ADVANCES IN SINGLE FRAME IMAGE RECOVERY

by

SAKINAH ALI PITCHAY

A thesis submitted to
University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
2013

Abstract

This thesis tackles a problem of recovering a high resolution image from a single compressed frame. A new image-prior that is devised based on Pearson type VII density is integrated with a Markov Random Field model which has desirable robustness properties. A fully automated hyper-parameter estimation procedure for this approach is developed, which makes it advantageous in comparison with alternatives. Although this recovery algorithm is very simple to implement, it achieves statistically significant improvements over previous results in under-determined problem settings, and it is able to recover images that contain texture.

This advancement opens up the opportunities for several potential extensions, of which we pursue two: (i) Most of previous work does not consider any specific extra information to recover the signal. Thus, this thesis exploits the similarity between the signal of interest and a consecutive motionless frame to address this problem. Additional information of similarity that is available is incorporated into a probabilistic image-prior based on the Pearson type VII Markov Random Field model. Results on both synthetic and real data of Magnetic Resonance Imaging (MRI) images demonstrate the effectiveness of our method in both compressed setting and classical super-resolution experiments. (ii) This thesis also presents a multi-task approach for signal recovery by sharing higher-level hyper-parameters which do not relate directly to the actual content of the signals of interest but only to their statistical characteristics. Our approach leads to a very simple model and algorithm that can be used to simultaneously recover multiple natural images with unrelated content. The advantages of this approach in relation to state-of-the-art multi-task compressed sensing are investigated and findings are discussed.

Dari Malaysia merantau ke negara luar,

Dengan harapan mempelajari ilmu,

Harapan di hati sangat besar,

Alhamdulillah lahirnya falsafah 'buku'.

ACKNOWLEDGEMENTS

Writing a dissertation and the entire journey can be challenging without mental and physical support from numerous people. First, I would like to express my deepest appreciation to Dr. Ata Kabán for being a dedicated supervisor and for her continuously guidance throughout my journey. She deserved my greatest gratitude as without her insightful advice and constructive comments, I would be nowhere. Secondly, thank you to my thesis committee members for their fruitful suggestions and comments.

I would also like to acknowledge my sponsor, Ministry of Higher Education, Malaysia and Universiti Sains Islam Malaysia (USIM) for turning this dream into a reality.

I would like to heartfully thank my family in Malaysia especially to my beloved parents, Hj Ali Pitchay and Hjh Zubaidah for their unfailing support and encouraged me to do my best. I am grateful to my siblings Shahriman, Shazelin and Sharina for their emotional support and understanding during my difficult times.

Special thanks to my friends and researchers in the School of Computer Science who have made my graduate school life more enjoyable. Warm thanks go to Sarah, Khulood, Catherine, Shehnila, Jo and Bob for sharing their knowledge and opinions.

As an international student, the life can be tough and I was blessed with invaluable experiences and loving characters, whom without them, I would not have been able to cope some of the difficult situations. Thank you to my dear friends, Nur Hana, Eti Fairudz, Sahana, Irma, K Yatie, K Sally, K Azah. There will always be a place for them in my heart.

Finally, thanks to all my colleagues and friends whom I did not mention here, for their endless motivation and encouragement throughout this journey.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The aim of this research is to recover a high resolution (HR) version of one dimensional (1D) (e.g: wave, spectra, spike signals) and two dimensional (2D) signals (e.g: image) from a compressed or a low resolutin (LR) frame. The LR version is always contaminated with some additive noise. This is a highly under-determined problem where the number of observations in LR is much smaller than the number of unknowns (the HR pixels). Less observations in LR data makes this problem difficult to be solved, because even if the transformation is linear and is known, there are infinitely many high resolution signals that are all compatible with the LR data. That is, the data alone cannot distinguish which one of these is a meaningful signal. For this reason, the problem is ill-posed. To overcome this problem, we need to specify additional information about the structure of the high resolution signals of interest. Figure 1.1 illustrates an example of the problem description. The following are the issues with existing approaches that motivated our research:

- Most capturing processes introduce additive noise. Yet, most compressed sensing algorithms assumed a noiseless setting.
- The transformation from the original unknown high resolution image to the observed low resolution image degrades the information content of the image. A naive upsampling is therefore inadequate to recover the full details of the high resolution image. However, the more advance methods that use some former prior-knowledge

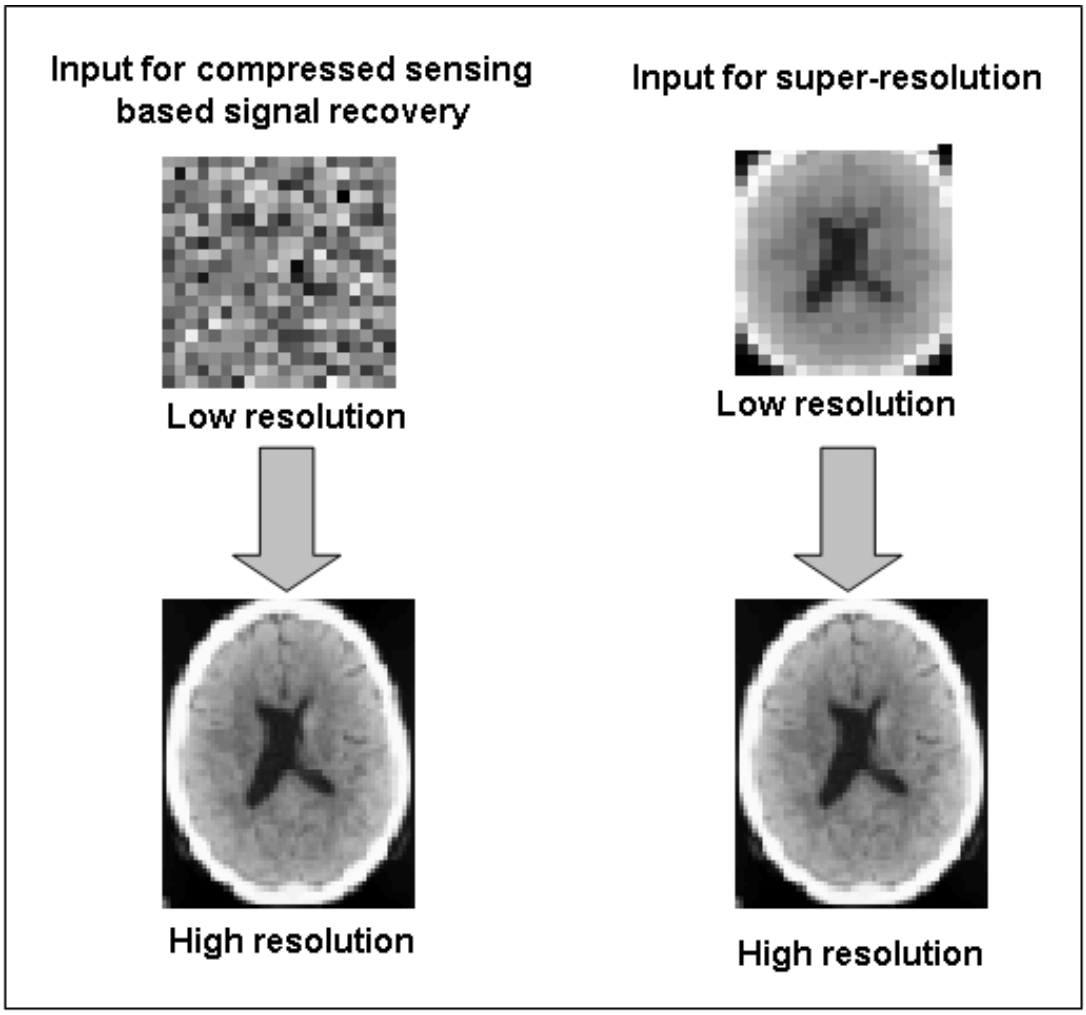


Figure 1.1: Examples of image recovery task from a single-frame of low resolution. The observed frame is compressed or has a low resolution and we try to recover its high resolution version. The left low resolution frame is generated using a random matrix with independent and identically distributed ($i.i.d$) standard Gaussian entries and the right frame is generated using the blur and down-sampling transformations. The high-resolution image is taken from the Matlab image database.

have various tuning parameters that are hard to set.

- Edges are generally hard to recover in high resolution because they have high information content that is hard to predict from the rest of the image.

This research develops novel alogorithms to recover the high resolution image, removes the additive noise and deals with the outliers simultaneously. To achieve this aim, we will formulate a flexible parameterised image-prior in a probabilistic framework that could be seen as a generalisation and extension of several existing approaches. A flexibility is achieved by automated estimation of higher level hyper-parameters.

1.1 Research Questions

This thesis is concerned with four main research questions; the first three questions are focusing on a single task (ST) recovery from a single low resolution frame and the final one addresses multi-task (MT) recovery.

- How to preserve the edges on the image when recovering the high resolution version of the image?

  In particular, different images contain different amounts of texture – How to devise a generic algorithm that is tuned automatically to the right propotion of texture versus smoothness? To solve this issue, we will devise a new flexible image model.

- How to estimate the parameters and hyper-parameters in the obtained model that is now non-convex and has many local optima?

- Intituitively any recovery methods should work better if it is given more specific extra information. Can we use available extra information to help recovering the high resolution image from fewer measurements? How to incorporate such extra information into our image-prior?

- Can we use a notion of statistical similarity, as described by our higher level parameters, to efficiently recover multiple high resolution images simultaneously that do not share semantic content?

Before proceeding to study these questions, we carried out an analysis of the natural image statistics to help us devise the new image-prior. Through study, this signal recovery application required the author to deal with several challenges.

1.2 Challenges

Development of the signal recovery algorithm is a challenging problem because it has to handle the following:

- An ill-posed condition - This generally means the solution does not exist or it is not unique, yielding highly noise sensitive solutions. In our case, it means the problem has too few observations, (measurements) on the low resolution frame because the high resolution image has under gone down-sampling, blurring and additive noise. Due to fewer pixels in the low resolution frame than in the high resolution image that we wish to recover, the single-frame version is under-determined. This makes the problem more challenging than classical super-resolution where several low resolution frames are available.

- Requirement of robustness - Since the problem is ill-posed, we will have to use an image-prior. It is crucial that this prior specifies correct information about the statistical characteristics of the high resolution image that we want to recover. There is a need to improve the existing prior-knowledge that is used in the existing state-of-the-art in terms of robustness and flexibility. Robustness means that the data contains outliers which corresponds to the edges on the image. Flexibility refers to the fact that we do not know beforehand the proportion of edges in a particular image. We want our approach to adapt itself to require level of texture within an

image.

