24) can be calculated by standard methods for solving linear systems, e.g., Gaussian elimination. On the other hand, explicit calculation of the matrix  $\Sigma_{\mathbf{x}}$  is needed in (9.28) and (9.33), which is computationally very intense, since  $\Sigma_{\mathbf{x}}$  is of size  $N \times N$ . Therefore, we first calculate the incomplete Cholesky factorization  $\Sigma_{\mathbf{x}}^{-1} \approx \mathbf{L}\mathbf{L}^T$  and approximate  $\Sigma_{\mathbf{x}}$  by  $(\mathbf{L}\mathbf{L}^T)^{-1}$ . This greatly improves the computational efficiency and decreases the numerical errors resulting from the direct calculation of  $\Sigma_{\mathbf{x}}$ .

To conclude this section, we investigate the special case of noiseless compressive sensing measurements, that is,  $\mathbf{y} = \Phi\mathbf{x}$ . In this case, the problem in (9.2) becomes

$$(9.34) \quad \hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{x}\|_p^p, \text{ subject to } \mathbf{y} = \Phi\mathbf{x}.$$

It can be seen from (9.24) and (9.25) that when  $\beta \rightarrow \infty$ , the estimate of  $\mathbf{x}$  is given by

$$(9.35) \quad \langle \mathbf{x} \rangle = \mathbf{W}^{-1} \Phi^t (\Phi \mathbf{W}^{-1} \Phi^t)^{-1} \mathbf{y}.$$

Let us further assume that the distribution  $q(\mathbf{x})$  is a degenerate distribution, that is, a distribution which takes the value  $\langle \mathbf{x} \rangle$  with probability one and the rest with probability zero. Then,  $\langle x_i^2 \rangle = \langle x_i \rangle^2$ , and therefore

$$(9.36) \quad \mathbf{W} = \operatorname{diag}(|\langle x_i \rangle|^{p-2}).$$

The estimate in (9.35) combined with (9.36) coincides with the IRLS algorithms [60] proposed to solve the problem (9.34). Clearly, the proposed method is a generalization of the IRLS method.

#### 9.4. Experiments

In this section we present numerical results with both one-dimensional signals and 2D images. We considered experimental setups commonly used in the literature, and compared the proposed algorithm with some of the state-of-the-art algorithms for compressive sensing recovery.

#### 9.4.1. 1D Synthetic Signals

We generate sparse vectors  $\mathbf{x}$  of size  $N = 256$  with 20 nonzero coefficients, which are drawn from a zero-mean Gaussian distribution of variance 1. The  $M \times 256$  measurement matrices  $\Phi$  are also generated from a zero-mean Gaussian distribution with variance 1, and their columns are scaled to have unit 2-norms. Other choices of both the signal and measurement matrix give similar results and therefore are not reported here.

For all the experiments reported here, we run the reconstruction algorithms by varying the number of measurements  $M$  from 60 to 120 in steps of 10. Each experiment is repeated 100 times and the average of these results is reported. Moreover, we study both noiseless and noisy observations, where in the latter case we added zero-mean white Gaussian noise with standard deviation 0.03 to the measurements  $\Phi\mathbf{x}$ . The least-squares solution  $(\Phi^t\Phi)^{-1}\Phi\mathbf{y}$  is used as the initial estimate of  $\mathbf{x}$  in the proposed algorithm, and the iterations are terminated when the change in the estimate from the previous iteration, calculated using the  $l_2$ -norm, is less than  $10^{-6}$ . Finally, the reconstruction error is calculated as  $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2$ , where  $\hat{\mathbf{x}}$  and  $\mathbf{x}$  are the estimated and true coefficient vectors, respectively.

We first study the effect of the variable  $p$  on the reconstruction performance. The proposed algorithm, which will be denoted by BCS-lp in the following, is run with different values within the interval  $0 < p \leq 1$ . Figure 9.2 shows the reconstruction errors for different values of  $p$  and while varying the number of measurements  $M$  both for noiseless and noisy observations. The errors are color-coded with lighter values corresponding to higher error values (with a maximum of 0.7), and darker values corresponding to lower error values (with a minimum of 0).

It is clear that values of  $p$  close to zero result in lower reconstruction errors for a given number of measurements. Moreover, it can be observed that perfect reconstruction is achieved

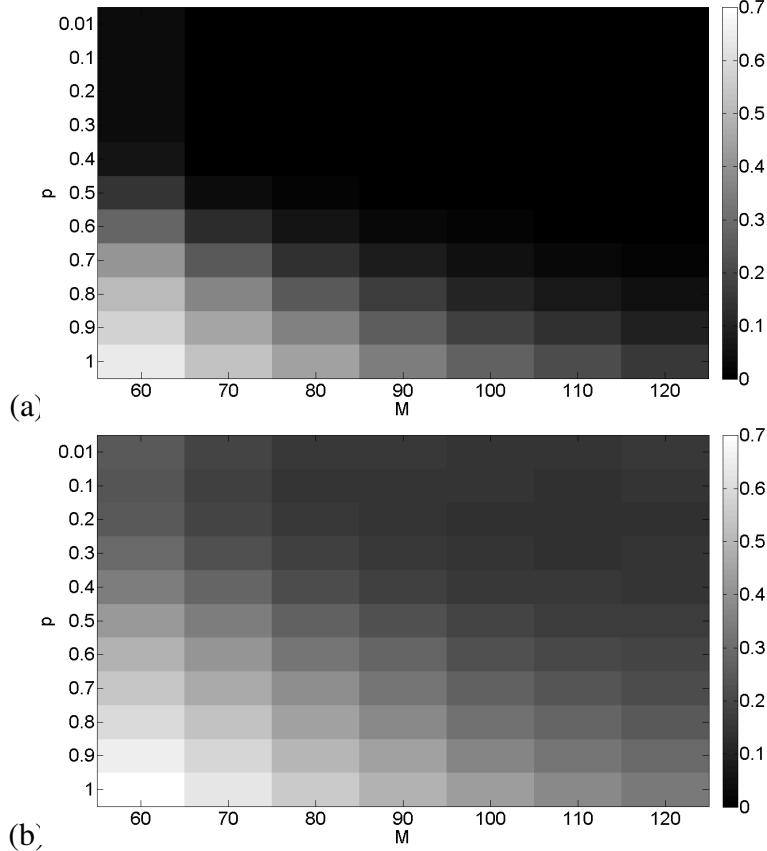


Figure 9.2. Reconstruction errors obtained by the proposed method for different values of  $p$  while varying the number of measurements  $M$ ; (a) when no observation noise is present, and (b) when Gaussian noise with standard deviation 0.03 is added to the measurements.

with a lower number of measurements when smaller  $p$  values are used. The smallest value of  $p$  we utilized is 0.01, which resulted in the best overall performance. However, values smaller and slightly higher than 0.01 provided similar results.

Figure 9.3 shows error rate comparisons between six selected  $p$ -values for both noiseless and noisy observation cases. It is clear that smaller values of  $p$  result in lower reconstruction errors for both noiseless and noisy observations. Note also that the performance increase while

decreasing  $p$  is diminishing, i.e., values close to  $p = 0.01$  results in similar reconstruction performance.

Next we compare the proposed method BCS-lp with the IRLS algorithm as proposed in [60]. We ran IRLS on the same dataset as BCS-lp using  $p$ -values within the interval  $0 \leq p \leq 1$ . The comparison for a selected set of  $p$  values is shown in Fig. 9.4. We run the proposed method BCS-lp both on the noisy and noiseless datasets with the same default setting, that is, as if observation noise is present. On the other hand, IRLS is originally designed for the reconstruction from noiseless observations.

It can be observed from Fig. 9.4 that in the case of noisy measurements, BCS-lp provides lower reconstruction errors for all values of  $p$  independent of the number of measurements. In fact, BCS-lp provides higher reconstruction accuracy even when a higher  $p$ -value is used than the  $p$ -value for IRLS. On the other hand, even though it is assumed in BCS-lp that the measurements are noisy, the reconstruction performance of BCS-lp is higher for  $M \leq 70$  and negligibly smaller for  $M > 70$  in the case of noiseless observations shown in Fig. 9.4(a). It is clear that overall, BCS-lp provides smaller reconstruction errors.

Next we compare the proposed algorithm with a selection of existing CS reconstruction algorithms from the literature, namely, the algorithms BCS [116], BP [64], BCS-Laplace (presented in Chapter 8) [9, 14], the GPSR method [83], and iterative hard thresholding (IHT) [35]. For all algorithms, their MATLAB implementations in the corresponding websites are used, and the required algorithm parameters are set according to their default setups. The algorithms BCS and BCS-Laplace are greedy constructive algorithms, whereas BP and GPSR are global optimization methods based on  $l_1$ -norms. We used  $p = 0$  for the IRLS method, and  $p = 0.01$

for BCLS-lp, as these provided the lowest reconstruction errors. The re-weighted  $l_1$  algorithm in [43] is not included in the comparison as it is expected to perform similarly to IRLS (see [60]).

The results are shown for both noiseless and noisy measurements in Fig. 9.5. It can be observed from Fig. 9.5(a) that in the case of noiseless measurements, BCS-lp outperforms other methods in terms of reconstruction error and it achieves perfect reconstruction with fewer number of measurements than other algorithms. On the other hand, when observation noise is present, BCS-lp provides the smallest reconstruction error when the number of measurements is very small, but its performance does not improve as much as with some other algorithms (such as IHT). From the experimental results, we stipulate that this is mostly due to numerical errors arising when solving the linear system in (9.24). Note that a similar behavior can also be observed with IRLS. The reason that the method GPSR provided poor reconstruction results is most likely due to the default parameter settings, and its performance is expected to increase with optimally selecting its parameters. On the other hand, note that the proposed algorithm does not require parameter-tuning, as all algorithm parameters are estimated simultaneously with the signal.

An interesting observation from Fig. 9.5(a) is that the reconstruction performance is improved as more heavy-tailed distributions are utilized as sparsity priors. Independent Gaussian-priors are utilized in BCS on each signal coefficient, whereas Laplace priors are used in BCS-Laplace. BCS-Laplace achieves lower reconstruction errors than BCS, and both IRLS and BCS-lp outperform these algorithms by utilizing  $l_p$  norms with  $p < 1$ .

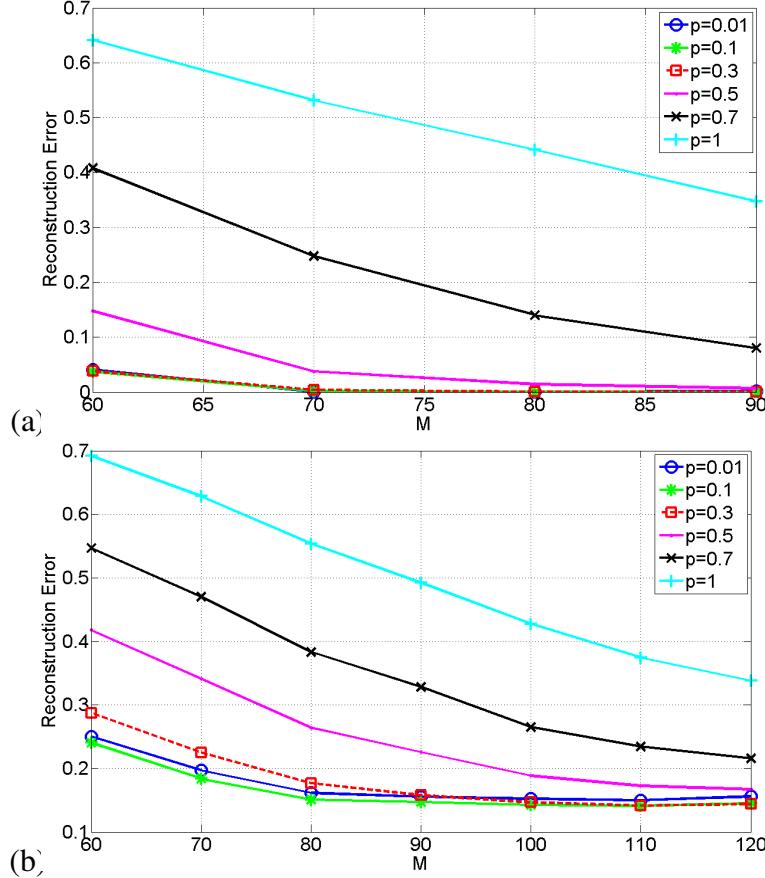


Figure 9.3. Reconstruction errors obtained by the proposed method with varying the number of measurements  $M$  for  $p$ -values 0.01, 0.1, 0.3, 0.05, 0.7 and 1. The measurements are noiseless in (a) and Gaussian noise with standard deviation 0.03 is added to the measurements.

#### 9.4.2. 2D Images

In this section we present reconstruction experiments for a widely used compressive image sensing setup, namely the multiscale CS reconstruction of the *Mondrian* image. In the experiments, incoherent measurements of the wavelet transform of the image are acquired at each wavelet scale, and a *symmlet8* wavelet with the coarsest scale 4 and finest scale 6 is utilized. We adapted the same parameters as in the SparseLab package [1], where the number of samples

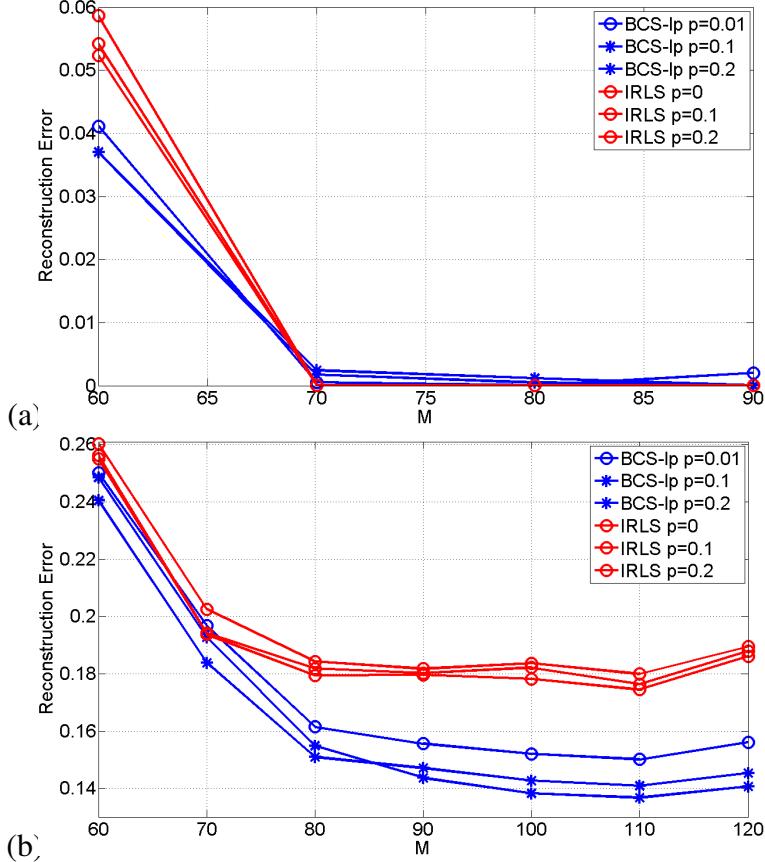


Figure 9.4. Comparison between the proposed method BCS-lp and IRLS algorithm for different  $p$  values when the observations are (a) noiseless and (b) degraded with Gaussian noise with standard deviation 0.03.

is  $N = 4096$ , the number of measurements is  $M = 2073$  and the measurement matrix is drawn from an uniform spherical distribution.

We compared the performance of the proposed method BCS-lp with the performance of the algorithms BP, GPSR, BCS and BCS-Laplace, as in the previous section. The IRLS method failed to provide reasonable reconstruction results since the increased number of signal coefficients caused instabilities in the Gaussian elimination process (see equation (9.35) in Sec. 9.3), and its running time increased beyond practical usage range.