- Sparsity property - A vector has the sparsity property when its elements are mostly zeros. For example, $v = [v_1, v_2, ..., v_i, ..., v_N]$ is sparse if many index of $v_i = 0$.

  The vector of neighbourhood features has many elements close to zeros because the intensity of many pixels is close to the average intensity of its neighbouring pixels. This is so because most images are locally smooth. It is a charatericstic of natural images, hence a good image-prior need to reflect this. Therefore, vector of the neighbourhood feature is almost a sparse vector. Smoothness or sparsity is over emphasised in compressed sensing Bayes method and as a result, the texture on the image cannot be recovered. In super-resolution approaches, the amount of free parameters is prohibitive.

- No universally accepted image model - There is no ideal image model that can both impose smoothness and preserve the edges in the image. This has motivated us to formulate an image-prior which covers those attributes and is appropriate for many natural images.

Once we are able to reconcile all of these requirements, the road will be open for many potential extensions and developements, as we shall see later.

1.3 Thesis Contributions

This thesis makes four significant contributions in the field of image recovery:

- Devising and formulating a novel robust image-prior which is capable of capturing the statistics of natural images using a probabilistic model based framework which allows a flexible approach. This model allows for the level of smoothness in the neighbourhood features to be estimated automatically. This is in contrast to previous methods which either fixed the hyper-parameters or required the user to set

them. Our new image-prior has been tested and compared with the existing methods and the results have been documented in conference and journal **Publications** in [**4**] and [**5**].

- A novel algorithm for single-frame image recovery that enables us to recover images with more texture than is currently possible using state-of-the-art methods. The performance of the proposed algorithm from our image recovery framework has demonstrated that our approach is superior to the state-of-the-art. This has been published in **Publications** [**4**] and [**5**].

- A similarity-prior is formulated and employed to include the similarity information between the scene of interest and a consecutive scene that differs in colouring or lighting. This prior enables a better accuracy from fewer measurements than a general-purpose prior would, and enables us to solve very under-determined problems. The results are published in **Publications** [**2**] and [**3**].

- A new approach to multi-task signal recovery is devised where the target signals need not have any overlap in their content but only share their higher level statistical characteristics. This can be used for simultaneous recovery of sets of natural images in a single run. Results and comparisons are presented in **Publication** [**1**].

1.3.1 Publications

Publications arising from the thesis are listed as follow:

- [**1**] S.A Pitchay and A Kabán. Multi-task Signal Recovery by Higher Level Hyperparameter Sharing. *Proc. 211st of International Conference on Pattern Recognition* (ICPR'2012), Tsukuba, Japan, 11-15 November 2012, IEEE Computer Press, pp. 2246-2249. (oral presentation - 15% acceptance rate)

- [**2**] S. Ali-Pitchay and A Kabán. Single-frame Signal Recovery Using a Similarity-Prior Based on Pearson type VII MRF. *In Proceedings of the 1st International*

Conference on Pattern Recognition Applications and Methods, ICPRAM'2012, pp. 123-133, Sci-TePress. (full paper, oral presentation - 24% acceptance rate)

- **[3]** S. Ali-Pitchay and A Kabán. Single-frame Signal Recovery Using a Similarity-Prior. *Springer Proceedings in Mathematics & Statistics 30*, Mathematical Methodologies in Pattern Recognition and Machine Learning, vol. 30, 2013, pp. 83-98, (c) Springer, (invited extension of ICPRAM'2012 selected contribution).

- **[4]** A Kabán and S.A Pitchay. Single-frame Image Recovery using a Pearson type VII MRF. *Neurocomputing*, 80: 111-119, 2012, (invited and referreed extension of MLSP'2010 paper, 14 in print for the special issue out of 78 papers.)

- **[5]** A Kabán and S.A Pitchay. Single-frame Image Superresolution using a Pearson type VII MRF. *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, MLSP, pp. 29-34, August 29 - September 1, 2010, Kittila, Finland, (oral presentation - 30% acceptance rate[1]).

- **[6]** S.A Pitchay. Non-linear Image Recovery from a Single Frame Super Resolution Using Pearson Type VII Density. *Intelligent Automation and Systems Engineering,* pp. 295-307, Lecture Notes in Electrical Engineering 103, (c) Springer, 2011.

1.4 Overview of the Thesis

This thesis has seven chapters. The remainder of this thesis is structured as follows.

Chapter 2: Introduction to Signal Recovery

- Chapter 2 presents an introduction to signal recovery and super-resolution approaches. The chapter also discusses the inverse problem and introduces the probabilistic model-based methodology for solving such problems that we will build upon later. Several existing image-priors are reviewed and the strength and weaknesses are discussed.

[1]Acceptance rates can be found at http://www.sciencedirect.com/science/article/pii/S0925231211005947

Chapter 3: Single-frame Image Recovery using a Pearson type VII MRF

- Chapter 3 proposes the novel image-prior, the Pearson type VII Markov Random Field, and provides details on constructing the main building blocks of this image-prior. This chapter outlines the overall image recovery framework from a single-frame. It also develops the signal recovery algorithm associated with the model and is based on **Publications** [**4**] and [**5**] as mentioned in section 1.3.1.

Chapter 4: Investigating Alternative Hyper-parameters Estimation Approaches

- Chapter 4 formulates and develops the multivariate Pearson type VII that acts on the entire image. We compare its performance with the univariate Pearson type VII which acts on the pixel level and existing image-priors. This chapter also investigates alternative approaches for estimating the hyper-parameters and demonstrates the efficiency of this new approach through extensive experiments. This is based on **Publication** [**6**] in section 1.3.1.

Chapter 5: Single-frame Signal Recovery Using a Similarity-Prior

- Chapter 5 considers specific extra information which we found to be useful for recovering the high resolution scene of interest and devises the similiarity-prior that incorporates the information into the Pearson type VII density model. The extra information consists of a notion of similarity between high resolution images that differ in colouring or lighting. An application to MRI images is presented where this method is shown to greatly reduce the number of measurements needed for a good recovery. This chapter is based on **Publications** [**2**] and [**3**] in section 1.3.1.

Chapter 6: Multi-task Recovery without Content Similarity

- Chapter 6 extends the work of single-task recovery into multi-task recovery by sharing the hyper-parameters. The chapter provides extensive results that are additional to **Publication** [**1**] in section 1.3.1.

Chapter 7: Summary and Conclusions

- Finally chapter 7 concludes the thesis by summing up the achievements and listing possible future avenues for further investigation.

CHAPTER 2

INTRODUCTION TO SIGNAL RECOVERY

This chapter assumes no expert knowledge in signal recovery and therefore a very brief introduction to signal recovery versus super-resolution approach is provided. This chapter will first differentiate and relate the signal recovery and the super-resolution task in section 2.1. Section 2.2 gives an overview of the sorts of problems when the additional information is required. Section 2.3 describes briefly some of the parameter estimation methods that will be utilised in the development. Next, in section 2.4 introduces a statistical model, Markov Random Fields, and describes an example construction of a neighbourhood matrix that will be utilised later in the additional information based on this model. This section also studies the types of the distribution of the neighbourhood features. A number of state-of-the-art image-priors are reviewed in section 2.5. Section 2.6 discusses the inadequacies of the previous work. Finally 2.7 summarises the chapter and raises the issue that motivated the author on a further investigation.

2.1 Signal Recovery versus Super-resolution

The basic idea behind super-resolution (SR) [5] is to combine the information from multiple low resolution frames to generate a high resolution image as shown in figure 2.1. These frames contain non-redundant information [5] typically because of subpixel shifts between them. In this case, the more details there are in the image the better. If we

have sufficient number of low resolution frames, the problem is rather easy to solve as the data contains all the necessary information for recovery. On the other hand, this thesis focuses on recovering the signals from a single low resolution noisy frame. This is what we referred to as the signal recovery problem.

Signal recovery is studied in this work in two instances: (i) Recovery from a randomly compressed version, as in Compressed Sensing [6] (ii) Recovery from a sub-sampling and blur as in the classical, but single-frame super-resolution setting. In both instances, we only consider linear transformation models, which is indeed the form used in the state-of-the-art in the literature since it is sufficient for many applications, and keeps the problem manageable. In the field of compressed sensing [7] (or also know as compressive sensing), the random matrix that compressed the high resolution image is also called the compressive matrix.

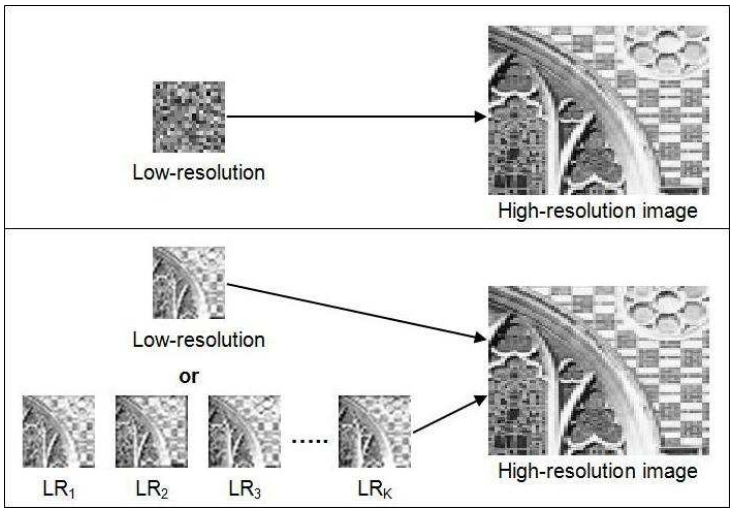


Figure 2.1: Examples of the low resolution input for a signal recovery and super-resolution application. The signal recovery shown in the top half of the figure utilises the compressive matrices. In the second half of the figure refers to the super-resolution where the LR frames are generated by utilising the transformation[1]operator.

Before we begin to outline the existing methods to recover the signals, an illustrative example for both of these image observation models is presented. The general form of the image observation model can be written as in equations (2.1) or (2.2) where $\boldsymbol{y}$ is the low resolution frame, $\boldsymbol{W}$ is the transformation operator that acts on the high resolution

[1]http://www.robots.ox.ac.uk/∼vgg/research/SR/synthdata.html

image $\boldsymbol{z}$, and η is the additive noise of M-dimensional vector with $i.i.d$ zero mean spherical Gaussian elements of variance σ^2.

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{z} + \eta \tag{2.1}$$

$$\boldsymbol{y}_k = \underbrace{\boldsymbol{D}_k\boldsymbol{B}_k\boldsymbol{M}_k}_{\boldsymbol{W}_k}\boldsymbol{z} + \eta_k, \ \forall \ k{=}1{,}2{,}...{,}K \tag{2.2}$$

where $\boldsymbol{D}_k$ encodes the down-sampling operator for the k-th low resolution frame, $\boldsymbol{B}_k$ models the blurring effects, $\boldsymbol{M}_k$ is the motion information for the k-th frame and η_k is the noise term. As already mentioned, in classical multi-frame super-resolution, K is bigger than one. But in this thesis, K is equal to one. Therefore these two equations (2.1) and (2.2) have the same form but differ in the structure of $\boldsymbol{W}$.