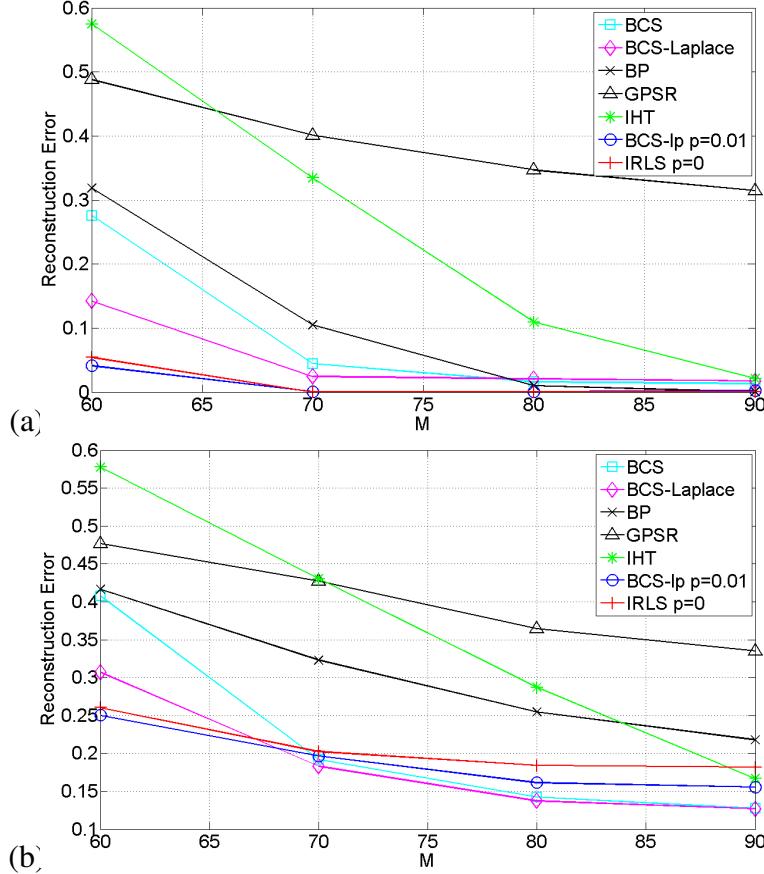


Figure 9.5. Comparison between a number of CS reconstruction algorithms with varying number of measurements  $M$  with (a) noiseless and (b) noisy measurements.

The experiments are repeated 50 times due to the randomness of the measurement matrices, and the average reconstruction errors of the algorithms are shown in Table 9.1. Note that BCS-lp provides the smallest reconstruction error among all algorithms. The high running time of the BCS-lp method is due to the difficulty in solving the linear system in (9.24) when the number of coefficients is high. This problem is expected to be overcome with the use of greedy approaches and therefore eliminating the need to solve the system in (9.24). Examples of reconstructed

images are shown in Fig. 9.6 where it can be observed that all methods provide fairly good reconstructions.

An interesting observation from Table 9.1 is that the proposed method provides along with BP the least-sparse solution among all methods. Initially, this might seem to contradict the expectation that the priors based on  $l_p$ -norms should provide sparser solutions than convex priors. However, note that when the original signal coefficients are nonzero, the reconstruction algorithm is not expected to provide zero estimates. Therefore, providing sparse solutions is an advantage when the underlying signal is actually sparse. Although most of the wavelet coefficients of the *Mondrian* image are very small, they are not exactly zero, and the reconstruction performance is increased if they can also be recovered.

To check the relevance of the small coefficients in the estimates of BCS-lp to the original signal, we keep 849 largest coefficients in the estimates provided by all algorithms and set the rest of them to zero. We calculate the reconstruction error with these signal estimates, and the results are shown in Table 9.2. It can be observed that although the reconstruction error provided by BCS-lp significantly increases, it is still the lowest among all algorithms. Moreover, the smaller coefficients in the estimates are clearly relevant to the original signal, and BCS-lp provides a non-sparse estimate because the original wavelet coefficients are not exactly sparse.

In summary, experimental results with both 1D synthetic signals and 2D images show that the proposed method provides high performance in reconstruction quality compared to existing algorithms.

Table 9.1. Average reconstruction errors, running times and number of nonzero components for multi-scale CS reconstruction of the *Mondrian* image.

	Mondrian		
	# Nonzeros	Time	Error
BP	4096	338.03	0.1400
GPSR	849	3.45	0.1802
BCS	1210	19.57	0.1432
BCS-Laplace	1128	26.00	0.1436
BCS-lp	4096	18974.49	0.1390

Table 9.2. Average reconstruction errors when only the 849 largest signal coefficients from the estimates in Table 9.1 are kept.

	Mondrian	
	# Nonzeros	Error
BP	849	0.1432
GPSR	849	0.1802
BCS	849	0.1445
BCS-Laplace	849	0.1452
BCS-lp	849	0.1426

## 9.5. Conclusions

In this chapter we developed a Bayesian framework utilizing non-convex sparsity priors for compressive sensing reconstruction. We proposed a majorization-minimization approach which allows using non-convex priors based on  $l_p$ -norms within a Bayesian formulation. By employing a variational Bayesian analysis, the reconstruction algorithm developed from this framework simultaneously estimates all unknowns and provides distribution estimates, which take the estimation uncertainties into account and can be used to ensure the accuracy of the estimation process. We have shown that the proposed formulation is a generalized version of some existing methods, such as reweighted least squares and sparse Bayesian methods, and therefore it can provide potential directions for improvement. Experimental results demonstrate

that using non-convex priors results in reconstructions with higher accuracy, and the unknown signal can be recovered with fewer measurements compared to methods using convex-priors. Finally, we have shown that the performance of the proposed algorithm is competitive compared to state-of-the-art methods.

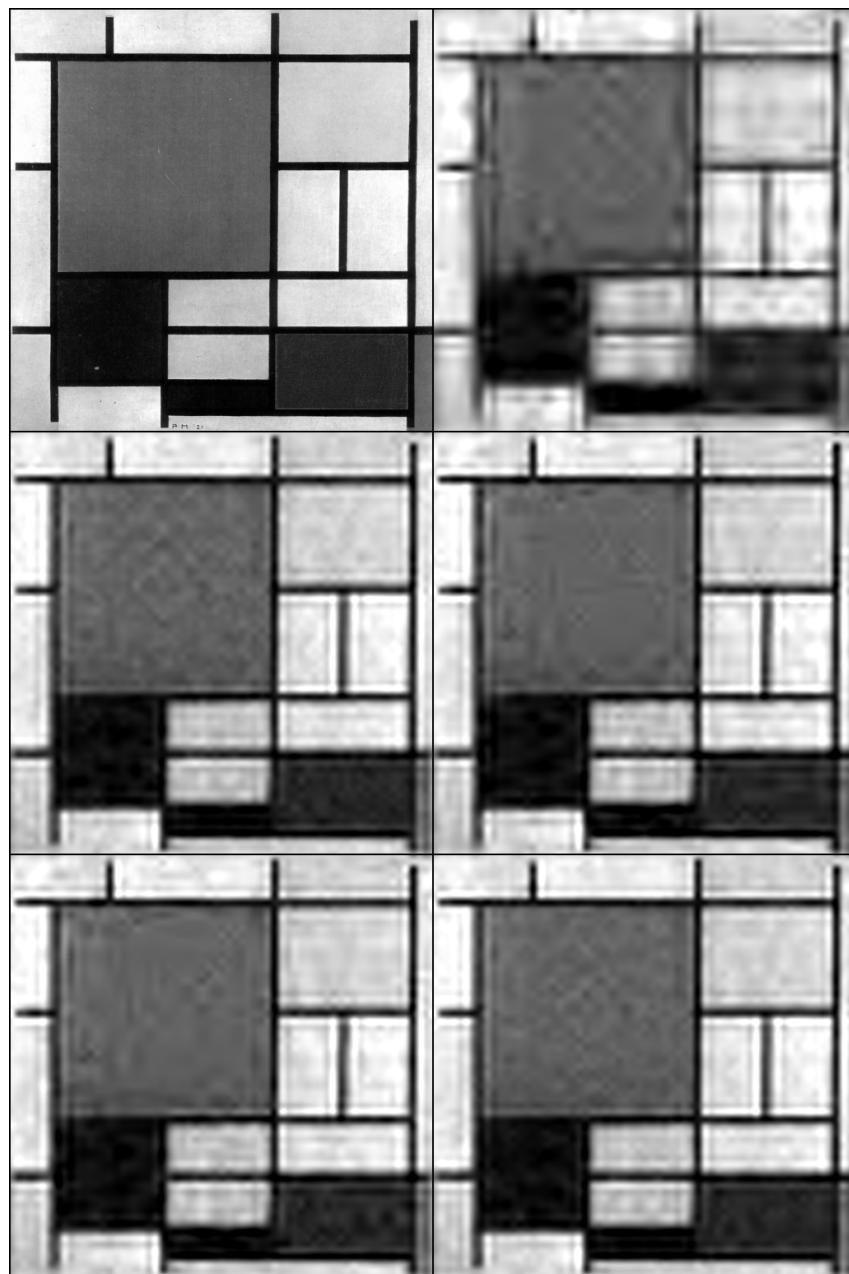


Figure 9.6. Examples of reconstructed Mondrian images using a multi-scale compressive sensing scheme. (Top-left) Original image; Reconstructed images by (Top-right) GPSR; (Middle-left) BP; (Middle-right) BCS; (Bottom-left) BCS-Laplace; (Bottom-right) BCS-lp.

## CHAPTER 10

**Compressive Light Field Imaging****10.1. Introduction**

This chapter presents an interesting application of compressive sensing, namely, a novel camera design for light field image acquisition. Before providing details about our proposed light field acquisition method, we will first briefly review the area of light field imaging and introduce some necessary concepts.

Recent advances in computational photography provided effective solutions to a number of photographic problems, and also resulted in novel methods to acquire and process images. Novel camera designs allow capturing information of the scene which is not possible to obtain using traditional cameras. This information can then be utilized for example to generate the three-dimensional scene geometry, or for novel post-processing methods, such as digital refocusing or synthetic aperture.

Light-field cameras are one of the most widely used class of computational cameras. The light-field expresses the radiance density function on the camera sensor, or the light energy of all rays in 3D space passing through the camera. For instance, a four-dimensional (4D) discrete light field image  $\mathbf{x}(i, k, m, n)$  with spatial dimensions  $i, k$  and angular dimensions  $m, n$  contains images of a scene from a variety of angles, which provide information about the 3D structure of the scene. Each 2D image  $\mathbf{x}(i, k, m_0, n_0)$  with fixed angular coordinates  $m_0, n_0$  is called an *angular image*. Traditional cameras integrate these angular images (or equivalently, light

rays) over their 2D aperture to obtain the image, which results in the loss of valuable depth information about the scene. On the other hand, light field cameras capture the angular data and provide means to work directly on the light-rays instead of pixels, allowing one to produce many views of the scene, or perform many photographic tasks after the acquisition is made. This provides a clear advantage for light-field imaging over traditional photography and makes many novel applications possible.

Compressive sensing (CS) [41, 72], on the other hand, has recently become very popular due to its interesting theoretical nature and wide area of applications. The theory of compressive sensing dictates that a signal can be recovered very accurately from a much smaller number of measurements than required by traditional methods, provided that it is *compressible* (or *sparse*) in some basis, i.e., only a few basis coefficients contain the major part of the signal energy. Besides sparsity, compressive sensing makes use of incoherent measurements and nonlinear reconstruction, and has led to many interesting theoretical results and novel applications (see, for instance, [45, 20]).

In this chapter, we present a novel application of compressive sensing, namely, a novel framework to acquire light-field images. We show that light field acquisition can be formulated using incoherent measurement principles. We then demonstrate that light-field images have a highly sparse nature, which, in combination with incoherent measurements, can be exploited to reconstruct the light-field images with much fewer image acquisitions than traditionally required. By exploiting this sparsity in light field images, we develop a novel reconstruction algorithm that recovers the original images from few compressive measurements with a very high degree of accuracy.

In addition, we propose a novel camera design based on the developed acquisition framework. We build our design on ideas from coded aperture imaging, computational photography and compressive sensing. By exploiting the fact that different regions of the aperture of a camera correspond to images of the scene from different angles, we incorporate a compressively coded mask placed at the aperture to obtain incoherent measurements of the incident light-field. These measurements are then decoded using the proposed reconstruction algorithm to recover the original light-field image. We exploit the highly sparse nature of the light-field images to obtain accurate reconstructions with only a few measurements compared to the high angular dimension of the light-field image. The proposed camera design provides images with high signal-to-noise ratio and does not suffer from resolution trade-off inherent in most existing light-field camera designs. Finally, we demonstrate the efficiency of the proposed framework with experimental results.

This chapter is organized as follows: First we review related prior work in light field and coded aperture imaging in Sec. 10.2. In Sec. 10.3 we present the proposed acquisition method to obtain incoherent measurements of the light field image. We model the acquisition system and the light field images using a Bayesian framework, which is described in Sec. 10.4. The Bayesian inference procedure used to develop the reconstruction algorithm is presented Sec. 10.5. We demonstrate the efficiency of the proposed system with experimental results in Sec. 10.6. Finally, conclusions are drawn in Sec. 10.7.

## 10.2. Prior Work

### 10.2.1. Light Field Acquisition

Light field acquisition, based on the principles of integral photography, was first proposed a long time ago [112, 145]. The same ideas appeared in computer vision literature first as the *plenoptic camera* [5], and then the potential of light field imaging is demonstrated in [142] and [99]. The original design in [5] is implemented in a hand-held camera in [179], where a microlens (lenticular) array is placed between the main lens and the camera sensor. A similar approach is proposed in [95], where instead of using microlenses, a lens array is placed in front of the camera main lens. In both approaches, the light field image is captured using one acquisition. The additional lens array is used to capture the angular information, and reordering the captured image results in images of different views of the scene. Other proposed light field camera designs include multi-camera systems [243] and mask-based designs [237, 94], which encode the angular information using frequency-multiplexing.

Many of these designs suffer from the spatio-angular tradeoff [95], that is, one cannot obtain light-field images with both high spatial- and high angular resolution. This problem is inherent in designs with one recording sensor (or film) and only one acquisition is made. If the captured light-field image has an angular resolution of  $N_h \times N_v$ , and a spatial resolution of  $P_h \times P_v$ , then  $N_h \times N_v \times P_h \times P_v$  can only be less than or equal to the resolution of the camera sensor. For instance, a typical light field image captured using the plenoptic camera in [179] provides 14x14 angular images of size app. 300x300 in a 16 megapixel camera. Multi-camera systems [243] are not affected from the spatio-angular tradeoff, but they are very costly to implement and cumbersome for practical usage.

Recently, a *programmable aperture camera* is proposed [143], where a binary mask is used to code the aperture. Angular images are *multiplexed* into single 2D images similar to the principle of coded aperture imaging. After multiple acquisitions are made, a linear estimation procedure is utilized to recover the full light-field image. Although this design captures images with both high spatial and angular resolution, the number of acquisitions is equal to the number of angular dimensions. Therefore, obtaining a light-field image with a high angular resolution is not practical.

### 10.2.2. Coded Aperture Imaging

Coded aperture imaging is developed in order to collect more light in situations where a lens system cannot be used, due to the measured wavelengths. Imaging systems with coded apertures are currently widely used in astronomy and medicine. The technique is based on the principle of pinhole cameras, but instead of utilizing only one pinhole which suffers from low SNR, a specially designed array of pinholes is used. This array of pinholes provide images that are overlapping copies of the original scene, which can then be decoded using computational algorithms to provide a sharp image. There is a vast literature on coded aperture methods in astronomy and medicine (see, for example, [80, 100]).