There are many theoretical works [2, 3, 6] in the literature of Compressed Sensing (CS) that guarantee the recovery if the transformation $\boldsymbol{W}$ is a certain random matrix (for example, i.i.d Gaussian entries) and $\boldsymbol{z}$ is sparse. However that theory does not fit the other, more structured and deterministic $\boldsymbol{W}$ in equation (2.2). Although the theory of CS is outside our scope, in this thesis we will utilise both of these $\boldsymbol{W}$ and will employ a probabilistic model based approach as opposed to most CS algorithms (e.g. the ℓ1-magic package[1], based in ℓ1-regularisation). As a byproduct we will gain some insights into whether the type of $\boldsymbol{W}$ makes any major difference to the ability of an algorithm to recover a good quality image.

Note, these linear systems in equations (2.1) or (2.2), when $K = 1$, are under-determined because the dimensionality of $\boldsymbol{y}$ is much smaller than the dimensionality of $\boldsymbol{z}$. Therefore, solving the system for $\boldsymbol{z}$ is ill-posed. The challenge is to come up with additional information that constrains the problem in the right way; and the procedure to do this will be pursued throughout the thesis.

We should mention that the matrix $\boldsymbol{W}$ may be partially unknown in certain applications and may need to be estimated. This can be done when $K >> 1$. Indeed, former

[1]The code can be found at http://users.ece.gatech.edu/∼justin/ℓ1magic/

works [8, 9, 10] and more recently Pickup [4] have already tackled this when there is sufficient data available. Since our scope is in the single-frame ($K = 1$) setting, we will consider $\boldsymbol{W}$ to be known, as in compressed sensing, and focus on recovering $\boldsymbol{z}$ from the under-determined system. The next section will introduce the reader to the problem that requires the additional information. Before we proceed further, we highlight some literature for the non-linear[1] problem.

2.2 Introduction to the Problem

We already disccused that conventional SR approaches require multiple low resolution frames. In reality, it is hard to find sufficient number of low resolution frames, so we are left to solve an under-determined system. To overcome this problem, some form of prior-knowledge is introduced to stabilize the inversion of the ill-posed system. A problem is ill-posed [13] when the system does not have the well-posed properties: (i) a solution exists, (ii) is unique and (iii) the solution depends continuously on the data.

The recovery approaches are divided into two major categories to stabilize the solution. One by numerically stabilising the solution and the second category by exploiting some additional information gained through building an image model (referred to as an image-prior). To numerically stabilise the solution, we can use the pseudo-inverse. From equation

[1]A non-linear model could be formulated as $\boldsymbol{y} = f(\boldsymbol{z}) + \eta$ where f($\boldsymbol{z}$) is some parameterised non-linear function of $\boldsymbol{z}$. However, signal recovery in non-linear compressed sensing (NLCS) literature is very scarce. Some of the works [11, 12] introduce the concept of non-linear measurements into CS theory. Xu *et al* [11] study sparse recovery by linearizing the equations and apply an iterative procedure to obtain a solution. Blumensath [12] shows in theory that sparse or structured singals from few non-linear observations are possible to be recovered under certain conditions.

(2.1), we can define an objective function (2.3) to find the most probable $\boldsymbol{z}$:

$$f(\boldsymbol{z}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z}||^2 \tag{2.3}$$

$$\begin{aligned}\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}} &= \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{1}{2}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z})\right) \\ &= \frac{1}{2}(-2\boldsymbol{W})^T(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z}) \\ &= -\boldsymbol{W}^T\boldsymbol{y} + \boldsymbol{W}^T\boldsymbol{W}\mathbf{z}\end{aligned} \tag{2.4}$$

By setting equation (2.4) to zero, the solution of $\boldsymbol{z}$ is as follows:

$$\begin{aligned}(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{z} - \boldsymbol{W}^T\boldsymbol{y}) &= 0 \\ \boldsymbol{W}^T\boldsymbol{W}\boldsymbol{z} &= \boldsymbol{W}^T\boldsymbol{y} \\ \boldsymbol{z} &= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{y}\end{aligned} \tag{2.5}$$

However the inverse of the $\boldsymbol{W}^T\boldsymbol{W}$ (written in equation (2.5)) does not exist because the determinant of this matrix is zero and the rank of this matrix is at most the dimensionality of $\boldsymbol{y}$ which is much more smaller than the dimensionality of $\boldsymbol{z}$. Therefore we could replace the inverse by the pseudo-inverse. Another possibility is to add a full rank matrix to $\boldsymbol{W}^T\boldsymbol{W}$ to make it invertible such as in equation (2.6).

$$\boldsymbol{z} = (\boldsymbol{W}^T\boldsymbol{W} + \underbrace{\sigma_{\boldsymbol{\eta}}^2\Omega}_{regularisation})^{-1}\boldsymbol{W}^T\boldsymbol{y} \tag{2.6}$$

where Ω is a full rank matrix, e.g. $N \times N$ identity matrix. Both of these ideas will give a unique solution, i.e stabilise numerically the solution of the original system. However, this solution is quite arbitrarily picked from the infinitely many solutions of the original system from equations (2.1) and (2.2).

For this reason, rather than choosing one of the above solutions, we should design how to stabilise the solution. We can do this by exploiting some information about the structure of high resolution images in general. For this, we will need to review building blocks

like Markov Random Fields and various probabilistic models that allow us to formulate prior-knowledge in a consistent framework and derives method for image recovery that exploits this framework.

The prior-knowledge will be built into an image-prior that plays an important role to specify what makes a solution to the unknown (i.e: pixels in the high resolution image) in general. An image-prior can help when (i) the number of samples is less than the number of unknowns, (ii) an accurate prior model exists and is required and (iii) accuracy without prior is poor. Before we proceed to formulise some image-prior, we will describe parameter estimation methods that will deal with those priors.

2.3 Parameter Estimation Methods

Extensive work in super-resolution and signal recovery methods employ probabilistic approaches and have been the main key to initiating modelling the data. Probabilistic models [14, 15] are a useful tool that allows us to infer data from the observed data based on Bayesian theory [16]. Maximum Likelihood and Maximum A Posteriori are the two estimators utilised in this approach. This section is restricted to methods of choosing parameter and hyper-parameter value that we will use in this thesis.

2.3.1 Maximum Likelihood (ML) estimation

Maximum Likelihood [17] estimation is the most widely used method for estimating the parameters of a statistical model. Examples of work that uses ML estimator can be found in [18, 19, 20, 21, 22, 23, 24]. The maximum-likelihood estimator is obtained by maximizing the log likelihood function. This is done in equation (2.7). In equation (2.7), $p(d|\Theta)$ denotes a parameterised probability model where d is the data and Θ are the

parameters, and we can maximise the log of $p(d|\Theta)$ with respect to Θ.

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \log p(d|\Theta) \tag{2.7}$$

Assume a particular form for the density (e.g: Gaussian, Poisson, Binomial), only the parameters such as the mean and variance need to be estimated. ML has the advantage of not requiring the expression of the prior distribution on the parameters. Nevertheless, overfitting turns out to be the disadvantage, if the number of data points are small [25]. This is equivalent to an ill-posed problem in image recovery where the quality of the recovered image becomes worse when there are infinitely many solutions. The factor [26] that contributes to this poor quality is that the image recovery suffers when the number of low resolution frames is small. To solve this problem, a solution is required to include some prior information. We can do this by adopting the Maximum A Posteriori estimator.

2.3.2 Maximum A Posteriori (MAP) estimation

In Maximum A Posteriori [25, 27] framework, it assumed a prior distribution for the parameters $p(\Theta)$ is available. Bayes' theorem shows the way for incorporating prior information in the estimation process:

$$p(\Theta|d) = \frac{p(d|\Theta)p(\Theta)}{p(d)} \tag{2.8}$$

The left hand side of the equation is called the posterior. The term on the right hand side is the numerator which is the product of the likelihood term and the prior term. The denominator serves as a normalization term so that the posterior probability density function (PDF) integrates to unity. Thus, Bayesian inference [28, 29] produces the maximum

a posteriori estimate:

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} \; p(d|\Theta)p(\Theta) \tag{2.9}$$

$$= \arg\max_{\Theta} \; \{\log p(d|\Theta) + \log p(\Theta)\} \tag{2.10}$$

When there is no available knowledge on Θ, this is equivalent assuming a non-informative prior or an improper prior [30]. For that assumption, equation (2.10) reduces to ML formulation. In our image recovery setting, MAP estimates is computed via numerical optimisation such as conjugate gradient which requires the derivative of its objective function. Many works [31, 32, 33, 34, 35, 24, 36, 37] have proposed to use MAP estimator in super-resolution and image enhancement area.

2.3.3 Cross validation (CV)

Cross validation [38] is a method for estimating the performance of a predictive model [39]. According to Arlot and Kei [40, 41], the main key behind cross validation is to divide data, once or several times, for estimating the risk of each model. Part of the data will be used to learn or train the model and the remaining part is used to validate the model. Finally, the CV method selects the smallest error. To estimate the hyper-parameters using this method is also challenging because we need to choose one from the three types of CV wisely. It can be categorized into simple CV or random subsampling, K-fold CV and leave-one-out CV(LOOCV) [41, 42]. CV method also can prevent the over-fitting problem because the training data is independent from the validation data [41].

Simple CV also known as hold-out validation [43], where a single data sample is divided into two groups, one for training and the remainder for testing. The advantage of hold-out method is that it is usually a preferred choice to the residual method and consumes less time to compute. However, its evaluation can have a large variance [40, 44]. For instance when the training data is small, we may get an unfortunate split which results in high error. Since this method is relies on a single division into training and test set, sometimes

we may get lucky split or the unlucky split. Therefore there is a large variability of the error. The evaluation relies on how the division is generated and therefore it could be remarkably different. Yet, most recent theoretical results on any diverse CV procedures are not accurate enough to differentiate which splitting strategy is the best [40].

In K-fold CV, the original sample is partitioned into K-samples randomly. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining K-1 subsamples are used as training data. The process is repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage [45] of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. Therefore, it provides an accurate performance estimation [46]. 10-fold cross-validation is commonly used, but in general K remains an unfixed parameter. A good choice of K depends on the dataset size.