Recent works considered coded aperture methods for developing novel image acquisition methods. In [197], the aperture is coded in time-domain to modify the exposure for motion deblurring. Spatially modifying the aperture has been utilized for a range of applications: Levin *et. al.* [141] proposed utilizing an aperture mask to reconstruct both the original image and the depth of the scene from a single snapshot. A lensless imaging system is proposed in [256] that allows manipulating the captured scene in ways not possible by traditional cameras, such

as splitting field of view. Nayar *et. al.* [174] utilized a spatial light modulator to control the exposure per pixel, which can be used to obtain high-dynamic range images. Other uses of coded apertures include super resolution [165] and range estimation [77].

Compressive sensing methods have also been applied in conjunction with coded apertures or compressively coded blocking masks. Novel imaging methods have been proposed for spectral imaging [89], dual-photography [213], and the design of structured light for recovering inhomogeneous participating media [101]. Most recently, compressively coded aperture masks are utilized for single-image super-resolution and shown to provide higher quality images than traditional coded apertures [156, 157].

A related approach to coded aperture imaging is *wavefront coding* [49], where the image is intentionally defocused using phase plates so that the defocus is uniform throughout the image. The captured image can then be deconvolved to obtain an image with an enlarged depth of field.

### 10.3. Compressive Sensing of Light-Fields

In this section, we will show that light-field image acquisition can be formulated within a compressive sensing framework. We first show that light-field images can be acquired by coding the camera aperture, and then present the proposed compressive acquisition system. In the following, a 4D light field image is denoted by  $\mathbf{x}$ , which is the collection of  $N$  angular images  $\mathbf{x}^j$ , such that  $\mathbf{x} = \{\mathbf{x}^j\}, j = 1, \dots, N$ .

#### 10.3.1. Light-Field Acquisition by Coded Apertures

A fundamental principle utilized in this work is that different regions of the aperture capture images of the scene from different angles. In other words, the main camera lens can be interpreted

as an array of multiple virtual lenses (or cameras). This concept is illustrated in Fig. 10.1(a)-(c), where only certain parts of the aperture are left open to acquire images exhibiting vertical and horizontal parallax. By separately opening one region of the aperture and blocking light in the others, the complete light-field with an angular dimension of  $N$  can be captured with  $N$  exposures. However, obtaining the light-field image in this fashion has two disadvantages: First, due to the very small amount of light arriving to the sensor at each exposure, the captured angular images have very low signal-to-noise ratios (SNR). Second, a high number of acquisitions has to be made in order to obtain high angular resolution. The programmable aperture camera design in [143] addressed the first problem by incorporating a multiplexing scheme, but the second problem remains as a serious drawback.

We address both of these issues by utilizing a randomly coded non-refractive mask in front of the aperture. Each image acquired in this fashion is a random linear combination (and therefore an incoherent measurement) of the angular images. An example image captured in this fashion is illustrated in Fig. 10.1(d), where the amount of light passing through regions of the aperture are randomly selected (shown at the bottom of Fig. 10.1(d)). As shown in the following, utilizing such a random mask overcomes both of the problems described above.

The mathematical principle behind this idea is formulated as follows. Let us assume that the aperture of the main camera lens is divided into  $N$  blocks, with  $N = N_h \times N_v$  where  $N_h$  and  $N_v$  represent the number of horizontal and vertical divisions. During each acquisition  $i$ , each block  $j$  is assigned a weight  $0 \leq a^{ij} \leq 1$  which controls the amount of light passing through this block. Therefore,  $a^{ij}$  represents the transmittance of the block  $j$ , i.e., it is the fraction of incident light that passes through the block. As mentioned above, each block  $j$  captures an angular image  $\mathbf{x}^j$  in the light-field image, and therefore the acquired image  $\mathbf{y}^i$  at the  $i^{\text{th}}$  acquisition can be

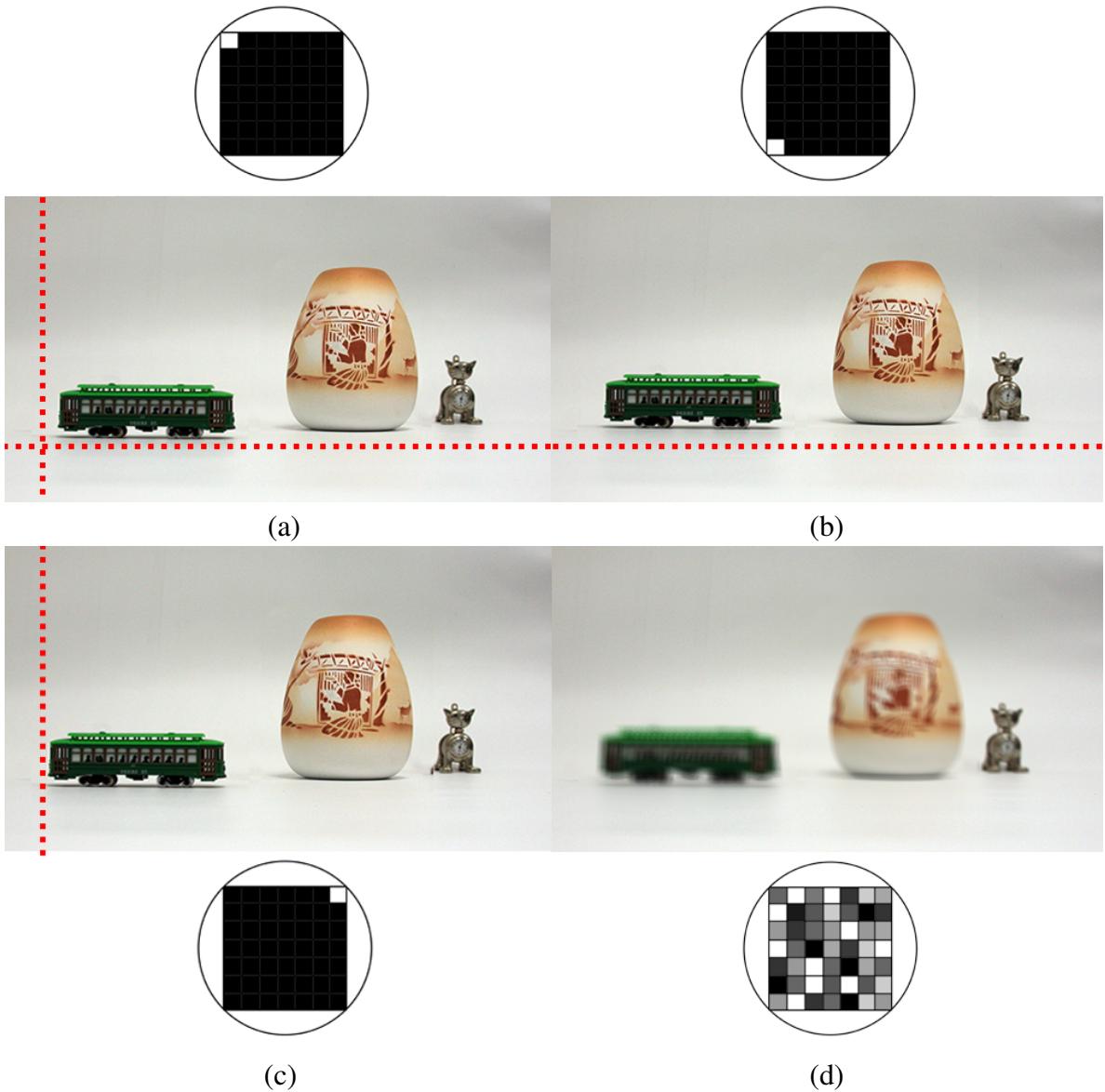


Figure 10.1. The basic principle of utilizing a coded aperture to obtain light field images. The angular images are shown in (a), (b) and (c) when only corner blocks of the aperture are left open. Both horizontal and vertical parallax can be observed between these images (Horizontal and vertical dashed lines are shown to clearly denote the vertical and horizontal parallax, respectively). Figure (d) shows a captured image with the randomly coded aperture used in the proposed compressive sensing light field camera. All images are from a synthetic light field image (see Sec. 10.6).

represented as a linear combination of the  $N$  angular images as

$$(10.1) \quad \mathbf{y}^i = \sum_{j=1}^N a^{ij} \mathbf{x}^j, \quad i = 1, \dots, M.$$

where we use the vector notation such that  $\mathbf{y}^i$  and  $\mathbf{x}^j$  are both  $P \times 1$  vectors, with  $P$  the number of pixels in each image. Note that in a traditional camera, the acquired image is the average of all angular images, i.e.,  $a^{ij} = \frac{1}{N}$ , since the aperture integrates all light rays coming from different directions.

After  $M$  acquisitions ( $M \leq N$ ), the complete set of observed images  $\{\mathbf{y}^i\}$  can be expressed in matrix-vector form as

$$(10.2) \quad \begin{pmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \\ \vdots \\ \mathbf{y}^M \end{pmatrix} = \begin{pmatrix} a^{11}\mathbf{I} & a^{12}\mathbf{I} & \dots & a^{1N}\mathbf{I} \\ a^{21}\mathbf{I} & a^{22}\mathbf{I} & \dots & a^{2N}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ a^{M1}\mathbf{I} & a^{M2}\mathbf{I} & \dots & a^{MN}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^N \end{pmatrix},$$

with  $\mathbf{I}$  the  $P \times P$  identity matrix. (10.2) is expressed in a more compact form as

$$(10.3) \quad \mathbf{y} = \mathbf{Ax}$$

with

$$(10.4) \quad \mathbf{A} = \begin{pmatrix} a^{11} & a^{12} & \dots & a^{1N} \\ a^{21} & a^{22} & \dots & a^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ a^{M1} & a^{M2} & \dots & a^{MN} \end{pmatrix} \otimes \mathbf{I} = \hat{\mathbf{A}} \otimes \mathbf{I},$$

where  $\otimes$  is the Kronecker product. Taking also the acquisition noise into account, the final observation model can be expressed as

$$(10.5) \quad \mathbf{y} = (\hat{\mathbf{A}} \otimes \mathbf{I}) \mathbf{x} + \mathbf{n} = \mathbf{Ax} + \mathbf{n}$$

with  $\mathbf{n}$  the  $PM \times 1$  noise vector.

### 10.3.2. Compressively Coded Apertures

If the linear measurement matrix  $\mathbf{A}$  satisfies certain properties dictated by the theory of compressive sensing [41], the light field acquisition system in (10.5) can be seen as a noisy incoherent measurement system. A sufficient condition for a matrix to be compressive sensing matrix is the *restricted isometry property* (RIP) [42, 41], which is proven to hold with a very high probability for a general class of matrices with their entries drawn from certain random probability distributions. For instance, if  $\hat{\mathbf{A}}$  in (10.5) is constructed by independently drawing its entries from a Gaussian distribution, then  $\hat{\mathbf{A}}$  satisfies RIP with an overwhelming probability.

It is straightforward to show that if  $\hat{\mathbf{A}}$  is a valid compressive sensing matrix, then  $\mathbf{A}$  is a valid compressive sensing matrix as well. A simple proof is as follows. The matrix  $\hat{\mathbf{A}}$  has the

restricted isometry property for  $k$  if the eigenvalues of  $\hat{\mathbf{A}}^T \hat{\mathbf{A}}$  lie between  $1 - \delta_k$  and  $1 + \delta_k$  where  $0 < \delta_k < 1$  [42]. Since  $\mathbf{A}^T \mathbf{A} = \hat{\mathbf{A}}^T \hat{\mathbf{A}} \otimes \mathbf{I}$ , from the properties of the Kronecker product,  $\mathbf{A}^T \mathbf{A}$  has the same eigenvalues as  $\hat{\mathbf{A}}^T \hat{\mathbf{A}}$ . Therefore,  $\mathbf{A}$  also has the restricted isometry property and is therefore a valid compressive sensing matrix.

Based on this, the acquisition system in (10.5) is an incoherent measurement system of angular images  $\mathbf{x}^j$ , where each acquired image is a random linear combination of the angular images. The theory of compressive sensing then dictates that if the unknown image  $\mathbf{x}$  can be represented sparsely in some transform domain, then it can be recovered with much fewer measurements than traditionally required ( $M \ll N$ ). Due to the nature of multi-view images and especially in the specific case considered in this work where the angular images are aligned in a small-baseline, the redundancy within the light field images is very high. In fact, there are multiple sources of sparsity inherent in light field images, arising both from within angular images and the high correlations between them (see Sec. 10.4.2 for details). Therefore, light field images can be very accurately reconstructed with very few acquisitions by utilizing the compressive acquisition system in (10.5) and by exploiting their sparse nature within nonlinear reconstruction frameworks.

An important design issue is the selection of the measurement matrix  $\mathbf{A}$ , which determines the level of incoherence of measurements and therefore the reconstruction performance. The design of the measurement matrices for compressive sensing is an active area of research, and many of the existing designs can be utilized for the proposed aperture mask. In this work, we specifically experimented with two different types of measurement matrices, namely, Gaussian ensembles and scrambled Hadamard ensembles [88]. If fractional values of the block transmittances are permitted, a very general class of matrices can be utilized, with positivity of matrix

entries as the only constraint. In this case, Gaussian ensembles are very suitable as measurement matrices. If the mask is limited to binary codes, scrambled Hadamard ensembles can be used to code the aperture. Moreover, the measurement matrices can also be selected depending on specific requirements of the optical systems, e.g., the expected amount of passing light can be varied by varying the mean value of the corresponding probability distribution, or choosing specific construction of the random measurement matrix.

It should be noted that since many (or possibly all) blocks are open in each exposure, each captured image has a high SNR due to the small amount of loss of light. In fact, the measurement matrices can be designed to optimize the amount of passing light while maintaining the random structure. Moreover, as shown in the experimental results section, incorporating a non-linear reconstruction mechanism provides images with much higher SNRs than those of linear reconstruction methods such as demultiplexing.

Finally, it should be noted that the coded aperture setup utilized in this work is a specific application for the acquisition system in (10.5). The proposed compressive sensing formulation for light field acquisition can be applied to a wider range of light-field imaging applications. For instance, multiple camera or multiple lens imaging systems such as camera arrays and stereo cameras can equally well incorporate the incoherent measurement system in (10.5) and significantly reduce the number of acquisitions without sacrificing spatial or angular resolution.

#### **10.4. Hierarchical Bayesian Model For Reconstruction**

In order to be able to reconstruct the angular images  $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N$  from the incoherent measurements  $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^M$  and  $\mathbf{A}$ , both the observation process (10.5) and the unknown light field image  $\mathbf{x}$  have to be modeled. We utilize an hierarchical Bayesian framework by employing

a conditional distribution  $p(\mathbf{y}|\mathbf{x}, \beta)$  for the observation model in (10.5) and a *prior* distribution  $p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)$  on the unknown light field image  $\mathbf{x}$ . These distributions depend on the model parameters  $\beta$ ,  $\boldsymbol{\alpha}_{\text{TV}}$  and  $\boldsymbol{\alpha}_c$ , which are called *hyperparameters*, and in the second stage of the hierarchical model we utilize additional prior distributions, called *hyperpriors*, to model them. This hierarchical Bayesian model and carrying out a fully-Bayesian inference procedure leads to a fully-automated reconstruction algorithm.

In this work, we utilize the following factorization of the joint distribution  $p(\mathbf{y}, \mathbf{x}, \beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)$  of all unknown and observed quantities

$$(10.6) \quad p(\mathbf{y}, \mathbf{x}, \beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c) = p(\mathbf{y}|\mathbf{x}, \beta) p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c) p(\beta) p(\boldsymbol{\alpha}_{\text{TV}}) p(\boldsymbol{\alpha}_c)$$

In the following subsections, we present the specific models utilized for each of these distributions.