LOOCV[46] is a special case of K-fold cross-validation where k equals the number of instances in the data. It involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This method is computationally expensive because it requires many repetitions of training. Algorithms for seeking the parameter and hyper-parameters using cross validation methods are presented in chapter 4.3.

In order to include the additional information that represents what the recovered image looks like, a construction on the neighbourhood matrix $\boldsymbol{D}$ is implemented. The following section describes a brief introduction to the Markov Random Fields which have been employed in the image-prior. A study on the distribution of the neighbourhood feature is conducted to analyse the type of the common distribution that is required for building the image-prior.

2.4 Introduction to Markov-Random Fields

Markov Random Fields (MRF) are kind of statistical model. It aims to model the local structure or the interactions among random variable in a particular set. An illustration of the neighbourhood system for MRF is shown in figure 2.2. The structure of the neighbourhood system determines the order of the MRF.

Figure 2.2: Neighbourhood system as function of Random Fields model order (right) and on the left is the 1^{st} order neighbourhood system. Adapted from:[1]

The 1^{st} order neighbourhood system is also called the 4-neighbourhood system. Every pixel has four neighbours except pixels that are on the top, bottom, the first left and the last right side of an image. A first-order MRF assumes that for any pixel i its intensity depends on the intensities of its closest cardinal neighbours but does not depends on any other pixel of the image as illustrated in figure 2.2. Cardinal neighbour means pixel that are nearby in the sense of location. The 2^{nd} order neighbourhood system is known as the 8-neighbourhood system.

MRF are identified as well suited and widely used models that are able to capture the local smoothness property. Therefore, in the image recovery case, local smoothness corresponds to the neighbourhood intensity of each pixel. MRF plays a significant role for modeling and estimation in a variety of fields within pattern recognition [47], object classification [48, 49], object matching [50], image segmentation[1] and image restoration [51].

We employ MRF as it allows us to built an image-model, taking advantage of what we know about the images of the kind we are looking for (i.e: prior-knowledge). It has the flexibility to exploit the prior-knowledge. There are some other image analysis method

z_1	z_4	z_7	z_{10}
z_2	z_5	z_8	z_{11}
z_3	z_6	z_9	z_{12}

Figure 2.3: An illustration of the $\boldsymbol{z}$ coordinate for size(3,4).

such as interpolation [52, 53] that do not build a model of the high resolution image. This method does not recover the images well compared to the methods [4] that employ prior-knowledge. An example recovery using bicubic interpolation is shown in figure 3.3.

In addition, by being a probabilistic model, the MRF can be combined with the noise model (i.e.: the likelihood model) to form our overall model of the SR problem. This overall model can be used to estimate the high resolution image using Bayes rule by maximising its posterior. Hence, the probabilistic framework of which the MRF of one element offers a principled framework to include both prior-knowledge about the unknown HR image as well as the characteristic of the observation noise and enables a principled estimation of the high resolution image using the MAP estimation technique.

2.4.1 Constructing the neighbourhood matrix

In this section, an algorithm of constructing the neighbourhood matrix $\boldsymbol{D}$ is described and an example of identifying the cardinal neighbours for each pixel $\boldsymbol{z}$ is presented. An image $\boldsymbol{z}$ of size (3,4) is illustrated in figure 2.3 and the coordinate of the neighbourhood $\boldsymbol{D}$ matrix is described in table 2.1. Each pixel of image is labelled from z_1 to z_{12}. Using the pixel location, the four cardinal neighbours are identified. As shown in table 2.1, the border of an image (e.g. z_1, z_2, z_3, z_4, z_6, z_7, z_9, z_{10}, z_{11} and z_{12}) can only consist of two or three cardinal neighbours. However for real data with higher dimensionality, there are more intensity values in $\boldsymbol{z}$ or the pixel that relies on the four neighbours compared to the pixels with two and three neighbours. Therefore, the algorithm neglects the minority of the pixels since it does not affect the majority difference. Table 2.2 represents the entries for all pixels in image ($\boldsymbol{z}$) according to the image size (3,4).

Table 2.1: An example of $\boldsymbol{z}(3,4)$ with its cardinal neighbours.

Pixel location	**Neighbours ID**
z_1	ID(2,4)
z_2	ID(1,3,5)
z_3	ID(2,6)
z_4	ID(1,5,7)
z_5	ID(2,4,6,8)
z_6	ID(3,5,9)
z_7	ID(4,8,10)
z_8	ID(5,7,9,11)
z_9	ID(6,8,12)
z_{10}	ID(7,11)
z_{11}	ID(8,10,12)
z_{12}	ID(9,11)

Figure 2.4 displays their cardinal neighbours for each pixel location of an image size (8,7). The red plot presents the pixel for each coordinate and the blue color indicates the primary neighbours according to the given algorithm. Following the steps in algorithm

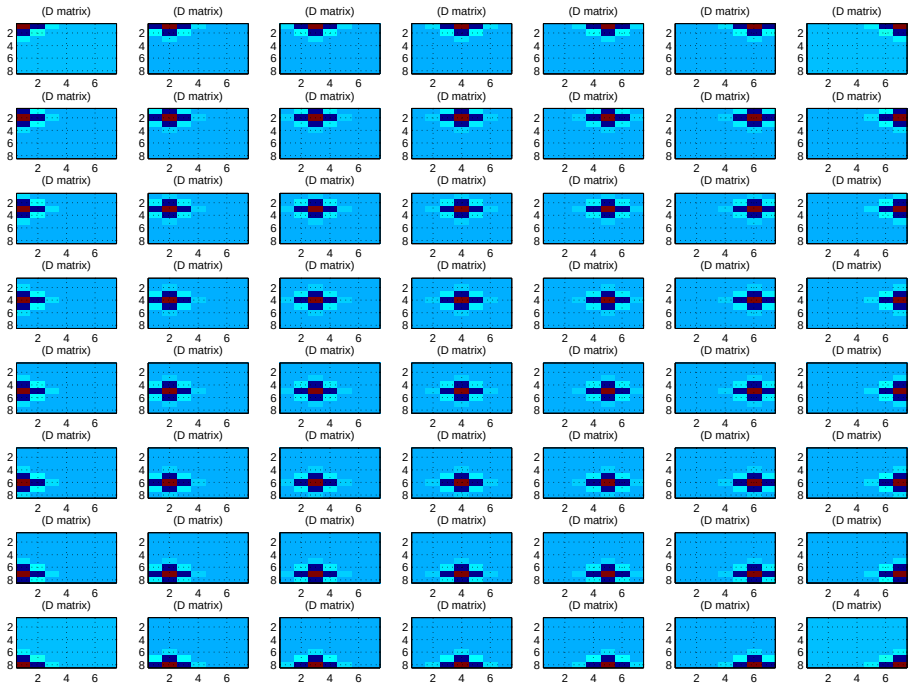


Figure 2.4: An illustration of the cardinal neighbours.

1, each entry in the constructed $\boldsymbol{D}$ matrix is now filled by the entries defined in equation

Table 2.2: The neighbourhood structure constructed for an image of size $\boldsymbol{z}(3\times4)$.

row/col	1	2	3	4	5	6	7	8	9	10	11	12
1	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0	0	0	0	0	0	0	0
2	$-\frac{1}{4}$	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0	0	0	0	0	0	0
3	0	$-\frac{1}{4}$	1	0	0	$-\frac{1}{4}$	0	0	0	0	0	0
4	$-\frac{1}{4}$	0	0	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0	0	0	0	0
5	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0	0	0	0
6	0	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1	0	0	$-\frac{1}{4}$	0	0	0
7	0	0	0	$-\frac{1}{4}$	0	0	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0	0
8	0	0	0	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1	$-\frac{1}{4}$	0	$-\frac{1}{4}$	0
9	0	0	0	0	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1	0	0	$-\frac{1}{4}$
10	0	0	0	0	0	0	$-\frac{1}{4}$	0	0	1	$-\frac{1}{4}$	0
11	0	0	0	0	0	0	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1	$-\frac{1}{4}$
12	0	0	0	0	0	0	0	0	$-\frac{1}{4}$	0	$-\frac{1}{4}$	1

(3.5).

Algorithm 1 : Constructing the neighbourhood matrix $\boldsymbol{D}$.

Step 1: Define the row and column of the recovered image $\boldsymbol{z}$
Step 2: Initialise the size of the neighbourhood matrix, row × column.
Step 3: Initialise the neighbourhood matrix as zeros entry.
for $i = 1$ **to** length(row) **do**
  for $j = 1$ **to** length(column) **do**
    Get the pixel ID
    Fill in D(ID,:) with entry 1 *if* $i = j$
    Obtain the ID's of the neighbor of the ij-the pixel of $\boldsymbol{z}$ using the following entries described in section 3.5.
  end for
end for

2.4.2 What is the distribution of neighbourhood features?

In this section, we build understanding of the distribution of neighbourhood features by looking at the histogram of neighbourhood features $\boldsymbol{D}_i\boldsymbol{z}$. $(\boldsymbol{D}_i\boldsymbol{z})$ means the difference between each pixel of $\boldsymbol{z}$ and its four nearest neighbours or it is equivalently in this mathematical form, $z_i - \frac{1}{4}\sum_{j\in 4neigb(i)} z_j$. The details for this form are written in Appendix A.

We investigate what the shape looks like and whether all the natural images exhibit the same pattern. Before we proceed further, we need to define what we mean by natural

images. The term itself is too subjective but examples include images of human beings, animals and plants. Images of some buildings also have the characteristic of the natural images. For that kind of images, if we look at the histogram of $\boldsymbol{D}_i\boldsymbol{z}$, we observe a heavy-tailed shape, a big peak at zero and almost symmetric on both sides. This big peak corresponds to the local smoothness of the data where it captures the homogeneous area in the image. Weiss and Freeman's [54] also claim that natural images have those properties.

We visualised the histogram of neighbourhood features for many images and we discovered that most of the tested natural images have a common property. The resulting histogram has a high peak at zero, almost symmetric and long tails. This histogram shape is likely to fit a heavy-tail density that we will propose later in Chapter 3 of this thesis. The high peak means it has a high probability of zeros and near-zeros, which represent many areas of local smoothness in the image. At the same time, the long tails allow outliers that correspond to the edges. From the observed data, we investigate and instantiate the functional form of the probability densities that describe the shape of the likely values of these features later in the thesis. Figure 2.5 shows a few examples of observed histograms of these features, from natural images.

In this study, some peculiar histogram are visualised in section 4.5.2 in figure 4.14. Those images represent examples where the images contain a lot of texture and a texture-based prior is more preferable in this case. Therefore, the proposed image-prior is not suitable for those kinds of image recovery.

2.5 Existing Image-priors

Throughout this section, relevant image-priors to this field are presented.