#### 10.4.1. Observation (Noise) Model

The observation noise is assumed to be independent and Gaussian with zero mean and variance equal to  $\beta^{-1}$ , that is, using (10.5),

$$(10.7) \quad p(\mathbf{y}|\mathbf{x}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{Ax}, \beta^{-1}).$$

#### 10.4.2. Light-Field Image Model

The choice of randomly programmed coded apertures makes the exact/approximate recovery of angular images possible through the use of sparsity inherent in light field images. There are multiple sources of sparsity within light field images that can be exploited. The first one

is sparsity within each angular image. It is already well known that 2D images can be very accurately represented by only a few coefficients of a *sparsifying* transform, such as wavelet transforms or total-variation (TV) functions on the image. In the case of light-field images, there is another fundamental source of sparsity, that is, the angular images are very closely related to each other. Specifically, each angular image can be accurately estimated from another one using dense warping (or correspondence) fields, as shown below.

Based on the above, we utilize the following factorization of the prior distribution

$$(10.8) \quad p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c) = C(\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}})p(\mathbf{x}|\boldsymbol{\alpha}_c)$$

where  $p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}})$  is the TV image prior employed on each angular image separately,  $p(\mathbf{x}|\boldsymbol{\alpha}_c)$  is the prior modeling the sparsity arising from the strong dependency between angular images and  $C(\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)$  is a function of the unknown hyperparameters needed for the image prior model to integrate to one.

Next we describe the specific models utilized for each of the prior distributions in this factorization.

**10.4.2.1. Total Variation Image Prior.** The angular images  $\mathbf{x}^i$  are natural images, therefore they are expected to be mostly smooth except at a number of discontinuities (e.g., at edges). As the spatial domain image priors, we utilize the total variation function because of its edge-preserving property by not over-penalizing discontinuities in the image while imposing smoothness [204]. Specifically,  $p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}})$  can be expressed as

$$(10.9) \quad p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}}) \propto \prod_{i=1}^N (\alpha_{\text{TV}}^i)^{P/2} \exp \left[ -\frac{1}{2} \alpha_{\text{TV}}^i \text{TV}(\mathbf{x}^i) \right],$$

where

$$(10.10) \quad \text{TV}(\mathbf{x}^i) = \sum_k \sqrt{(\Delta_k^h(\mathbf{x}^i))^2 + (\Delta_k^v(\mathbf{x}^i))^2},$$

where  $\Delta_k^h$  and  $\Delta_k^v$  correspond to, respectively, horizontal and vertical first order differences, at pixel  $k$ , that is,  $\Delta_k^h(\mathbf{x}^i) = (\mathbf{x}^i)_k - (\mathbf{x}^i)_{l(k)}$  and  $\Delta_k^v(\mathbf{x}^i) = (\mathbf{x}^i)_k - (\mathbf{x}^i)_{a(k)}$ , where  $l(k)$  and  $a(k)$  denote the nearest neighbors of pixel  $k$ , to the left and above, respectively.

**10.4.2.2. Cross-image prior.** As mentioned above, there is a high correlation between the angular images in the light field image. Specifically, disregarding the occlusions, each angular image  $\mathbf{x}^i$  can be very closely approximated from another angular image  $\mathbf{x}^j$  by the dense warping field  $\mathbf{M}^{ij}$ , that is  $\mathbf{x}^i \approx \mathbf{M}^{ij}\mathbf{x}^j$ . Therefore, the dependency of each angular image on another is very strong and can be exploited while modeling  $\mathbf{x}$ . Based on this, we utilize the following cross-image prior between the angular images

$$(10.11) \quad p(\mathbf{x}|\boldsymbol{\alpha}_c) \propto \exp\left(\sum_{i=1}^N \sum_{j \in \mathcal{N}(i)} -\frac{\alpha_c^{ij}}{2} \|\mathbf{x}^i - \mathbf{M}^{ij}\mathbf{x}^j\|^2\right),$$

where  $\alpha_c^{ij}$  is the precision of the registration error, and  $\mathcal{N}(i)$  defines a neighborhood of  $\mathbf{x}^i$ , which consists of the angular images with closest viewpoints to that of  $\mathbf{x}^i$ . In other words, angular images captured by nearby regions in the aperture are treated as neighboring images. This neighborhood is imposed in (10.11) for several reasons. First, angular images far apart in the aperture can be less accurately related by dense warping fields due to the 3D structure of the scene and increased size of the occluded areas. Second, incorporating a cross-image prior between each pair of angular images in  $\mathbf{x}$  largely increases memory requirements and therefore computationally not efficient during the reconstruction phase. Finally, since  $\mathbf{x}^i$  is part of at

least one neighborhood defined on  $\mathbf{x}$ , the warping constraint is propagated to all angular images during the reconstruction algorithm.

The cross-image prior in (10.11) can be written in matrix-vector form as

$$(10.12) \quad p(\mathbf{x}|\boldsymbol{\alpha}_c) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T \Pi \mathbf{x}\right),$$

where the matrix  $\Pi$  is a sparse  $NP \times NP$  matrix constructed by  $N \times N$  blocks of size  $P \times P$ .

Specifically, it is given by

$$(10.13) \quad \Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} & \dots & \Pi_{1N} \\ \Pi_{21} & \Pi_{22} & \dots & \Pi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_{N1} & \Pi_{N2} & \dots & \Pi_{NN} \end{pmatrix}$$

The  $P \times P$  blocks  $\Pi_{ij}$  can be found from (10.11) as

$$\Pi_{ij} = \begin{cases} \sum_{s \in \mathcal{N}(i)} \alpha_c^{is} \mathbf{I} + \alpha_c^{si} (\mathbf{M}^{si})^T \mathbf{M}^{si} & \text{if } i = j \\ -\alpha_c^{ij} \mathbf{M}^{ij} - \alpha_c^{ji} (\mathbf{M}^{ji})^T & \text{if } j \neq i, j \in \mathcal{N}(i) \\ 0 & \text{else} \end{cases}$$

The form of the matrix  $\Pi$  makes the calculation of the partition function of the distribution in (10.12) intractable. To overcome this difficulty, we approximate the partition function by a

quadratic form, and utilize the following improper prior as the cross-image prior

$$(10.14) \quad p(\mathbf{x}|\boldsymbol{\alpha}_c) = c \left[ \prod_{i,j} (\alpha_c^{ij})^{P/2} \right] \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Pi} \mathbf{x}\right),$$

It is clear that incorporating the cross-image prior requires knowledge of the dense warping fields  $\mathbf{M}^{ij}$ , which cannot be directly obtained from the compressive measurements. In this work, we overcome this problem by acquiring two additional images from two opposite sides of the aperture. These images exhibit full horizontal and vertical parallax, and a dense registration algorithm based on graph-cuts is utilized to obtain the warping field [38]. Due to the uniform partitioning of the aperture, this warping field can be used to obtain approximate intermediate warping fields between all angular images. The disadvantage of this approach is that two additional exposures have to be taken with small apertures (i.e., with a low SNR), and combined with the approximate calculation of the intermediate warping fields, the constraints imposed in the cross-image prior might not fully characterize the actual relations within the light field image. However, our experiments have shown that this approach provide accurate enough warping fields so that no major difficulties arise in the reconstruction phase. Moreover, estimating the precision variables  $\alpha_c^{ij}$  along with the image compensates for the inaccuracies in the warping fields during the reconstruction.

On the other hand, note that the observation model in (10.5) can be seen as a convolution of the light field image with a filter constructed by the transmittances in the aperture in the four-dimensional space. Utilizing this fact, one can apply deconvolution of  $\mathbf{y}$  in the 4D space by this filter, and obtain approximations to the angular images. These images can then be used to estimate the warping fields  $\mathbf{M}^{ij}$ . This is in principle similar to the idea of segmenting depth layers by identifying the filter scales in [141].

Another alternative method is to use  $\mathbf{x}^i \approx \mathbf{x}^j$ , which is similar to the approximation utilized in compressive video sensing algorithm in [155]. Although this method does not require knowledge about the warping fields, it is a very crude approximation and therefore does not provide reconstruction results comparable to the ones reported here. However, it can be utilized with relatively high performance in the case of very densely packed angular images, since the variation between two neighboring angular images will be very small.

#### 10.4.3. Hyperpriors on the Hyperparameters

The form of the hyperprior distributions on the hyperparameters  $\beta$ ,  $\boldsymbol{\alpha}_{\text{TV}}$  and  $\boldsymbol{\alpha}_c$  determines the ease of calculation of the posterior distribution  $p(\mathbf{y}, \mathbf{x}, \beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)$ . A desired property for each hyperprior is to be conjugate [23], that is, to have the same functional form with the product  $p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c)p(\mathbf{y}|\mathbf{x}, \beta)$ . Since the distributions  $p(\mathbf{y}|\mathbf{x}, \beta)$  and  $p(\mathbf{x}|\boldsymbol{\alpha}_c)$  are Gaussian distributions, and we will approximate the distribution  $p(\mathbf{x}|\boldsymbol{\alpha}_{\text{TV}})$  by a Gaussian distribution (shown in Section 10.5), we chose to utilize Gamma distributions for all hyperparameters, as it is the conjugate prior for the inverse variance (precision) of the Gaussian distribution.

The Gamma distribution is defined by

$$(10.15) \quad p(\omega) = \Gamma(a_\omega^o, b_\omega^o) = \frac{(b_\omega^o)^{-a_\omega^o}}{\Gamma(a_\omega^o)} \omega^{a_\omega^o - 1} \exp\left[-\frac{\omega}{b_\omega^o}\right],$$

where  $\omega > 0$  denotes a hyperparameter,  $b_\omega^o > 0$  is the scale parameter, and  $a_\omega^o > 0$  is the shape parameter, both of which are assumed to be known and introduce our prior knowledge on the

hyperparameters. We utilize identical *a priori* shape and scale parameters ( $a^o$  and  $b^o$ , respectively) for each hyperparameter, so using (10.15), the hyperpriors are defined by

$$(10.16) \quad p(\beta) = \Gamma(\beta | a^o, b^o)$$

$$(10.17) \quad p(\alpha_{\text{TV}}^i) = \Gamma(\alpha_{\text{TV}}^i | a^o, b^o), \quad i = 1, \dots, N$$

$$(10.18) \quad p(\alpha_c^{ij}) = \Gamma(\alpha_c^{ij} | a^o, b^o), \quad i = 1, \dots, N, \quad j \in \mathcal{N}(i)$$

We use small values for the parameters  $a^o$  and  $b^o$  (e.g.,  $10^{-3}$ ) to make the *a priori* hyperpriors vague, which makes the estimation process depend more on observations than prior knowledge. Note, however, that if some prior knowledge about the hyperparameters is available (for example, the noise variances in the observations), this knowledge can easily be incorporated into the estimation procedure using appropriate values of the shape and scale parameters (see, for example, [13]).

## 10.5. Reconstruction Algorithm

Let us denote the set of the hyperparameters by  $\Omega = \{\beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c\}$  and the set of unknowns by  $\Theta = \{\Omega, \mathbf{x}\} = \{\beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c, \mathbf{x}\}$ . The Bayesian inference is based on the posterior distribution

$$(10.19) \quad p(\Theta | \mathbf{y}) = p(\beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c, \mathbf{x} | \mathbf{y}) = \frac{p(\beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c, \mathbf{x}, \mathbf{y})}{p(\mathbf{y})},$$

with  $p(\beta, \boldsymbol{\alpha}_{\text{TV}}, \boldsymbol{\alpha}_c, \mathbf{x}, \mathbf{y})$  is given by (10.6). Unfortunately, the posterior  $p(\Theta | \mathbf{y})$  is intractable (since  $p(\mathbf{y})$  is intractable), and therefore approximations are utilized. A common approximation is approximating the posterior by a delta function at its mode. Then, using  $p(\mathbf{x} | \mathbf{y}, \Omega) \propto p(\Theta, \mathbf{y})$ ,

the unknowns can be found by

$$\Theta = \operatorname{argmax}_{\Theta} p(\Theta | \mathbf{y}) = \operatorname{argmax}_{\Theta} p(\Theta, \mathbf{y})$$

Note that this formulation results in the well-known maximum *a posteriori* estimates of  $\Theta$ . Specifically, assuming uniform hyperpriors on the hyperparameters, the estimates found by this inference procedure are equivalent to the solution of the following regularized inverse problem:

$$(10.20) \quad \Theta = \operatorname{argmin}_{\Theta} \beta \| \mathbf{y} - \mathbf{Ax} \|^2 + \sum_{i=1}^N \alpha_{\text{TV}}^i \text{TV}(\mathbf{x}^i) + \sum_{i=1}^N \sum_{j \in \mathcal{N}(i)} \alpha_c^{ij} \| \mathbf{x}^i - \mathbf{M}^{ij} \mathbf{x}^j \|^2$$

Therefore, existing methods for TV-regularized optimization can also be employed for solving the recovery problem (see, for example, [53, 151]).

However, even with this approximation, the distribution  $p(\Theta, \mathbf{y})$  is hard to calculate due to the use of the TV prior  $p(\mathbf{x} | \boldsymbol{\alpha}_{\text{TV}})$ . This is especially evident when calculating the hyperparameter estimates. Therefore, we resort to the majorization-minimization method developed in Chapter 3. We provide here a brief outline of its application to solve (10.20). We first define the following functional using a vector  $\mathbf{w}_{\text{TV}}^i \in (R^+)^P$  of length  $P$  as

$$(10.21) \quad \mathbf{M}(\boldsymbol{\alpha}_{\text{im}}, \mathbf{x}, \mathbf{w})(\mathbf{x}^i, \alpha_{\text{TV}}^i, \mathbf{w}_{\text{TV}}^i) = c (\alpha_{\text{TV}}^i)^{P/2} \exp \left[ -\frac{\alpha_{\text{TV}}^i}{2} \sum_k \frac{(\Delta_k^h(\mathbf{x}^i))^2 + (\Delta_k^v(\mathbf{x}^i))^2 + (\mathbf{w}_{\text{TV}}^i)_k}{\sqrt{(\mathbf{w}_{\text{TV}}^i)_k}} \right].$$

Using this functional, a lower bound of the prior  $p(\mathbf{x}^i | \alpha_{\text{TV}}^i)$  can be defined as

$$(10.22) \quad p(\mathbf{x}^i | \alpha_{\text{TV}}^i) \geq \mathbf{M}(\boldsymbol{\alpha}_{\text{im}}, \mathbf{x}, \mathbf{w})(\mathbf{x}^i, \alpha_{\text{TV}}^i, \mathbf{w}_{\text{TV}}^i)$$

which leads to the following lower bound of the joint probability distribution  $p(\Theta | \mathbf{y})$

$$(10.23) \quad p(\Theta, \mathbf{y}) \geq p(\Omega) p(\mathbf{y} | \mathbf{x}, \beta) p(\mathbf{x} | \boldsymbol{\alpha}_c) \prod_{i=1}^N \mathbf{M}(\alpha_{\text{im}}, \mathbf{x}, \mathbf{w})(\mathbf{x}^i, \alpha_{\text{TV}}^i, \mathbf{w}_{\text{TV}}^i) \\ = \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{y}).$$

Therefore, instead of utilizing  $p(\Theta, \mathbf{y})$ , we utilize its lower bound  $\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{y})$  for inference. Specifically, since  $p(\mathbf{x} | \mathbf{y}, \Omega) \propto p(\Theta, \mathbf{y}) \geq \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{y})$ , the unknowns are found by solving the following optimization problem

$$(10.24) \quad \Theta = \underset{\Theta}{\operatorname{argmax}} \mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{y})$$

In other words, instead of finding the mode of the joint distribution  $p(\Theta, \mathbf{y})$ , we find the mode of the approximating distribution  $\mathbf{F}(\Theta, \mathbf{w}, \mathbf{y}_1, \mathbf{y}_2)(\Theta, \mathbf{y})$  which bounds  $p(\Theta, \mathbf{y})$  from below. As shown in [13] and [10], this bound is made tighter at each iteration by recalculating the parameters  $\alpha_{\text{TV}}^i$  and the vectors  $\mathbf{w}_{\text{TV}}^i$  (see [13] and [10] for details).