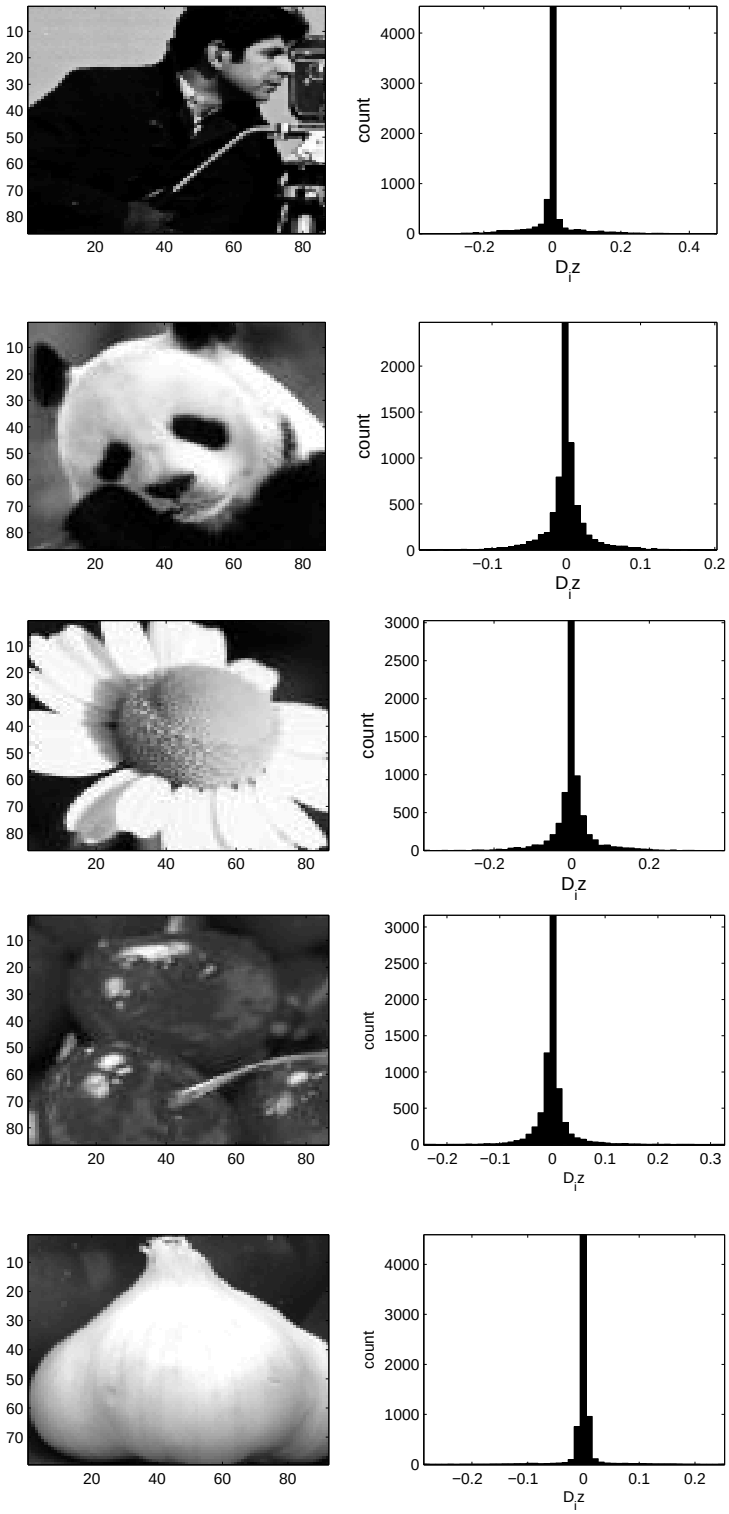


Figure 2.5: Examples of histograms of the distribution of neighbourhood features $\boldsymbol{D}_i \boldsymbol{z}, i = 1, \cdots, N$ from natural images.

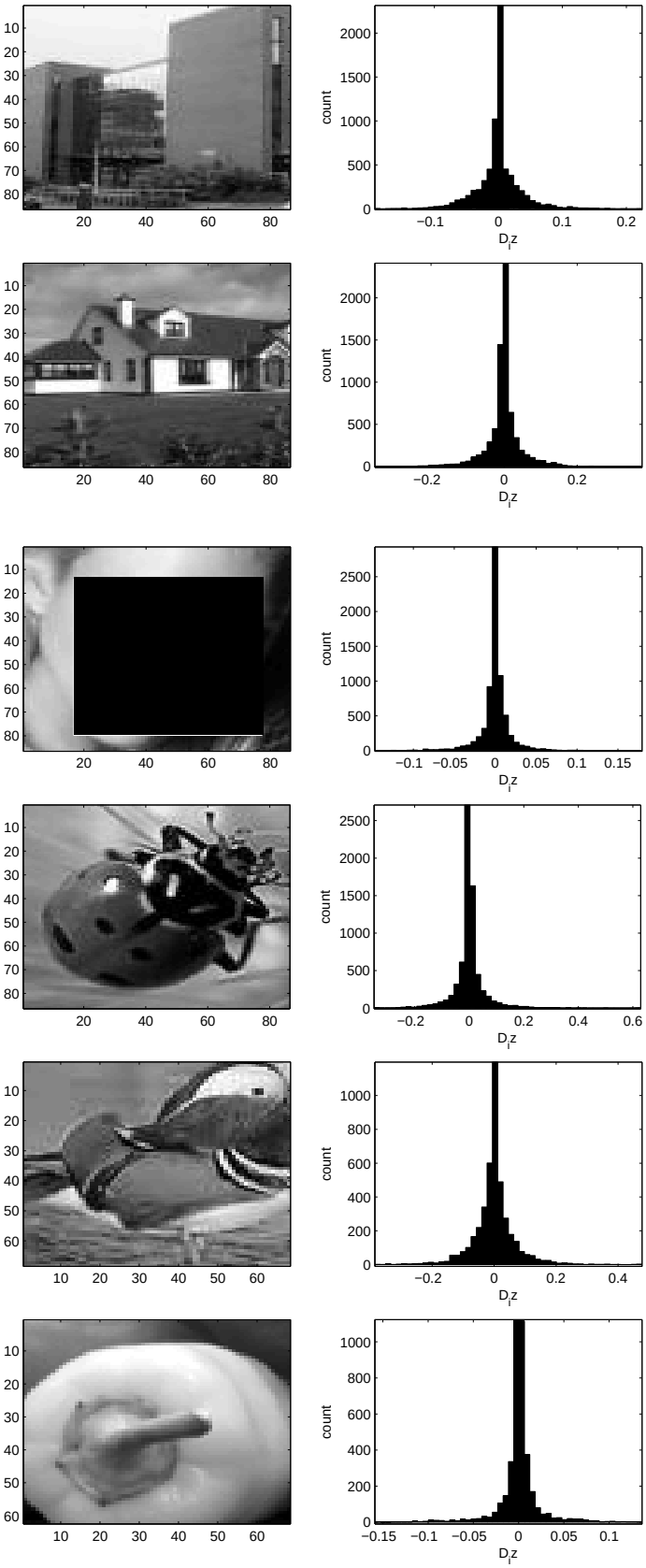


Figure 2.6: Examples of histograms of the distribution of neighbourhood features $\boldsymbol{D}_i \boldsymbol{z}, i = 1, \cdots, N$ from natural images.

2.5.1 Gaussian MRF

The Gaussian Markov Random Field (GMRF) is extensively used in statistical modelling [55, 56, 57, 58] and most widely used as an image-prior density in [32, 59, 60, 61, 62, 63, 64]. It has the following form [32]:

$$Pr(\mathbf{z}) \propto \prod_{i=1}^{N} \exp\left\{-\frac{1}{2\lambda}(\boldsymbol{D}_i \boldsymbol{z})^2\right\} \tag{2.11}$$

$$= \exp\left\{-\frac{1}{2\lambda}\sum_{i=1}^{N}(\boldsymbol{D}_i \boldsymbol{z})^2\right\} \tag{2.12}$$

$$= \exp\left\{-\frac{1}{2\lambda}\boldsymbol{z}^T \mathbf{D}^T \boldsymbol{D}\mathbf{z}\right\} \tag{2.13}$$

where λ is the variance parameter and $\propto$ denotes proportionality. Comparing the latter expression to that of a multivariate Gaussian density we can read off the covariance matrix induced by the employed neighbourhood definition:

$$Cov = (\boldsymbol{D}^T \boldsymbol{D})^{-1} \tag{2.14}$$

Figure 2.7 illustrates the negative log Gaussian prior with several values of hyper-parameter λ. The most horizontal with the strongest curve plots the larger λ and it is far from the solution. The horizontal line denotes smaller value of λ are more appropriate to obtain a better solution. Despite the GMRF image-prior has many advantages including its unique solution [65], it always tends to smooth and penalise the sharp edges that we wish to recover. GMRF does not model edges well and in some applications such as in image restoration, it blurs the edges and leaves excessive amounts of noise.

2.5.2 Huber MRF

Huber Markov Random Field (HMRF) image-prior is studied in [36, 66, 67, 68, 31]. Huber density is defined with the aid of the Huber function as in equation (2.15). The Huber function is quadratic around the centre, and linear in the tails, with no gradient

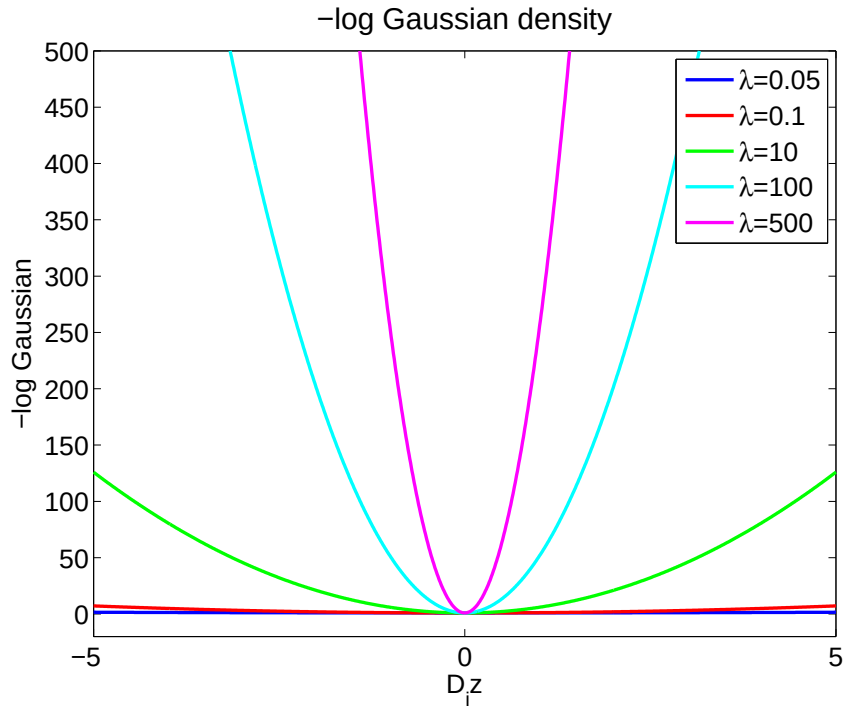


Figure 2.7: Illustration of Gaussian plot of 1D density for five values of λ.