Next we proceed by giving the specific solutions for each of the unknown variables. The estimate for the light field image  $\hat{\mathbf{x}}$  can be calculated as

$$(10.25) \quad \hat{\mathbf{x}} = \Sigma_{\mathbf{x}} \boldsymbol{\beta} \mathbf{A}^T \mathbf{y}$$

$$(10.26) \quad \Sigma_{\mathbf{x}}^{-1} = \operatorname{diag} \left( \alpha_{\text{TV}}^i (\Delta^h)^t \mathbf{W}_{\text{TV}}^i (\Delta^h) + \alpha_{\text{TV}}^i (\Delta^v)^t \mathbf{W}_{\text{TV}}^i (\Delta^v) \right) + \Pi$$

where the first matrix term in (10.26) is a  $NP \times NP$  block diagonal matrix created by  $P \times P$  blocks  $\alpha_{\text{TV}}^i(\Delta^h)^t \mathbf{W}_{\text{TV}}^i(\Delta^h) + \alpha_{\text{TV}}^i(\Delta^v)^t \mathbf{W}_{\text{TV}}^i(\Delta^v)$ . The matrices  $\mathbf{W}_{\text{TV}}^i$  are calculated by

$$(10.27) \quad \mathbf{W}_{\text{TV}}^i = \text{diag} \left( \frac{1}{\sqrt{(\mathbf{w}_{\text{TV}}^i)_k}} \right), k = 1, \dots, P$$

where

$$(10.28) \quad (\mathbf{w}_{\text{TV}}^i)_k = (\Delta_k^h(\hat{\mathbf{x}}^i))^2 + (\Delta_k^v(\hat{\mathbf{x}}^i))^2.$$

It is clear that the vector  $\mathbf{w}_{\text{TV}}^i$  in (10.28) represents the local spatial activity in each angular image  $\mathbf{x}^i$  using its total variation. Consequently, the matrix  $\mathbf{W}_{\text{TV}}^i$  in (10.27) is the spatial adaptivity matrix which controls the trade-off between the smoothness of the solutions and their fidelity to the measurements.

Similar to the light field estimate, the hyperparameters can be found by solving (10.24) with respect to each hyperparameter by keeping the others constant. Proceeding in this fashion, the corresponding estimates can be given by

$$(10.29) \quad \beta = \frac{\frac{1}{2}NP + b^o}{\frac{1}{2} \| \mathbf{y} - \mathbf{Ax} \|^2 + a^o}$$

$$(10.30) \quad \alpha_{\text{TV}}^i = \frac{\frac{1}{2}P + b^o}{\sum_k (\mathbf{w}_{\text{TV}}^i)_k + a^o}$$

$$(10.31) \quad \alpha_c^{ij} = \frac{\frac{1}{2}P + b^o}{\frac{1}{2} \| \mathbf{x}^i - \mathbf{M}^{ij}\mathbf{x}^j \|^2 + a^o}$$

Finally, the algorithm iterates among estimating the light field image using (10.25), the spatial adaptivity vectors using (10.28), and the hyperparameters using (10.29)-(10.31) until convergence.

## 10.6. Experimental Results

We generated a synthetic light-field image shown in Fig. (10.1) with known warping fields. The light field image has a spatial resolution of  $250 \times 125$  and an angular resolution of  $7 \times 7$ . As the measurement matrix  $\mathbf{A}$  we chose the uniform spherical ensemble, that is, its entries are drawn from a uniform distribution and are between 0 and 1. Since the mean of this distribution is 0.5, using this measurement matrix, the expected amount of light passing through the aperture in each acquisition is half of the maximum possible. Finally, we add zero-mean Gaussian noise with variance 0.1 to the measurements to obtain the final observations.

We vary the number of acquired images  $M$  from 1 to 49 and apply the proposed reconstruction algorithm using the incoherent observations to obtain estimates of the original light-field image. Each experiment is repeated 50 times and the average is reported. The reconstruction error is calculated according to  $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2$ , where  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  are the original and estimated images, respectively.

Average reconstruction errors over 50 runs are shown in Fig. (10.2). It is clear that very accurate reconstructions can be obtained using very few measurements. The minimum reconstruction error that can be achieved is around  $0.5 \times 10^{-5}$  with 49 measurements, due to the presence of observation noise. The proposed algorithm provides average reconstruction errors of around  $1 \times 10^{-4}$  and  $1.5 \times 10^{-5}$  from 11 and 21 measurements, respectively. In fact, an average error of  $6 \times 10^{-4}$  is already obtained with only 7 measurements. Examples of reconstructed images using 11 and 21 measurements are shown in Fig. 10.3(b) and Fig. 10.3(c), respectively. Note that the reconstructed images are nearly indistinguishable from the original image, which is shown in Fig. 10.3(a). It can be observed that using the proposed design the number of acquisitions can be significantly reduced (by a factor between 1/7 to 1/4). Furthermore, the reduction

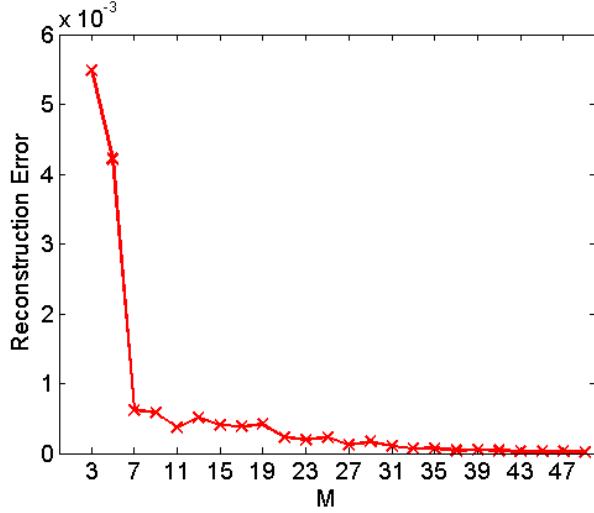


Figure 10.2. Number of measurements  $M$  vs relative reconstruction error (average over 50 runs).

in the number of acquisitions is expected to be much higher with larger light-field images, due to the increased level of sparsity.

### 10.7. Conclusions

In this chapter, we proposed a novel application of compressive sensing to a novel camera design to acquire 4D light-field images. We have shown that incoherent measurements of the angular images can be collected by using a randomly coded mask in front of the aperture of a traditional camera. These measurements are then used to reconstruct the original light field image. We developed a reconstruction algorithm which exploits the high degree of sparsity inherent in the light field images, and have shown that the complete light field image can be reconstructed using only a few acquisitions. Moreover, the captured images have high signal-to-noise ratios due to small amount of loss of light. The proposed design provides high spatial and angular resolution light field images, and does not suffer from limitations of many existing

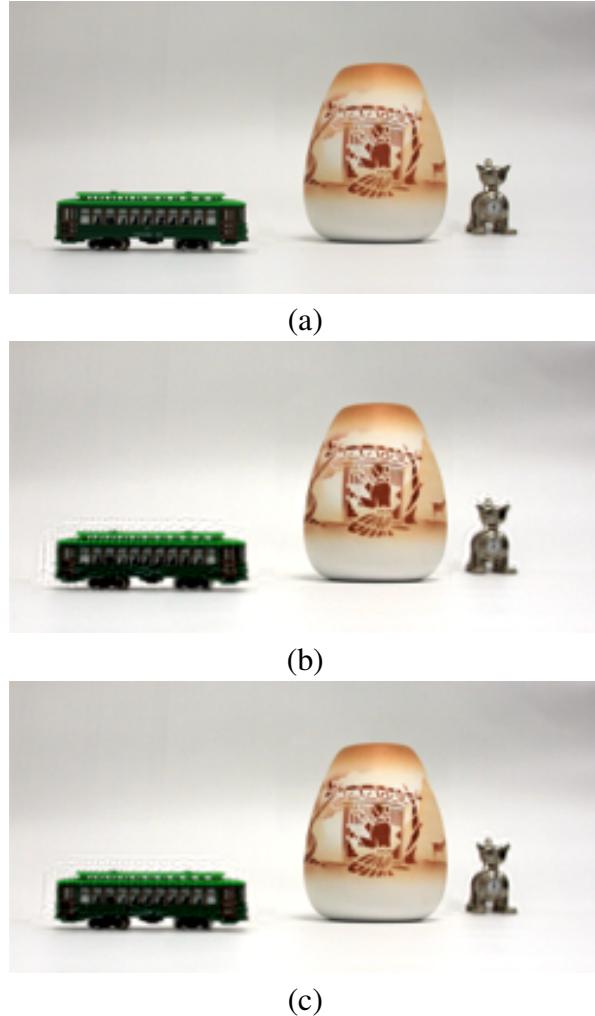


Figure 10.3. Reconstruction examples. (a) Original angular image, reconstructed images from (b) 11 measurements (relative reconstruction error =  $3.4 \times 10^{-4}$ ) and (c) 21 measurements (relative reconstruction error =  $1.4 \times 10^{-5}$ ).

light field images. Finally, the proposed design can be implemented by simple modifications of traditional cameras.

## CHAPTER 11

### Conclusions

Imaging systems are ubiquitous. From photography to medical diagnostic systems like magnetic resonance (MR) and computer assisted tomography (CT), they are an integral part of everyday life. A common problem with all imaging systems is that they can only acquire a degraded version of the original scene, and in most cases, these degradations are unavoidable. Therefore, computational approaches such as signal processing provide a powerful alternative way of restoring images. Moreover, as the nature of imaging is changing with new devices and application areas, such as computational photography, image-based rendering, novel camera designs that can capture more than single photographs, and new types of medical and nanoscale imaging systems such as multi-coil MRs and atomic force microscopes, the need for generally applicable frameworks and algorithms for image recovery is growing.

This thesis addresses the general field of image recovery, and devises a number of novel frameworks and algorithms for several problems in image recovery. We formulated the inverse problems from a Bayesian perspective, where the observed and unknown images, and all the model parameters are treated as stochastic quantities, and systematic stochastic models are constructed using prior knowledge about the degrading systems and unknown data. In this thesis, we first developed a new mathematical framework for image recovery using total variation-priors where the posterior distributions of the unknowns are approximated by simple, mathematically tractable forms. Our approach, which utilizes variational distribution approximations, can be seen as between the methods utilizing point estimates and sampling methods,

and it combines the positive aspects of these approaches: Very complex models of the imaging systems and images can be utilized within our framework, and memory and time requirements of the resulting algorithms are very low compared to the sampling approaches, which renders them practical in a wide range of applications.

A powerful aspect of our approach is that all algorithm parameters are embedded in our framework using a fully Bayesian formulation, and are estimated simultaneously with the unknown image. Therefore, the resulting algorithms are fully automated and do not require user intervention. This is very crucial in applications where the users do not have a technical background and parameter-tuning is inconvenient. We have also shown that despite the lack of user supervision the performance of the algorithms is comparable and in many cases higher than the performance of methods that require a high amount of supervision. Therefore, our methods can easily be applied to a wider range of applications of interest to people such as creative professionals in photography and advertising, medical staff, and scientists, among others. We applied this framework to image restoration, blind deconvolution and super resolution.

Because of the theoretical nature of the image recovery problem, our framework is applicable to a variety of areas. An important application area is compressive sensing (CS). We developed two novel algorithms for CS reconstruction. First, using a Bayesian framework, we proposed the use of Laplace priors for imposing sparsity to a higher extent than existing methods. Based on this model, we also developed a very efficient algorithm that reconstructs the unknowns in a greedy fashion. Furthermore, unlike other existing reconstruction methods, the resulting algorithm is fully automated with all required model parameters being estimated along with the unknown signal coefficients. The proposed method provides state-of-the-art reconstruction performance with very low computational complexity.

Second, we investigated a non-convex sparsity model for CS reconstruction, namely, a signal prior based on  $l_p$ -norms with  $0 < p < 1$ . We have shown that this prior captures the sparsity of the unknown signal better than  $l_1$ -norms. We have proposed a global optimization algorithm based on variational Bayesian inference. We have shown that the proposed formulation is a generalized version of some existing methods, such as reweighted least squares and sparse Bayesian methods. Experimental results have shown that although the developed algorithm is computationally expensive, it provides state-of-the-art reconstruction error performance, and therefore it can provide potential directions for improvement.

Finally, we have developed a novel application of compressive sensing for light-field image acquisition. Using the physical model of the camera lens and aperture, we have shown that the light field image acquisition process can be formulated as a incoherent measurement system. Moreover, we have shown that light field images inherently exhibit a large degree of sparsity. Combining these properties, we have shown that a traditional camera with a randomly coded aperture mask can be used to acquire light-field images with much fewer image captures than traditionally needed. Moreover, we have developed a novel reconstruction method which provides high-quality light field images with surprisingly low number of exposures. We believe that the real power of compressive sensing lies at the development of novel applications of such as this, and other interesting applications will be developed in the near future.

## References

- [1] Sparselab. <http://sparselab.stanford.edu/>.
- [2] StereoPhoto Maker. <http://stereo.jpn.org/eng/stphmkr/>.
- [3] MDSP super-resolution and demosaicing datasets. <http://www.soe.ucsc.edu/~milanfar/software/sr-datasets.html>, 2007.
- [4] K. Z. Adami. Variational methods in Bayesian deconvolution. *PHYSTAT2003 ECONF*, C030908:TUGT002, 2003.
- [5] T. Adelson and J. Wang. Single lens stereo with a plenoptic camera. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(2):99–106, Feb 1992.
- [6] G. L. Anderson and A. N. Netravali. Image restoration based on a subjective criterion. *IEEE Trans. Syst., Man, Cybern. 6*, pages 845–853, 1976.
- [7] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, Jan. 2003.
- [8] S. Babacan, L. Mancera, R. Molina, and A. Katsaggelos. Bayesian compressive sensing using non-convex priors. In *European Signal Processing Conference 2009 (EUSIPCO'09)*, Glasgow, Scotland, 2009.
- [9] S. Babacan, R. Molina, and A. Katsaggelos. Fast Bayesian compressive sensing using Laplace priors. In *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, April 2009.
- [10] S. Babacan, R. Molina, and A. Katsaggelos. Variational Bayesian blind deconvolution using a total variation prior. *IEEE Trans. Image Processing*, 18:12 – 26, January 2009.
- [11] S. Babacan, R. Molina, and A. Katsaggelos. Variational Bayesian super resolution. *submitted to IEEE Trans. Image Processing*, 2009.

- [12] S. D. Babacan, R. Molina, and A. Katsaggelos. Total variation blind deconvolution using a variational approach to parameter, image, and blur estimation. In *2007 European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, Sept. 2007.
- [13] S. D. Babacan, R. Molina, and A. Katsaggelos. Parameter estimation in TV image restoration using variational distribution approximation. *IEEE Trans. Image Processing*, (3):326–339, March 2008.
- [14] S. D. Babacan, R. Molina, and A. Katsaggelos. Bayesian compressive sensing using Laplace priors. *to appear in IEEE Trans. on Image Processing*, 2010.
- [15] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Total variation image restoration and parameter estimation using variational posterior distribution approximation. In *IEEE International Conf. on Image Processing (ICIP) 2007*, San Antonio, USA, Sept. 2007.
- [16] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Generalized Gaussian Markov field image restoration using variational distribution approximation. In *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, Nevada, February 2008.
- [17] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Total variation super resolution using a variational approach. In *IEEE International Conf. on Image Processing 2008*, October 2008.
- [18] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.
- [19] M. R. Banham and A. K. Katsaggelos. Digital image restoration. *IEEE Signal Processing Mag.*, 14(2):24–41, March 1997.
- [20] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, July 2007.
- [21] M. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.
- [22] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on bandspectroscopic data. *Technometrics*, 16:145–185, 1974.
- [23] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*, chapter 3 and 4. New York, Springer Verlag, 1985.