discontinuity. It takes a threshold parameter δ, specifying the value at which it diverts from being quadratic to being linear. If the threshold δ is large, then the HMRF prior reduces to a GMRF image-prior. On the other hand, if $\delta \to 0$, the Huber prior is equivalent to the total variation prior. A generic variable u in the definition of this function is used and will be instantiated later as a neighbourhood-feature in chapter 3.

$$H(u|\delta) = \begin{cases} u^2, & \text{if } |u| < \delta \\ 2\delta|u| - \delta^2, & \text{otherwise.} \end{cases} \tag{2.15}$$

The Huber-MRF prior is then defined as:

$$Pr(\boldsymbol{z}) \propto \prod_{i=1}^{N} \exp\left\{-\frac{1}{2\lambda} H(\boldsymbol{D}_i \boldsymbol{z}|\delta)\right\} \tag{2.16}$$

$$= \exp\left\{-\frac{1}{2\lambda} \sum_{i=1}^{N} H(\boldsymbol{D}_i \boldsymbol{z}|\delta)\right\} \tag{2.17}$$

where λ is similar to a variance parameter and acts as the tuning parameter to control the smoothness of the regions data. Figure 2.8 plots the Huber function for several different

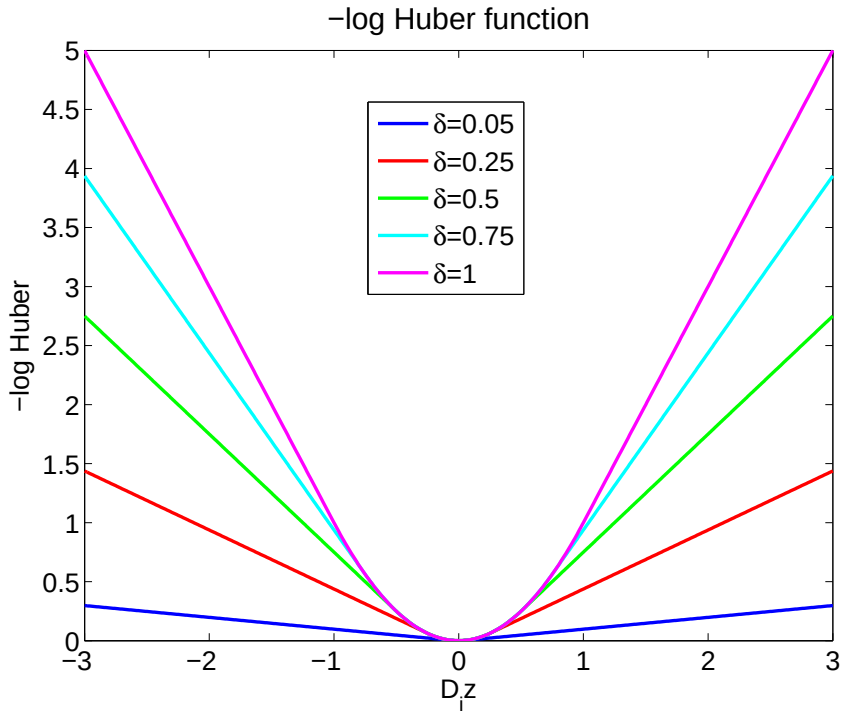


Figure 2.8: 1D Huber function plot for five values of δ.

values of the threshold, δ. This function is then applied to the Huber image-prior as the edge penalty term in a distribution over the image edges. A major disadvantage of using this prior is that it requires the choice of an edge threshold. Several examples of the previous works [69, 33, 70, 21, 71, 72, 37] on estimating this parameter are discussed in section 2.6.

2.5.3 Heavy-tail prior in Bayesian Compressed Sensing (BCS)

BCS is a recent state-of-the-art method for image recovery method in the compressed sensing literature. Ji *et al.* [2] compare their method with the existing work (ℓ1-magic [73]) and demonstrated superior results. They used the following heavy-tail prior as

defined in equation (2.18).

$$p(\boldsymbol{\theta}|a,b) = \prod_{i=1}^{N} \int_0^{\infty} \mathcal{N}(\theta_i|0, u_i^{-1}) \mathcal{G}(u_i|a,b) du_i \tag{2.18}$$

where $\mathcal{N}$ is the Gaussian probability density function, $\mathcal{G}$ is the Gamma density, $\boldsymbol{\theta}$ is the sparse features of the high resolution image (for example we can use $\boldsymbol{Dz}$ or the wavelet transform) and a, b are the hyper-parameters. This integral evaluates the t-distribution as discussed in [2, 74].

However in the BCS work [2], the authors used the so-called ML type II estimation which means they compute the most likely value of u_i where $i = 1, ..., N$. This is equivalent to taking the limit when $a = b \to 0$. Therefore this setting removes a lot of the flexibility of this distribution. That is, there is no flexibility to vary the tails and the general shape of the prior. As a result, as we shall see in the later chapters, the recovered image will be too smooth and will loose part of the edges. Hence, there is no flexibility to deal with texture in the image. The estimation for hyper-parameters in t-prior could be automated. However the authors of [2] did not proceed this way and they fixed the degrees of freedom instead. We found in other literature about outlier detection [75] that it is hard to automate the hyper-parameter estimation of the t-prior because of its complex expression and estimating those hyper-parameters are time consuming as claimed in [75]. Therefore, we proceed to formularise another heavy-tail prior that has not been employed in signal recovery and we found that our Pearson prior is superior than t-prior in BCS work [2].

2.6 Inadequacies of Previous Work

We close this chapter by listing some open problems of interest in this area. There are several shortcomings from the previous works in the area to be highlighted in this section that relate to the existing working area. Various image-priors were developed in previous works. Nevertheless, it is still vague which method is the best way to construct the edge-

preserving image-prior. It is unclear which procedures can lead on preserving the edges well in the image. These works [35, 36, 76, 77] and [78] claim that their method can preserve the edges and smooth the noise and we review some of them.

He and Kondi [35] claim their method using Huber MRF preserves the image discontinuities better than the Gaussian prior when compared to a work by Hardie *et al.* [32]. Both works [32, 35] lack automated hyper-parameter estimation which our work overcomes. Hardie *et al.* [32] introduce *proper choice* term on the hyper-parameter. This can be classified as trial and error by repeating the experiments. They then extended the work in [62] but, they still do not provide a principle method on the hyper-parameter estimation. They assume the value can be set between some range by conducting several experiments to obtain the best result with good trial and error value. Therefore, it takes time and is not feasible in practice. Later, He and Kondi [36] extend their work by introducing a method of choosing the threshold for the HMRF prior. They used a heuristic method to estimate the threshold T from several synthetic tests and although it works on two test images, there is no principled methodology behind it. Moreover, those works are specialised for a selected image which limits their general use. The major drawback of a manual search [79] is the difficulty in reproducing results.

Recent work, Pickup *et al.* [76] responded to this matter by employing the non-Gaussian prior (Huber prior) and obtained better results by optimising SR image and registration parameters simultaneously. They learnt prior parameters by using cross validation which can be time consuming. In their setting, several low resolution frames are considered where they can hold enough data when employing the cross validation. Conversely in our setting, these multiple frames do not exist. We conject that the quality of the recovered image might not be as good as theirs when more data is available. However, in chapter 4 we will investigate a cross validation method for estimating the hyper-parameters in our image-prior in order to see the outcome of the recovered image.

The Huber MRF has state-of-the-art performance, provided that its parameters are well chosen according to Pickup *et al.* [67]. A fixed value of the threshold, 0.4 is found

from their provided code[1]. Nevertheless, automating this choice in a principled way is not straightforward, and although the work in [80] has been able to develop an approximate solution to estimating λ, the determination of the threshold parameter δ remains somewhat problematic since the probabilitiy density function is not differentiable in δ.

Fixed estimation refers to a method that estimates the regularization hyper-parameters manually. There is no rigorous or formal definition of a hyper-parameter. Yet, Bergstra and Bengio [79] and Molina *et al.* [81] interpret the hyper-parameter term as a parameter of a prior distribution to differentiate them from parameters of the model. According to Statisticat *LLC* [82], '*the parameters of a prior distribution are called hyper-parameters, to distinguish them from the parameters* Θ *of the model*'. The hyper-parameter is used to control the actual parameter and according to fixed estimation, it is manually tuned and the best value is found based on a particular random grid search. But consider the t-prior in section 2.5.3, the authors [2] call the u_i the hyper-parameters while fixing a and b. There is no reason for this. We now indeed interpret a and b as hyper-parameters. This will obviously increase flexibility. On the other hand, this is more easily said than done with the complicated equations that result. However, as we shall see, a more convenient alternative is offered by the Pearson type VII.

2.7 Conclusions and Motivation

In summary, the main components to differentiate signal recovery and super-resolution applications have been described in this chapter. Both applications allow the recovery of images whilst facing some challenging issues such as the ill-posed condition and robustness aspect. In such situations where the low resolution frame suffers from the down-sampling, the problem is under-determined and harder to be solved. Hence, there is a need for employing a prior-knowlegde or an image-prior. The probabilistic model is applied in formulating the image-prior as it is well-known to be the best on integrating an image-

[1]http://www.robots.ox.ac.uk/ ∼vgg/software/SR/index.html

prior. Its flexibility on capturing any neighbourhood order of Markov Random Field giving the advantage of modifying the feature according to the observed problem.

A study on existing image-priors are also discussed to understand the advantage and disadvantage of each image-prior. The existing image-priors such as Gaussian, Huber and t-prior are described. Lack of a robustness property in Gaussian image-prior, difficulties on estimating the threshold parameter in Huber image-prior, the fixed hyper-parameters in t-prior works motivate the author to improve the recovery edges and at the same time estimating the hyper-parameters in image-prior model automatically.

This thesis aims to contribute towards understanding to the ill-posed problem, a scenario when exploiting the additional information is necessary and highlights the existing image-priors issues. It also points out the significance on having and formulating the image-prior. The following chapter describes the proposed image recovery framework and presents the outcome with supportive results in comparison to the existing image-prior.