- [24] M. Bertalmio, V. Caselles, B. Roug , and A. Sol . TV based image restoration with local constraints. *J. Sci. Comput.*, 19(1-3):95–122, Dec. 2003.
- [25] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.
- [26] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira. Adaptive total-variation image deconvolution: A majorization-minimization approach. In *Proceedings of EUSIPCO'2006*, Florence, Italy, Sept. 2006.
- [27] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira. Total-variation image deconvolution: A majorization-minimization approach. In *ICASSP'2006*, Toulouse, France, May 2006.
- [28] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [29] C. Bishop and M. Svensen. Bayesian hierarchical mixtures of experts. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 57–64, San Francisco, CA, 2003. Morgan Kaufmann.
- [30] C. Bishop and M. Tipping. Variational relevance vector machine. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53. Morgan Kaufmann Publishers, 2000.
- [31] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.
- [32] C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [33] T. E. Bishop, S. D. Babacan, B. Amizic, A. K. Katsaggelos, T. Chan, and R. Molina. Blind image deconvolution: problem formulation and existing approaches. In P. Campisi and K. Egiazarian, editors, *Blind image deconvolution: theory and applications*, chapter 1. CRC press, 2007.
- [34] T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Trans. Signal Processing*, 56(6):2370–2382, June 2008.
- [35] T. Blumensath and M. E. Davies. A simple, efficient and near optimal algorithm for compressed sensing. In *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, 2009.
- [36] N. Bose, S. Lertrattanapanich, and J. Koo. Advances in superresolution using L-curve. *IEEE International Symposium on Circuits and Systems*, 2:433–436, 2001.

- [37] C. A. Bouman and K. Sauer. Generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Processing*, 2(3):296–310, 1993.
- [38] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(11):1222–1239, Nov 2001.
- [39] T. Bretschneider, P. Bones, S. McNeill, and D. Pairman. Image-based quality assessment of SPOT data. In *Proceedings of the American Society for Photogrammetry & Remote Sensing*, 2001. unpaginated CD-ROM.
- [40] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 2, pages 60–65, June 2005.
- [41] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, Feb. 2006.
- [42] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.
- [43] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted  $\ell_1$  minimization. *Journal of Fourier Analysis and Applications (Special issue on sparsity)*, 14(5):877–905, December 2008.
- [44] F. Candocia and J. Principe. Superresolution of images with learned multiple reconstruction kernels. In L. Guan, S.-Y. Kung, and J. Larsen, editors, *Multimedia Image and Video Processing*, pages 67–95. CRC Press, 2000.
- [45] E. Cands. Compressive sampling. In *Int. Congress of Mathematics 3*, pages 1433–1452, Madrid, Spain, 2006.
- [46] D. Capel and A. Zisserman. Super-resolution enhancement of text image sequence. In *International Conference on Pattern Recognition*, pages 600–605, 2000.
- [47] D. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 627–634, 2001.
- [48] D. Capel and A. Zisserman. Computer vision applied to super resolution. *IEEE Signal Processing Magazine*, 20(3):75–86, May 2003.

- [49] W. T. Cathey and E. R. Dowski. New paradigm for imaging systems. *Appl. Opt.*, 41(29):6080–6092, 2002.
- [50] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167 – 88, 1997.
- [51] R. Chan, T. Chan, L. Shen, and Z. Shen. Wavelet algorithms for high-resolution image reconstruction. *SIAM Journal of Scientific Computing*, 24:1408–1432, 2003.
- [52] R. H. Chan, T. F. Chan, and C.-K. Wong. Cosine transform based preconditioners for total variation deblurring. *IEEE Trans. Image Processing*, 8(10):1472–1478, Oct 1999.
- [53] T. Chan, S. Esedoglu, F. Park, and A. Yip. Recent developments in total variation image restoration. In N. Paragios, Y. Chen, and O. Faugeras, editors, *Handbook of Mathematical Models in Computer Vision*. Springer Verlag, 2005.
- [54] T. C. Chan and J. Shen. *Image Processing And Analysis: Variational, PDE, Wavelet, And Stochastic Methods*. SIAM, 2005.
- [55] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20(6):1964–1977, November 1999.
- [56] T. F. Chan, M. K. Ng, A. C. Yau, and A. M. Yip. Superresolution image reconstruction using fast inpainting algorithms. *Applied and Computational Harmonic Analysis*, 23(1):3 – 24, 2007. Special Issue on Mathematical Imaging.
- [57] T. F. Chan, N. Ng, A. Yau, and A. Yip. Superresolution image reconstruction using fast inpainting algorithms. *Applied and Computational Harmonic Analysis*, 23(1):3–24, July 2007.
- [58] T. F. Chan and C.-K. Wong. Total variation blind deconvolution. *IEEE Trans. Image Processing*, 7(3):370–375, Mar. 1998.
- [59] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Processing Letters, IEEE*, 14(10):707–710, Oct. 2007.
- [60] R. Chartrand and W. Tin. Iteratively reweighted algorithms for compressive sensing. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3869–3872, 31 2008-April 4 2008.
- [61] S. Chaudhuri and J. Manjunath. *Motion-free super-resolution*. Springer, 2005.

- [62] J. Chen, L. Yuan, C.-K. Tang, and L. Quan. Robust dual motion deblurring. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, June 2008.
- [63] L. Chen and K.-H. Yap. A soft double regularization approach to parametric blind image deconvolution. *IEEE Trans. Image Processing*, 14(5):624–633, 2005.
- [64] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.
- [65] B. A. Chipman and B. D. Jeffs. Blind multiframe point source image restoration using MAP estimation. *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, 2:1267–1271, 1999.
- [66] S. Dai, M. Yang, Y. Wu, and A. K. Katsaggelos. Tracking motion-blurred targets in video. In *International Conference on Image Processing (ICIP'06)*, Atlanta, GA, October 2006.
- [67] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Gunturk. Iteratively re-weighted least squares minimization for sparse recovery, 2009.
- [68] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 369–378, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.
- [69] A. D. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the E-M algorithm. *Journal of the Royal statistical Society, Series B*, 39:1–37, 1977.
- [70] H. Derin and H. Elliott. Modelling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-9(1):39–55, Jan 1987.
- [71] R. A. DeVore. Deterministic constructions of compressed sensing matrices. *J. of Complexity*, 23(4-6):918–925, 2007.
- [72] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [73] D. L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences*, 102(27):9446–9451, 2005.

- [74] D. L. Donoho and Y. Tsaig. Sparse solution of underdetermined linear equations. by stagewise orthogonal matching pursuit. *Preprint*, March 2006.
- [75] S. N. Efstratiadis and A. K. Katsaggelos. Adaptive iterative image restoration with reduced computational load. *Optical Engineering*, 29:1458–1468, 1990.
- [76] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans. Image Processing*, 6(12):1646–1658, Dec 1997.
- [77] H. Farid and E. P. Simoncelli. Range estimation by optical differentiation. *JOSA A*, 15:1777–1786, 1998.
- [78] S. Farsiu. *MDSP resolution enhancement software*. University of California at Santa Cruz, 2004.
- [79] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Trans. Image Processing*, 13(10):1327–1344, Oct. 2004.
- [80] E. E. Fenimore and T. M. Cannon. Coded aperture imaging with uniformly redundant arrays. *Appl. Opt.*, 17(3):337–347, 1978.
- [81] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. Freeman. Removing camera shake from a single photograph. *ACM Transactions on Graphics, SIGGRAPH 2006 Conference Proceedings, Boston, MA*, 25:787–794, 2006.
- [82] M. Figueiredo, J. Bioucas-Dias, and R. Nowak. Majorizationminimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Processing*, 16(12):2980–2991, Dec. 2007.
- [83] M. Figueiredo, R. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Trans. on Selected Topics in Signal Processing*, 1(4):586–597, December 2007.
- [84] M. A. T. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1150–1159, 2003.
- [85] W. Freeman, T. Jones, and E. Pasztor. Example based super-resolution. *IEEE Computer Graphics and Applications*, 22:56–65, 2002.
- [86] N. P. Galatsanos, V. Z. Mesarovic, R. Molina, and A. K. Katsaggelos. Hierarchical Bayesian image restoration for partially-known blur. *IEEE Trans. Image Processing*, 9(10):1784–1797, Aug. 2000.

- [87] N. P. Galatsanos, V. Z. Mesarovic, R. Molina, A. K. Katsaggelos, and J. Mateos. Hyper-parameter estimation in image restoration problems with partially-known blurs. *Optical Eng.*, 41(8):1845–1854, Aug. 2002.
- [88] L. Gan, T. Do, and T. Tran. Fast compressive imaging using scrambled block Hadamard ensemble. In *EUSIPCO 2008*, Lausanne, Switzerland, August 2008.
- [89] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express*, 15(21):14013–14027, 2007.
- [90] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall., 2003.
- [91] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383, 1992.
- [92] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing*, 4(7):932–946, 1995.
- [93] S. Geman and D. Geman. Stochastic Relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-6(6):721–741, 1984.
- [94] T. Georgiev, C. Intwala, S. D. Babacan, and A. Lumsdaine. Unified frequency domain analysis of lightfield cameras. In *ECCV*, Marseille, France, December 2008.
- [95] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala. Spatio-angular resolution tradeoffs in integral photography. In *EGSR*, pages 263–272, june 2006.
- [96] F. S. Gibson and F. Lanni. Experimental test of an analytical model of aberration in an oil-immersion objective lens used in three-dimensional light microscopy. *Journal of the Optical society of America-A*, 8:1601–1613, 1991.
- [97] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Comp.*, 13(11):2517–2532, 2001.
- [98] G. H. Golub and C. F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*. The Johns Hopkins University Press, October 1996.
- [99] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. *ACM Trans. Graph.*, pages 43–54, 1996.

- [100] S. R. Gottesman and E. E. Fenimore. New family of binary arrays for coded aperture imaging. *Appl. Opt.*, 28(20):4344–4352, 1989.
- [101] J. Gu, S. K. Nayar, E. Grinspun, P. N. Belhumeur, and R. Ramamoorthi. Compressive Structured Light for Recovering Inhomogeneous Participating Media. In *European Conference on Computer Vision (ECCV)*, Oct 2008.
- [102] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Trans. Image Processing*, 12:597–606, 2003.
- [103] S. Gurevich, R. Hadani, and N. Sochen. On some deterministic dictionaries supporting sparsity. *Special issue on sparsity, the Journal of Fourier Analysis and Applications*, 14:859–876, December 2008.
- [104] A. Hamza, H. Krim, and G. Unal. Unifying probabilistic and variational estimation. *IEEE Signal Processing Magazine*, 19:37–47, Sep. 2002.
- [105] R. Hardie, K. Barnard, and E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Trans. Image Processing*, 6(12):1621–1633, 1997.
- [106] R. Hardie, K. Barnard, J. Bognar, E. Armstrong, and E. Watson. High resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 73:247–260, 1998.
- [107] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak. Compressed sensing for networked data [a different approach to decentralized compression]. *IEEE Signal Processing Magazine*, 25(2):92–101, March 2008.
- [108] H. He and L. Kondi. An image super-resolution algorithm for different error levels per frame. *IEEE Trans. Image Processing*, 15(3):592–603, 2006.
- [109] Y. He, K. H. Yap, L. Chen, and L. P. Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Trans. Image Processing*, (11):2830–2841, 2007.
- [110] F. Humblot and A. Mohammad-Djafari. Super-resolution using hidden Markov model and Bayesian detection estimation framework. *EURASIP Journal on Applied Signal Processing*, 2006:Article ID 36971, 16 pages, 2006.
- [111] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, 1993.

- [112] F. Ives. Parallax stereogram and process of making same. *Patent US 725,567*, 1903.
- [113] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.
- [114] F.-C. Jeng and J. W. Woods. Compound Gauss-Markov random fields for image estimation. *IEEE Transactions on Signal Processing*, 39(3):683–697, 1991.
- [115] S. Ji, D. Dunson, and L. Carin. Multitask compressive sensing. *IEEE Trans. Signal Processing*, 57(1):92–106, Jan. 2009.
- [116] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Trans. Signal Processing*, 56(6):2346–2356, June 2008.
- [117] M. I. Jordan, Z. Ghahramani, T. S. Jaakola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in Graphical Models*, pages 105–162. MIT Press, 1998.
- [118] A. Kanemura, S.-I. Maeda, and S. Ishii. Superresolution with compound Markov random fields via the variational EM algorithm. *Neural Networks*, 22(7):1025 – 1034, 2009.
- [119] M. Kang and S. Chaudhuri (Eds.). Super-resolution image reconstruction. *IEEE Signal Processing Magazine*, 20(3), 2003.
- [120] A. Katsaggelos, editor. *Digital Image Restoration*. Springer Series in Information Sciences, vol. 23, Springer-Verlag, 1991.
- [121] A. Katsaggelos, K. T. Lay, and N. P. Galatsanos. A general framework for frequency domain multi-channel signal processing. *IEEE Trans. Image Processing*, 2(3):417–420, July 1993.
- [122] A. K. Katsaggelos. *Iterative Image Restoration Algorithms*. PhD thesis, Georgia Institute of Technology, School of Electrical Engineering, August 1985.
- [123] A. K. Katsaggelos. A multiple input image restoration approach. *Journal of Visual Comm. and Image Representation*, 1:93–103, September 1990.
- [124] A. K. Katsaggelos, J. Biemond, R. W. Schafer, and R. M. Mersereau. A regularized iterative image restoration algorithm. *IEEE Trans. Signal Processing*, 39(4):914–929, April 1991.
- [125] A. K. Katsaggelos and M. G. Kang. A spatially adaptive iterative algorithm for the restoration of astronomical images. *International Journal of Imaging Systems and*

*Technology, special issue on "Image Reconstruction and Restoration in Astronomy"*, 6(4):305–313, 1995.

- [126] A. K. Katsaggelos and K. T. Lay. Maximum likelihood blur identification and image restoration using the EM algorithm. *IEEE Trans. Signal Processing*, 39(3):729–733, 1991.
- [127] A. K. Katsaggelos and K. T. Lay. Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In A. K. Katsaggelos, editor, *Digital Image Restoration*. Springer-Verlag, 1991.
- [128] A. K. Katsaggelos, R. Molina, and J. Mateos. *Super Resolution of Images and Video*. Morgan and Claypool, 2007.
- [129] C. Kerfrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *IEEE Trans. Image Processing*, 15(10):2866–2878, October 2006.
- [130] H. Kim, J.-H. Jang, and K.-S. Hong. Edge-enhancing super-resolution using anisotropic diffusion. In *Proceedings of the IEEE Conference on Image Processing*, volume 3, pages 130–133, 2001.
- [131] J. Krist. Simulation of HST PSFs using Tiny Tim. In R. A. Shaw, H. E. Payne, and J. J. E. Hayes, editors, *Astronomical Data Analysis Software and Systems IV*, pages 349–353, San Francisco, USA, 1995. Astronomical Society of the Pacific.
- [132] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel. Adaptive noise smoothing filter for images with signal-dependent noise. *IEEE Trans. Pattern Anal. Machine Intell.*, 7(2):165–177, March 1985.
- [133] S. Kullback. *Information Theory and Statistics*. New York, Dover Publications, 1959.
- [134] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
- [135] R. L. Lagendijk, J. Biemond, and D. E. Boekee. Regularized iterative image restoration with ringing reduction. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(12):1874–1888, December 1988.
- [136] R. L. Lagendijk, J. Biemond, and D. E. Boekee. Identification and restoration of noisy blurred images using the expectation-maximization algorithm. *IEEE Trans. Acoust., Speech, Signal Processing*, 38:1180–1191, July 1990.
- [137] K. Lange. *Optimization*. Springer Texts in Statistics, Springer-Verlag, 2004.