CHAPTER 3

SINGLE-FRAME IMAGE RECOVERY USING A PEARSON TYPE VII MRF

In this chapter,[1] a general framework for single image recovery using an image-prior is presented. The primary motivation of this work is to provide a higher accuracy alternatively on the image recovery. This is achieved by devising and employing a heavy-tail image-prior. Section 3.1 introduces the scenario problem for single image recovery and highlights the significance of the proposed image-prior. In section 3.2, the proposed image-prior is described in a mathematical form with reference to Pearson type VII density and details the construction of the neighbourhood matrix. Section 3.3 describes the pseudo-likelihood approximation and section 3.4 outlines the overall framework for the image recovery including the observation and the joint model. In section 3.5, the estimation for the high resolution image and hyper-parameters in the image-prior are explained. Section 3.6 describes the algorithm and all experimental results are presented in section 3.7. Finally, section 3.8 concludes the contribution of this chapter.

[1] A slightly shorter version of the work presented in this chapter appears in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, MLSP, pp. 29-34, 2010 and in *Neurocomputing*, 80: 111-119, 2012, (invited and refereed extension of MLSP'2010 paper)

3.1 Introduction

Compressive imaging and image super-resolution aim to recover a high-resolution scene from its compressed or low resolution measurements. The main difficulty lies with the ill-posedness of the problem, and there is no consensus as to how best to formulate image models that can both impose smoothness and preserve the edges in the image. Here we devise a new image-prior based on the Pearson type VII density integrated with a Markov Random Field model, which has desirable robustness properties. We develop a fully automated hyper-parameter estimation procedure for this approach, which makes it advantageous in comparison with alternatives. Our recovery algorithm, although very simple to implement, achieves statistically significant improvements over previous results in under-determined problem settings, and it is able to recover images that contain texture.

The loss of resolution is often inevitable due to limitations of the camera source. In addition, the capturing process introduces additive noise. Depending on the number of low resolution frames of the scene available, we may talk about single-frame or multi-frame version of the problem. In both cases, most often the observed frames are scarce and noisy, which makes restoration an ill-posed problem. The single-frame version is necessarily under-determined too. Therefore, additional information is required to obtain an adequate solution. In a probabilistic model-based framework, this additional information may be specified in the form of a prior distribution on the salient statistics that images are known to have. The two main characteristics are somewhat conflicting ones: local smoothness and the existence of edges. This makes the specification of a good image-prior challenging.

3.1.1 Previous work of Pearson type VII and motivation

In this chapter, we develop and investigate a perhaps less well-known, but quite convenient robust density, the Pearson type VII, formulated as Markov Random Field (MRF) for image recovery and super-resolution. The Pearson type VII has been used previously in situations where robust, heavy-tail behaviour is required, such as in stock market

modelling [83] and X-ray measurements [84], and for robust density estimation [85] as a more convenient and numerically stable alternative to the t-mixtures. Yet this density has never been formulated in a super resolution field or any image enhancement. Heavy-tail behaviour means the non negligible probability to point far away from the bulk of the density.

Moreover, this simple form of 1^{st} order MRF has been previously employed with success in [32, 35]. Based on this reasonable property, we have studied and formulated a novel image-prior, Pearson type VII based MRF. This allows for greater variability by having larger tails than the standard normal distribution. In this work we exploit the robustness of this density to balance predominant smoothness of images with some allowance for edges or discontinuities. Besides, at the same time we have focused on estimating the hyper-parameters automatically. We devise an alternative robust image-prior, the 1^{st} order of MRF made of univariate Pearson type VII distribution. As shown in chapter 2, figure 2.2 illustrates the region of 1^{st} order neighbourhood and its neighbourhood size.

We employ a Maximum A Posteriori method for parameter estimation, which gives us the most probable high resolution image. Once we have formulated and employed these new priors, we then study which one preserves the image better in terms of mean square error. We also exploit the property of the new research in compressive sensing application [86] to find out how good our proposed image-prior is. More experimentation on this is presented and discussed in the following chapter. Coping with multiple tasks by employing this image-prior will be presented in chapter 6.

3.2 The formulation of the Pearson type VII density

According to Pearson [87], the N-dimensional zero-mean Pearson type VII density is defined as follows:

$$p(\boldsymbol{u}|\boldsymbol{C}, m) = \frac{\boldsymbol{\Gamma}(m)}{\pi^{\frac{N}{2}}\boldsymbol{\Gamma}(m-\frac{N}{2})}|\boldsymbol{C}|^{-\frac{1}{2}}\left[1+\boldsymbol{u}^T\boldsymbol{C}^{-1}\boldsymbol{u}\right]^{-m} \tag{3.1}$$

where $\boldsymbol{u}$ denotes a random variable, $\boldsymbol{C}$ is a $N \times N$ matrix (i.e: covariance matrix), m is the degree of freedom represented by the Gamma $\boldsymbol{\Gamma}$ function that controls the degree of robustness that must satisfy $2m > N$, and N is the dimensionality of the recovered image of $\boldsymbol{z}$. It subsumes the Gaussian when m approaches infinity and the Student-t density. For convenience, the ν notation is denoted as $\nu := 2m - 1$, so that the parameter ν is subject to positivity constraint only, and the univariate Pearson type VII density is written (so, $N = 1$, $m = (\nu + 1)/2$) as:

$$p(u|\lambda, \nu) = \frac{\boldsymbol{\Gamma}(\frac{1+\nu}{2})\lambda^{\nu/2}(\lambda + u^2)^{-\left(\frac{1+\nu}{2}\right)}}{\boldsymbol{\Gamma}(\nu/2)\sqrt{\pi}} \tag{3.2}$$

where the parameter λ replaces $\boldsymbol{C}$ and controls the width of the density, and ν is the degrees of freedom. An illustration of the effect when using the hyper-parameters of λ and ν is shown in figure 3.1.

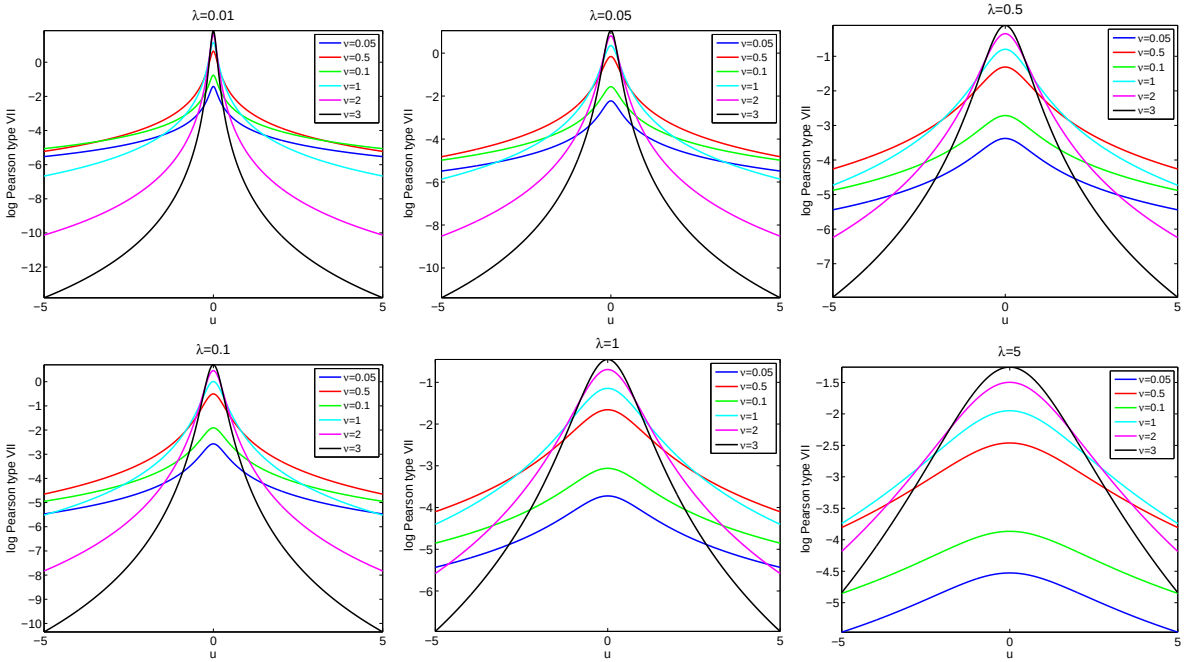


Figure 3.1: Plot of the log Pearson type VII for various values of ν when the hyper-parameter of λ is fixed to a certain value.

As we can see in figure 3.1, either λ is getting smaller or larger (i.e: λ=0.01 and 5) and the ν is increasing, highest peak is obtained. The hyper-parameters λ and ν play an

important role to shrink the density and controls the level of the peak and the tails. It becomes less heavy-tail when the ν is large and vice versa. Therefore, these two hyper-parameters are connected to each other so that the density is balance to cover the edges when it is translated on the neighbourhood feature histogram.

3.2.1 The Pearson type VII MRF as an image-prior

The main characteristic of any natural image is a local smoothness. That is, the intensities of neighbouring pixels tend to be similar. Any reasonable image model needs to be able to capture this property. As mentioned in chapter 2, Markov Random Fields are well suited and broadly used models that formalise this. For this reason, the Pearson type VII is devised and formulated as an image-prior to capture the local smoothness property and with its heavy-tail property allows to preserve the edges on an image.

A very simple form of 1^{st} order MRF, previously employed with success for image recovery in e.g. [35, 32], is to condition each pixel intensity on its four cardinal neighbours in the following way. For any one pixel z_i define:

$$\begin{aligned} p(z_i|\boldsymbol{z}_{-i}) &= p(z_i|z_{4\text{neighb}(i)}) \\ &\propto g(z_i - \frac{1}{4}\sum_{j\in 4\text{neighb}(i)} z_j) \end{aligned} \tag{3.3}$$

where the notation $\boldsymbol{z}_{-i}$ means all the pixels excluding the i-th, and the set of four cardinal neighbours of z_i was denoted as 4neighb(i). g is a function that does not to be a normalised density. For example in case of Pearson type VII density, the function g will take the following form $g(u)= (u^2 + \lambda)^{-\left(\frac{1+\nu}{2}\right)}$. These are univariate probability distributions. We should mention that alternatives include the so-called total variation model, employed in [67], which is based on image gradients. The experimental comparison in [35] suggests that the model in eq.(3.3) and total variation behave in a very similar manner, the former being slightly superior however.