- [138] C. L. Lawson. *Contributions to the theory of linear least maximum approximations*. PhD thesis, UCLA, 1961.
- [139] K. T. Lay and A. K. Katsaggelos. Image identification and image restoration based on the expectation-maximization algorithm. *Optical Eng.*, 29(5):436–445, May 1990.
- [140] E. S. Lee and M. G. Kang. Regularized adaptive high-resolution image reconstruction considering inaccurate subpixel registration. *IEEE Trans. Image Processing*, 12(7):826–837, July 2003.
- [141] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM Transactions on Graphics, SIGGRAPH 2007 Conference Proceedings*, page 70, New York, NY, USA, 2007. ACM.
- [142] M. Levoy and P. Hanrahan. Light field rendering. *ACM Trans. Graph.*, pages 31–42, 1996.
- [143] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen. Programmable aperture photography: multiplexed light field acquisition. *ACM Trans. Graph.*, pages 1–10, 2008.
- [144] A. Likas and N. Galatsanos. A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 52(8):2222–2233, 2004.
- [145] G. Lippmann. Epreuves reversibles donnant la sensation du relief. *J. Phys.* 7, pages 821–825, 1908.
- [146] A. Lopez, R. Molina, A. Katsaggelos, A. Rodriguez, J. Lpez, and J. Llamas. Parameter estimation in bayesian reconstruction of spect images: An aide in nuclear medicine diagnosis. *International Journal of Imaging Systems and Technology*, 14(1):21–27, June 2004.
- [147] P.-Y. Lu, T.-H. Huang, M.-S. Wu, Y.-T. Cheng, and Y.-Y. Chuang. High dynamic range image reconstruction from hand-held cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 509–516, June 2009.
- [148] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981.
- [149] L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astron. J.*, 79:745+, June 1974.

- [150] M. Lustig, D. L. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, December 2007.
- [151] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty. An efficient algorithm for compressed MR imaging using total variation and wavelets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, June 2008.
- [152] D. J. C. MacKay. Bayesian interpolation. *Neural Comput.*, 4(3):415–447, 1992.
- [153] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
- [154] S. Mann. Compositing multiple pictures of the same scene. In *Proceedings of the 46th Annual IS&T Conference*, pages 50–52, Cambridge, Massachusetts, May 1993. The Society of Imaging Science and Technology.
- [155] R. Marcia and R. Willett. Compressive coded aperture video reconstruction. In *EUSIPCO 2008*, Lausanne, Switzerland, August 2008.
- [156] R. Marcia and R. Willett. Compressive coded aperture superresolution image reconstruction. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 833–836, 31 2008–April 4 2008.
- [157] R. F. Marcia, Z. T. Harmany, and R. M. Willett. Compressive coded aperture imaging. In C. A. Bouman, E. L. Miller, and I. Pollak, editors, *Computational Imaging VII*, volume 7246, page 72460G. SPIE, 2009.
- [158] K. Mardia, J. Kent, and J. Bibby. *Multivariate analysis*. Academic Press; New York, 1979.
- [159] J. Mateos, A. Katsaggelos, and R. Molina. Simultaneous motion estimation and resolution enhancement of compressed low resolution video. In *IEEE International Conference on Image Processing*, volume 2, pages 653–656, 2000.
- [160] J. Mateos, A. K. Katsaggelos, and R. Molina. A Bayesian approach to estimate and transmit regularization parameters for reducing blocking artifacts. *IEEE Trans. Image Processing*, 9(7):1200–1215, July 2000.
- [161] O. Michailovich and D. Adam. A novel approach to the 2-D blind deconvolution problem in medical ultrasound. *IEEE Trans. Med. Imag.*, 24(1):86–104, 2005.
- [162] J. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, Astrophysics Group, University of Cambridge, 2000.

- [163] J. W. Miskin and D. J. C. MacKay. Ensemble learning for blind image separation and deconvolution. In M. Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag Scientific Publishers, July 2000.
- [164] A. F. J. Moffat. A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. *Astronomy and Astrophysics*, 3:455–461, 1969.
- [165] A. Mohan, X. Huang, J. Tumblin, and R. Raskar. Sensing increased image resolution using aperture masks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, June 2008.
- [166] R. Molina. On the hierarchical Bayesian approach to image restoration. Applications to Astronomical images. *IEEE Trans. Pattern Anal. Machine Intell.*, 16(11):1122–1128, Nov. 1994.
- [167] R. Molina, A. K. Katsaggelos, J. Abad, and J. Mateos. A Bayesian approach to blind deconvolution based on Dirichlet distributions. In *1997 International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume IV, pages 2809–2812, Munich (Germany), 1997.
- [168] R. Molina, A. K. Katsaggelos, and J. Mateos. Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Trans. Image Processing*, 8(2):231–246, 1999.
- [169] R. Molina, J. Mateos, and A. Katsaggelos. Blind deconvolution using a variational approach to parameter, image, and blur estimation. *IEEE Trans. Image Processing*, 15(12):3715–3727, Dec. 2006.
- [170] R. Molina, J. Mateos, and A. Katsaggelos. Super resolution reconstruction of multispectral images. In *Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing*, pages 211–220. Heron Press, 2006.
- [171] R. Molina and B. D. Ripley. Using spatial models as priors in astronomical image analysis. *Journal of Applied Statistics*, 16:193–206, 1989.
- [172] R. Molina, M. Vega, J. Abad, and A. Katsaggelos. Parameter estimation in Bayesian high-resolution image reconstruction with multisensors. *IEEE Trans. Image Processing*, 12(12):1655–1667, 2003.
- [173] R. Molina, M. Vega, J. Mateos, and A. Katsaggelos. Hierarchical Bayesian super resolution reconstruction of multispectral images. In *2006 European Signal Processing Conference (EUSIPCO 2006)*. Florence (Italy), September 2006.

- [174] S. K. Nayar and V. Branzoi. Adaptive dynamic range imaging: Optical control of pixel exposures over space and time. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1168, Washington, DC, USA, 2003. IEEE Computer Society.
- [175] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, University of Toronto, 1993. available online at <http://www.cs.toronto.edu/~radford/res-mcmc.html>.
- [176] M. Ng, T. Chan, M. Kang, and P. Milanfar. Super-resolution imaging: Analysis, algorithms, and applications. *EURASIP Journal on Applied Signal Processing*, 2006:Article ID 90531, 2 pages, 2006.
- [177] M. Ng, J. Koo, and N. Bose. Constrained total least-squares computations for high-resolution image reconstruction with multisensors. *International Journal of Imaging Systems and Technology*, 12(1):35–42, 2002.
- [178] M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang. A total variation regularization based super-resolution reconstruction algorithm for digital video. *EURASIP Journal on Advances in Signal Processing*, (74585), 2007.
- [179] R. Ng, M. Levoy, M. Brdif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Stanford Tech. Rep.*, 2005.
- [180] N. Nguyen. *Numerical Algorithms for superresolution*. PhD thesis, Stanford University, 2001.
- [181] N. Nguyen, P. Milanfar, and G. Golub. Blind superresolution with generalized cross-validation using Gauss-type quadrature rules. In *33rd Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 1257–1261, 1999.
- [182] N. Nguyen, P. Milanfar, and G. Golub. A computationally efficient superresolution image reconstruction algorithm. *IEEE Trans. Image Processing*, 10:573–583, 2001.
- [183] N. Nguyen, P. Milanfar, and G. Golub. Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Trans. Image Processing*, 10:1299–1308, 2001.
- [184] P. Nisenson and R. Barakat. Partial atmospheric correction with adaptive optics. *Journal of the Optical society of America-A*, 4:2249–2253, 1991.
- [185] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

- [186] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao. Variational EM algorithms for non-Gaussian latent variable models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1059–1066. MIT Press, Cambridge, MA, 2006.
- [187] G. Parisi. *Statistical Field Theory*. Addison-Wesley, Redwood City, CA, 1988.
- [188] S. Park, M. Park, and M. Kang. Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, 20:21–36, 2003.
- [189] W. Pearlman and W. Song. A robust method for restoration of photon-limited blurred images. In *Proc. SPIE*, volume 504, 1984.
- [190] L. Pickup, S. Roberts, and A. Zisserman. A sampled texture prior for image super-resolution. In *Advances in Neural Information Processing Systems*, pages 1587–1594, 2003.
- [191] L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman. Bayesian methods for image super-resolution. *The Computer Journal*, 2007.
- [192] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Processing*, 12(11):1338–1351, November 2003.
- [193] M. Protter, M. Elad, H. Takeda, and P. Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. Image Processing*, 18(1):36–51, Jan. 2009.
- [194] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business, Administration, Harvard University, Boston, 1961.
- [195] D. Rajan and S. Chaudhuri. Generation of super-resolution images from blurred observations using an MRF model. *Journal of Mathematical Imaging and Vision*, 16:5–15, 2002.
- [196] B. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, 47(1):187–200, Jan 1999.
- [197] R. Raskar, A. Agrawal, and J. Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 795–804, New York, NY, USA, 2006. ACM.
- [198] A. Rav-Acha and S. Peleg. Two motion-blurred images are better than one. *Pattern Recogn. Lett.*, 26(3):311–317, 2005.

- [199] H. W. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, January 1972.
- [200] B. D. Ripley. *Spatial Statistics*. John Wiley, 1981.
- [201] M. Roggemann. Limited degree-of-freedom adaptive optics and image reconstruction. *Applied Optics*, 30:4227–4233, 1991.
- [202] J. J. Ruanaidh and W. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer Series in Statistics and Computing. Springer, New York, 1 edition, 1996.
- [203] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. In *Proceedings of the eleventh annual international conference of the Center for Nonlinear Studies on Experimental mathematics: computational issues in nonlinear science*, pages 259–268, Amsterdam, The Netherlands, The Netherlands, 1992. Elsevier North-Holland, Inc.
- [204] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, pages 259–268, 1992.
- [205] P. Sarder and A. Nehorai. Deconvolution methods for 3D fluorescence microscopy images: an overview. *IEEE Signal Processing Mag.*, 23:32–45, May 2006.
- [206] R. Schultz and R. Stevenson. Extraction of high resolution frames from video sequences. *IEEE Trans. Image Processing*, 5:996–1011, 1996.
- [207] R. R. Schultz and R. L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Trans. Image Processing*, 5(6):996–1011, 1996.
- [208] T. J. Schultz. Multiframe blind deconvolution of astronomical images. *Journal of the Optical society of America-A*, 10:1064–1073, 1993.
- [209] M. W. Seeger and H. Nickisch. Compressed sensing and Bayesian experimental design. In *International Conference on Machine Learning (ICML)*, pages 912–919, 2008.
- [210] C. Segall, R. Molina, and A. Katsaggelos. High-resolution images from low-resolution compressed video. *IEEE Signal Processing Magazine*, 20:37–48, 2003.
- [211] C. Segall, R. Molina, A. Katsaggelos, and J. Mateos. Bayesian resolution enhancement of compressed video. *IEEE Trans. Image Processing*, 13(7):898–911, 2004.

- [212] C. A. Segall, R. Molina, and A. K. Katsaggelos. High-resolution images from low-resolution compressed video. *IEEE Signal Processing Mag.*, 20(3):37–48, 2003.
- [213] P. Sen and S. Darabi. Compressive Dual Photography. *Computer Graphics Forum*, 28(2):609 – 618, 2009.
- [214] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics, SIGGRAPH 2008 Conference Proceedings*, 2008.
- [215] V. Smidl and A. Quinn. *The variational Bayes method in Signal Processing*. Springer Verlag, 2005.
- [216] F. Šroubek, G. Cristobal, and J. Flusser. A unified approach to superresolution and multichannel blind deconvolution. *IEEE Trans. Image Processing*, 16(9):2322–2332, Sept. 2007.
- [217] F. Šroubek and J. Flusser. Multichannel blind iterative image restoration. *IEEE Trans. Image Processing*, 12(9):1094–1106, 2003.
- [218] F. Šroubek and J. Flusser. Multichannel blind deconvolution of spatially misaligned images. *IEEE Trans. Image Processing*, 7:45–53, July 2005.
- [219] H. Stark and P. Oskoui. High resolution image recovery from image-plane arrays, using convex projections. *Journal of the Optical Society of America A*, 6:1715–1726, 1989.
- [220] E. Sudderth, A. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 605–612, 2003.
- [221] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Trans. Image Processing*, 18(9):1958–1975, Sept. 2009.
- [222] A. Tekalp, M. Ozkan, and M. Sezan. High-resolution image reconstruction from lower-resolution image sequences and space varying image restoration. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 169–172, 1992.
- [223] A. M. Tekalp and G. Pavlovic. *Restoration of Scanned Photographic Images*,. Chp. 8 in Digital Image Restoration, Ed. A. Katsaggelos, Springer-Verlag, Berlin, 1991.
- [224] M. Tico and M. Vehvilainen. Image stabilization based on fusing the visual information in differently exposed images. *IEEE International Conference on Image Processing (ICIP 2007)*, 1:117–120, 16 2007-Oct. 19 2007.

- [225] M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, pages 211–244, 2001.
- [226] M. Tipping and C. Bishop. Bayesian image super-resolution. In S. Thrun, S. Becker, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1279–1286, Cambridge, MA, 2003. MIT Press.
- [227] M. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.
- [228] M. E. Tipping and C. M. Bishop. Bayesian image super-resolution. In *Advances in Neural Information Processing Systems 15 (NIPS)*. MIT Press, 2003.
- [229] B. Tom and A. Katsaggelos. Reconstruction of a high resolution image from multiple-degraded and misregistered low-resolution images. In *Proceedings of SPIE Conference on Visual Communications and Image Processing*, volume 2308, pages 971–981, 1994.
- [230] B. Tom and A. Katsaggelos. Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images”. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 539–542, 1995.
- [231] B. Tom and A. Katsaggelos. Resolution enhancement of monochrome and color video using motion compensation. *IEEE Trans. Image Processing*, 10:278–287, 2001.
- [232] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, Dec. 2007.
- [233] Y. Tsaig and D. L. Donoho. Extensions of compressed sensing. *Signal Process.*, 86(3):549–571, 2006.
- [234] P. Vandewalle, L. Sbaiz, J. Vandewalle, and M. Vetterli. Super-Resolution from Unregistered and Totally Aliased Signals Using Subspace Methods. *IEEE Trans. Signal Processing*, 55(7, Part 2):3687–3703, 2007.
- [235] P. Vandewalle, S. Ssstrunk, and M. Vetterli. A Frequency Domain Approach to Registration of Aliased Images with Application to Super-Resolution. *EURASIP Journal on Applied Signal Processing (special issue on Super-resolution)*, 2006:Article ID 71459, 14 pages, 2006.
- [236] P. Vandewalle, P. Zbinden, and C. Perez. Superresolution v2.0. <http://1cavwww.epfl.ch/software/superresolution/index.html>, 2006.

- [237] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69:1–69:12, July 2007.
- [238] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996.
- [239] C. R. Vogel and M. E. Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Image Processing*, 7(6):813–824, June 1998.
- [240] Q. Wang, X. Tang, and H. Shum. Patch based blind image super resolution. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 709–716, 2005.
- [241] G. Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of graphics, gpu, and game tools*, 8(2):17–30, 2003.
- [242] J. Weickert. A review of nonlinear diffusion filtering. In *SCALE-SPACE '97: Proceedings of the First International Conference on Scale-Space Theory in Computer Vision*, pages 3–28, London, UK, 1997. Springer-Verlag.
- [243] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, 2005.
- [244] R. Willett, I. Jermyn, R. Nowak, and J. Zerubia. Wavelet-based superresolution in Astronomy. In *Proc. Astronomical Data Analysis Software and Systems*, Strasbourg, France, October 2003.
- [245] D. Wipf, J. Palmer, and B. D. Rao. Perspectives on sparse Bayesian learning. *Advances in Neural Information Processing Systems*, (16), 2004.
- [246] D. Wipf, J. Palmer, B. D. Rao, and K. Kreutz-Delgado. Performance analysis of latent variable models with sparse priors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, USA, May 2007.
- [247] D. Wipf and B. Rao. Sparse Bayesian learning for basis selection. *IEEE Trans. Signal Processing*, 52(8):2153–2164, Aug. 2004.
- [248] C. S. Won and R. M. Gray. *Stochastic Image Processing*. Information Technology: Transmission, Processing, and Storage. Kluwer Academic / Plenum Publishers, 2004.

- [249] N. Woods, N. Galatsanos, and A. Katsaggelos. Stochastic methods for joint registration, restoration, and interpolation of multiple undersampled images. *IEEE Trans. Image Processing*, 15:201–213, 2006.
- [250] Y. L. You and M. Kaveh. A regularization approach to joint blur and image restoration. *IEEE Trans. Image Processing*, 5(3):416–428, 1996.
- [251] Y.-L. You and M. Kaveh. Ringing reduction in image restoration by orientation-selective regularization. *IEEE Signal Processing Letters*, 3(2):29–31, 1996.
- [252] Y. L. You and M. Kaveh. Blind image restoration by anisotropic regularization. *IEEE Trans. Image Processing*, 8(3):396–407, 1999.
- [253] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum. Blurred/non-blurred image alignment using sparseness prior. In *IEEE International Conference on Computer Vision (ICCV 2007)*, pages 1–8, October 2007.
- [254] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum. Image deblurring with blurred/noisy image pairs. In *ACM Transactions on Graphics, SIGGRAPH 2007 Conference Proceedings*, page 1, New York, NY, USA, 2007. ACM.
- [255] J. Zhang. The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE Trans. Image Processing*, 2(1):27–40, 1993.
- [256] A. Zomet and S. K. Nayar. Lensless imaging with a controllable aperture. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 339–346, Washington, DC, USA, 2006. IEEE Computer Society.
- [257] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 645–650, 2001.

## APPENDIX A

**Calculation of the image estimates in Algorithms 1 and 2**

To obtain the image estimates, the mean of the distribution  $q^k(\mathbf{x})$  in Eq. (3.33) is used in Algorithm1 and the point estimate in Eq. (3.48) is used in Algorithm2. The estimation of the quantities can be carried out by the Gradient Descent (GD) or the Conjugate Gradient (CG) methods. Note that by using the GD or the CG methods we avoid the calculation of the inverse of the covariance matrix. Our descriptions will be specifically for Algorithm 1. However, the same results apply to Algorithm 2. We next describe the specific GD steps applied to the solution of

$$(A.1) \quad \mathbf{A}^k \mathbf{x} = E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{y}$$

where

$$(A.2) \quad \mathbf{A}^k = \left( E_{q^k(\beta)}[\beta] \mathbf{H}' \mathbf{H} + E_{q^k(\alpha)}[\alpha] (\Delta^h)^t W(\mathbf{u}^k) (\Delta^h) + E_{q^k(\alpha)}[\alpha] (\Delta^v)^t W(\mathbf{u}^k) (\Delta^v) \right).$$

In the description that follows we use the notation  $i : +1, i : +2, i : +3, i : +4$  to denote the four pixels around pixel  $i$  (if  $i = (u, v)$  they correspond to  $(u+1, v), (u, v+1), (u-1, v)$ , and  $(u, v-1)$ , respectively) .

We now expand the matrix  $E_{Q^k(\alpha)}[\alpha](\Delta^h)^t W(\mathbf{u}^k)(\Delta^h) + E_{Q^k(\alpha)}[\alpha](\Delta^v)^t W(\mathbf{u}^k)(\Delta^v)$  and calculate  $[\mathbf{A}^k \mathbf{x}]$  at position  $i$ ,  $[\mathbf{A}^k \mathbf{x}]_i$ . We have

$$\begin{aligned}
 [\mathbf{A}^k \mathbf{x}]_i &= [E_{Q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} \mathbf{x}]_i + 2E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_i^k}} \mathbf{x}_i + E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+3}^k}} \mathbf{x}_i + E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+4}^k}} \mathbf{x}_i \\
 &\quad - E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+1}^k}} \mathbf{x}_{i+1} - E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+2}^k}} \mathbf{x}_{i+2} - E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+3}^k}} \mathbf{x}_{i+3} \\
 (A.3) \quad &\quad - E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_{i+4}^k}} \mathbf{x}_{i+4}.
 \end{aligned}$$

Let us now define

$$(A.4) \quad \mathbf{z}_j(\mathbf{u}^k) = E_{Q^k(\alpha)}[\alpha] \frac{1}{\sqrt{u_j^k}}.$$

Using this, we obtain

$$\begin{aligned}
 [\mathbf{A}^k \mathbf{x}]_i &= [E_{Q^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} \mathbf{x}]_i + 2\mathbf{z}_i(\mathbf{u}^k) \mathbf{x}_i + \mathbf{z}_{i+3}(\mathbf{u}^k) \mathbf{x}_i + \mathbf{z}_{i+4}(\mathbf{u}^k) \mathbf{x}_i \\
 (A.5) \quad &\quad - \mathbf{z}_i(\mathbf{u}^k) (\mathbf{x}_{i+1} + \mathbf{x}_{i+2}) - \mathbf{z}_{i+3}(\mathbf{u}^k) \mathbf{x}_{i+3} - \mathbf{z}_{i+4}(\mathbf{u}^k) \mathbf{x}_{i+4}.
 \end{aligned}$$

Combining with Eq. (A.1), we obtain

$$\begin{aligned}
 2\mathbf{z}_i(\mathbf{u}^k) \mathbf{x}_i + \mathbf{z}_{i+3}(\mathbf{u}^k) \mathbf{x}_i + \mathbf{z}_{i+4}(\mathbf{u}^k) \mathbf{x}_i &= \mathbf{z}_i(\mathbf{u}^k) (\mathbf{x}_{i+1} + \mathbf{x}_{i+2}) \\
 &\quad + \mathbf{z}_{i+3}(\mathbf{u}^k) \mathbf{x}_{i+3} + \mathbf{z}_{i+4}(\mathbf{u}^k) \mathbf{x}_{i+4} \\
 (A.6) \quad &\quad + E_{Q^k(\beta)}[\beta] [\mathbf{H}^t (\mathbf{y} - \mathbf{H} \mathbf{x})]_i.
 \end{aligned}$$

Adding  $E_{\mathbf{q}^k(\beta)}[\beta]\mathbf{x}_i$  to both sides of the above equation we have

$$\begin{aligned}
 & 2\mathbf{z}_i(\mathbf{u}^k)\mathbf{x}_i + \mathbf{z}_{i+3}(\mathbf{u}^k)\mathbf{x}_i + \mathbf{z}_{i+4}(\mathbf{u}^k)\mathbf{x}_i + E_{\mathbf{q}^k(\beta)}[\beta]\mathbf{x}_i \\
 &= \mathbf{z}_i(\mathbf{u}^k)(\mathbf{x}_{i+1} + \mathbf{x}_{i+2}) + \mathbf{z}_{i+3}(\mathbf{u}^k)\mathbf{x}_{i+3} + \mathbf{z}_{i+4}(\mathbf{u}^k)\mathbf{x}_{i+4} \\
 (A.7) \quad &+ E_{\mathbf{q}^k(\beta)}[\beta][\mathbf{H}^t(\mathbf{y} - \mathbf{H}\mathbf{x})]_i + \mathbf{x}_i.
 \end{aligned}$$

Let

$$(A.8) \quad \mathbf{r}_i(\mathbf{u}^k) = 2\mathbf{z}_i(\mathbf{u}^k) + \mathbf{z}_{i+3}(\mathbf{u}^k) + \mathbf{z}_{i+4}(\mathbf{u}^k) + E_{\mathbf{q}^k(\beta)}[\beta].$$

Finally, using this, we have to find the solution of

$$\begin{aligned}
 \mathbf{x}_i &= \frac{\mathbf{z}_i(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}(\mathbf{x}_{i+1} + \mathbf{x}_{i+2}) + \frac{\mathbf{z}_{i+3}(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}\mathbf{x}_{i+3} + \frac{\mathbf{z}_{i+4}(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}\mathbf{x}_{i+4} \\
 (A.9) \quad &+ \frac{E_{\mathbf{q}^k(\beta)}[\beta]}{\mathbf{r}_i(\mathbf{u}^k)}[[\mathbf{H}^t(\mathbf{y} - \mathbf{H}\mathbf{x})]_i + \mathbf{x}_i],
 \end{aligned}$$

from which the GD iteration is obtained, that is,

$$\begin{aligned}
 \mathbf{x}_i^{k+1} &= \frac{\mathbf{z}_i(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}(\mathbf{x}_{i+1}^k + \mathbf{x}_{i+2}^k) + \frac{\mathbf{z}_{i+3}(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}\mathbf{x}_{i+3}^k + \frac{\mathbf{z}_{i+4}(\mathbf{u}^k)}{\mathbf{r}_i(\mathbf{u}^k)}\mathbf{x}_{i+4}^k \\
 (A.10) \quad &+ \frac{E_{\mathbf{q}^k(\beta)}[\beta]}{\mathbf{r}_i(\mathbf{u}^k)}[[\mathbf{H}^t(\mathbf{y} - \mathbf{H}\mathbf{x}^k)]_i + \mathbf{x}_i^k], \quad \forall i
 \end{aligned}$$

Alternatively, a CG method can be applied. In our experiments we used the basic CG version shown in [185] to solve Eq. (A.1). Note that several methods can be used (see, for instance, [52] [239] [57]) to calculate the TV image estimate without the use of the majorization of the TV prior.

## APPENDIX B

### **Calculation of required expected values in Algorithm 1**

In this section we show how the calculations of  $\mathbf{u}_i^{k+1}$  in Eq. (3.36) and  $E_{\mathbf{q}^{k+1}(\beta)}[\beta]$  in Eq. (3.41) are carried out. We first expand Eq. (3.36) to obtain

$$(B.1) \quad \begin{aligned} E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2] &= (\Delta_i^h(E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2 + ((\Delta_i^v(E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2 \\ &+ E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2] + E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^v(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2]. \end{aligned}$$

For Eq. (3.41) we have

$$(B.2) \quad E_{\mathbf{q}^k(\mathbf{x})} [\|\mathbf{y} - \mathbf{Hx}\|^2] = \|\mathbf{y} - \mathbf{H}E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]\|^2 + \text{trace} \left( \text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}] \mathbf{H}' \mathbf{H} \right).$$

Therefore,  $\text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]$  is explicitly needed to calculate these quantities. However, since the calculation of  $\text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]$  is very intense, we propose the following approximation of the covariance matrix. We first approximate  $W(\mathbf{u}^k)$  using

$$(B.3) \quad W(\mathbf{u}^k) \approx z(\mathbf{u}^k) \mathbf{I},$$

where  $z(\mathbf{u}^k)$  is calculated as the mean value of the diagonal values in  $W(\mathbf{u}^k)$ , that is,

$$(B.4) \quad z(\mathbf{u}^k) = \frac{1}{N} \sum_i \frac{1}{\sqrt{u_i^k}}.$$

We then approximate  $\text{cov}_{\mathbf{Q}^k(\mathbf{x})}$  using

$$\begin{aligned} \text{cov}_{\mathbf{Q}^k(\mathbf{x})} &\approx \left( E_{\mathbf{Q}^k(\beta)}[\beta] \mathbf{H}^t \mathbf{H} + E_{\mathbf{Q}^k(\alpha)}[\alpha] z(\mathbf{u}^k) (\Delta^h)^t (\Delta^h) + E_{\mathbf{Q}^k(\alpha)}[\alpha] z(\mathbf{u}^k) (\Delta^v)^t (\Delta^v) \right)^{-1} \\ (B.5) \quad &= \mathbf{B}^{-1}. \end{aligned}$$

Note that the matrix  $\mathbf{B}$  is a block circulant matrix with circulant blocks (BCCB), thus, computing its inverse can be performed in Fourier domain, which is very efficient [121].

Using this approximation, the last two terms in Eq. (B.1) can be expressed as

$$\begin{aligned} E_{\mathbf{Q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x} - E_{\mathbf{Q}^k(\mathbf{x})}[\mathbf{x}]))^2] + E_{\mathbf{Q}^k(\mathbf{x})}[(\Delta_i^v(\mathbf{x} - E_{\mathbf{Q}^k(\mathbf{x})}[\mathbf{x}]))^2] \\ (B.6) \quad \approx \frac{1}{N} \text{trace} \left[ \mathbf{B}^{-1} \times \left( (\Delta^h)^t (\Delta^h) + (\Delta^v)^t (\Delta^v) \right) \right]. \end{aligned}$$

Finally we can approximate the last term in Eq. (B.2) as follows:

$$(B.7) \quad \text{trace}[\text{cov}_{\mathbf{Q}^k(\mathbf{x})}[\mathbf{x}] \mathbf{H}^t \mathbf{H}] \approx \text{trace}[\mathbf{B}^{-1} \mathbf{H}^t \mathbf{H}].$$

## APPENDIX C

### Calculation of required expected values in Algorithm 5

In this section we show how the calculations of  $\mathbf{u}_i^{k+1}$ ,  $\text{cov}^k(\mathbf{h})$ , and  $E_{\mathbf{q}^k(\mathbf{x})}[\|\mathbf{y} - \mathbf{Hx}\|^2]$  in Chapter 5 are carried out. We first expand Eq. (5.23) to obtain the following expression for  $\mathbf{u}_i^{k+1}$

$$(C.1) \quad \begin{aligned} \mathbf{u}_i^{k+1} = E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2] &= (\Delta_i^h(E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}])))^2 + ((\Delta_i^v(E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}])))^2 \\ &+ E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2] + E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^v(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2], \end{aligned}$$

where

$$(C.2) \quad \begin{aligned} E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^h(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2] + E_{\mathbf{q}^k(\mathbf{x})}[(\Delta_i^v(\mathbf{x} - E_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]))^2] \\ = \frac{1}{N} \text{trace} \left[ \text{cov}(\mathbf{x}) \times \left( (\Delta^h)^t(\Delta^h) + (\Delta^v)^t(\Delta^v) \right) \right]. \end{aligned}$$

Note that  $\text{cov}_{\mathbf{q}^k(\mathbf{x})}[\mathbf{x}]$  is explicitly needed to calculate the quantities  $\mathbf{u}_i^{k+1}$ ,  $\text{cov}^k(\mathbf{h})$ , and  $E_{\mathbf{q}^k(\mathbf{x})}[\|\mathbf{y} - \mathbf{Hx}\|^2]$  in Eqs. (5.23), (5.20), and (5.27), respectively. However, calculating this matrix is computationally very inefficient, since it requires the inversion of an  $N \times N$  matrix. We utilize an approximation to this inverse which is proposed for the image restoration problem in [15] (explained in Chapter 3), where  $W(\mathbf{u}^k)$  in Eq. (5.17) is replaced by  $z(\mathbf{u}^k)\mathbf{I}$  with  $z(\mathbf{u}^k)$

being the mean value of the diagonal values in  $W(\mathbf{u}^k)$ . Specifically,

$$\begin{aligned} \text{cov}_{\mathbf{q}^k(\mathbf{x})} &\approx \left( \beta^k \mathbf{H}^t \mathbf{H} + \alpha_{\text{im}}^k z(\mathbf{u}^k) (\Delta^h)^t (\Delta^h) + \alpha_{\text{im}}^k z(\mathbf{u}^k) (\Delta^v)^t (\Delta^v) \right)^{-1} \\ (C.3) \quad &= \mathbf{B}^{-1}. \end{aligned}$$

With this approximation the matrix  $\mathbf{B}$  becomes block circulant with circulant blocks (BCCB), thus, computing its inverse can be performed in Fourier domain, which is very efficient [121]. We therefore replace  $\text{cov}_{\mathbf{q}^k(\mathbf{x})}$  by  $\mathbf{B}^{-1}$  where its explicit calculation is needed, i.e., in Eqs. (C.2), (5.20) and (5.27).