Using eq.(3.3), for an image $\boldsymbol{z}$ of N pixels, the MRF represents the joint probability

over all the pixels on the image — a multivariate probability distribution:

$$p(\boldsymbol{z}) = \frac{1}{Z}\prod_{i=1}^{N} g(z_i - \frac{1}{4}\sum_{j\in 4\text{neighb}(i)} z_j) \tag{3.4}$$

where $Z = \int d\boldsymbol{z} \prod_{i=1}^{N} g(z_i - \frac{1}{4}\sum_{j\in 4\text{neighb}(i)} z_j)$ is the normaliser (or partition function) of the MRF. This is independent of $\boldsymbol{z}$ but depends on the hyper-parameters of the constituent probability density building blocks. Equation (3.4) needs to normalise because the multivariate density of $p(\boldsymbol{z})$ has to integrate to one with respect to $\boldsymbol{z}$. Even if the g would be normalised separately for each pixel, it would not be sufficient because the pixels are not statistically independent from each other.

The simplicity of (3.4) is also intuitively appealing. One can think of the difference between a pixel intensity and the average intensity of its neighbours, i.e. $z_i - \frac{1}{4}\sum_{j\in 4\text{neighb}(i)} z_j$, as a *feature*. However, the partition function Z is intractable to compute analytically, except for a very few specific cases. Therefore, approximations may be employed. For notational convenience, it is handy to create the symmetric $N \times N$ matrix $\boldsymbol{D}$ to encode the above neighbourhood structure, with the following entries:

$$d_{ij} = \begin{cases} 1 & \text{if } i = j; \\ -1/4 & \text{if i and j are neighbours;} \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Then we may write the i-th feature in a vector form, with the aid of the i-th row of this matrix (denoted $\boldsymbol{D}_i$) as the following:

$$z_i - \frac{1}{4}\sum_{j\in 4\text{neighb}(i)} z_j = \sum_{j=1}^{N} d_{ij} z_j = \boldsymbol{D}_i \boldsymbol{z} \tag{3.6}$$

In Appendix A, the example of image size (3,4) derives the final equation (3.6). $\boldsymbol{D}_i\boldsymbol{z}$ is the difference between the i_{th} pixel of $\boldsymbol{z}$ and the average of its 4 neighbours. $\boldsymbol{D}_i\boldsymbol{z}$ gives

the smaller value of its difference when the average of four neighbours are almost similar to the pixel itself. It means that the z i_{th} neighbourhood tends to be similar and this characteristic known as local smoothness. It will produce a bigger difference when the average of four neighbours are not the same. This shows that neighbourhood of the pixel itself has different edges. Essentially in this image model, each image is represented by a histogram of $\boldsymbol{D}_i\boldsymbol{z}$ with the aid of the neighbourhood feature. Previously in section 2.4.2, figures 2.5 and 2.6 present some of the histograms of the natural images that have been investigated.

3.3 The Pearson type VII MRF

We now propose to employ the Pearson type VII density with an MRF to provide a novel robust image model. One option would be to use its multivariate form as given in eq.(3.1) by encoding the neighbourhood structure in $\boldsymbol{C}^{-1} = \boldsymbol{D}^T\boldsymbol{D}$. However, this multivariate heavy-tail distribution would then be asserted on whole images (or possibly image patches) rather than tiny pixel neighbourhoods. We do not pursue this option here since our goal is to give non-zero probability to edges in the image, which requires a pixel-level modelling. Neighbourhood features that correspond to pixels that are situated at an edge may be thought of as spikes or outliers that our heavy-tail prior will account for, and this is what enables us to preserve the edges in the recovered image. To achieve this, we build up our MRF prior from univariate PearsonVII densities, as the following:

$$p(\mathbf{z}) = \frac{1}{Z_P(\lambda,\nu)} \prod_{i=1}^{N} \left\{\lambda + (\boldsymbol{D}_i\boldsymbol{z})^2\right\}^{-\left(\frac{\nu+1}{2}\right)} \tag{3.7}$$

where $Z_P(\lambda,\nu) = \int d\boldsymbol{z} \prod_{i=1}^{N} \{\lambda + (\boldsymbol{D}_i\boldsymbol{z})^2\}^{-\left(\frac{\nu+1}{2}\right)}$ is the partition function, and this multivariate integral does not have an analytic form.

As with all MRF priors, the partition function may be neglected as long as we are interested in a *maximum a posteriori* estimate of $\boldsymbol{z}$ with some known and fixed hyperpa-

rameters. However, the partition function does depend on the hyper-parameters, hence for an automated estimation of these based on the model, the partition function must be approximated and taken into account. Notice that, in the case of a Pearson type VII MRF, the partition function is smooth w.r.t. both λ and ν — unlike the Huber MRF, which is non-smooth in δ. Hence, with a suitable analytic approximation of $Z_P(\lambda, \nu)$ this may be used for hyper-parameter estimation.

3.3.1 Pseudo-likelihood approximation

We shall employ a pseudo-likelihood approximation to the partition function $Z_P(\lambda, \nu)$. It consists of taking each $\boldsymbol{D}_i\boldsymbol{z}$ as if it were independent of $\boldsymbol{D}_j\boldsymbol{z}$, for all $j \neq i$ to break down the intractable multivariate integral into tractable univariate integrals. Thus, we have the following:

$$Z_P(\lambda, \nu) \approx \prod_{i=1}^{N} \int dz_i p(z_i|\boldsymbol{z}_{-i}) = \left\{ \frac{\Gamma(\nu/2)\sqrt{\pi}}{\Gamma(\frac{1+\nu}{2})\lambda^{\nu/2}} \right\}^N \tag{3.8}$$

i.e. the inverse of the product of the normalising terms of the univariate Pearson type VII density building blocks.

Replacing this into the definition (3.7), we have the following approximate image model:

$$p(\boldsymbol{z}|\lambda, \nu) \approx \prod_{i=1}^{N} \frac{\Gamma(\frac{1+\nu}{2})\lambda^{\nu/2}((\boldsymbol{D}_i\boldsymbol{z})^2 + \lambda)^{-\frac{1+\nu}{2}}}{\Gamma(\nu/2)\sqrt{\pi}} \tag{3.9}$$

We are now ready to employ this in the overall model for super-resolution, and use this to infer $\boldsymbol{z}$ simultaneously with estimating our hyperparameters λ and ν.

3.4 The Overall Framework for Image Recovery

3.4.1 Observation model

Denoting the vectorised high-resolution image by $\boldsymbol{z}$, as before, this is now a hidden variable. Instead, some low resolution version of it is observed. The degradation process will

be taken as a linear transform, and we should note that, although this is a simplifying assumption, it has worked well in many super-resolution application so far [32, 35, 67].

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\eta} \tag{3.10}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ y_m \\ y_M \end{pmatrix} = \begin{pmatrix} W_{11}, & W_{12}, & . & . & W_{1n} \\ ... & & & & \\ W_{m1}, & W_{m2}, & . & . & W_{mn} \\ ... & & & & \\ W_{M1}, & W_{M2}, & . & . & W_{Mn} \end{pmatrix} \times \begin{pmatrix} z_1 \\ z_2 \\ . \\ z_n \\ z_N \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ . \\ \eta_m \\ \eta_M \end{pmatrix}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ is an additive noise. Equivalently, we can write $p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{W}\boldsymbol{z}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{y}$ is the observed version of the image, with $M < N$ pixels, and σ^2 is the observation noise variance. In single-frame super-resolution, the transform $\boldsymbol{W}$ typically contains blur and down-sampling. In the multi-frame case we also have shift that varies between the observed frames and in that case $\boldsymbol{y}$ is a concatenation of all the vectorised low resolution frames observed from the scene of interest. The single-frame problem is more challenging in that the system is under-determined (i.e. there are less observed pixel intensities than there are unknown ones).

3.4.2 Joint model

In this section, a joint probability is adopted to find out how likely it is that two (or more) events happen at the same time. The overall model is the joint model of the observations $\boldsymbol{y}$ and the unknowns $\boldsymbol{z}$. That is, $\boldsymbol{Pr(y, z)}$. To assemble this from the previously presented components, we first rewrite the observation model given in equation (2.2) in the form of a probability distribution of the observations $\boldsymbol{y}$ given the ground truth $\boldsymbol{z}$. That is, $\boldsymbol{Pr(y|z)}$. Using this, the joint probability is written as follows:

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{W}, \sigma^2, \lambda, \nu) = p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{W}, \sigma^2)p(\boldsymbol{z}|\lambda, \nu) \tag{3.11}$$

The overall working model consists of the first term which is the observation model and the second term which is the image-prior model. So we now have the pseudo-joint likelihood, assuming $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \boldsymbol{I})$ the additive observation noise was assumed to be zero mean spherical Gaussian.

3.5 MAP-based Estimation in the model with Pearson type VII MRF

We will use the joint probability (3.11) as the objective to be maximised. Maximising this w.r.t. $\boldsymbol{z}$ is also equivalent to finding the most probable image $\boldsymbol{z}$, i.e. the maximum a posteriori (MAP) estimate, since (3.11) is proportional to the posterior $p(\boldsymbol{z}|\boldsymbol{y})$. Equivalently, the negative log of this expression will be defined as our minimisation objective:

$$Obj(\boldsymbol{z}, \sigma^2, \lambda, \nu) = -\log[p(\boldsymbol{y}|\boldsymbol{z}, \sigma^2)] - \log[p(\boldsymbol{z}|\lambda, \nu)] \tag{3.12}$$

Plugging in the functional forms of the two density functions, we then minimise this w.r.t. $\boldsymbol{z}$ and the hyper-parameters in turn.

3.5.1 Estimating the most probable $\boldsymbol{z}$

Putting all the terms in objective (3.12) will yield equation (3.13). The terms in the objective (3.12) that depend on $\boldsymbol{z}$ are the following:

$$\begin{aligned} Obj(\boldsymbol{z}, \sigma^2, \lambda, \nu) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z})^2 + N\log(\pi) + \frac{1}{2}\log|\Sigma| + N\log\Gamma\left(\frac{\nu+1}{2}\right) + N\log(\lambda)^{\nu/2} \\ - N\log\Gamma\left(\frac{\nu}{2}\right) - N\log(\pi)^{1/2} - \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{N}\log\left\{\lambda + (\boldsymbol{D}_i\boldsymbol{z})^2\right\} \end{aligned} \tag{3.13}$$

By collecting the terms that depend on $\boldsymbol{z}$, the objective function of $\boldsymbol{z}$ is written as follows:

$$Obj_z(\boldsymbol{z}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z})^2 + \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{N}\log\left\{\lambda + (\boldsymbol{D}_i\boldsymbol{z})^2\right\} \tag{3.14}$$

The optimisation of (3.14) w.r.t. $\boldsymbol{z}$ may be done employing any nonlinear optimiser, the objective is differentiable. We employed a conjugate gradient type method[1], which requires gradient information. The gradient is derived as the following.

$$\nabla_{\boldsymbol{z}} Obj_{\boldsymbol{z}} = \frac{1}{\sigma^2}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{z} - \boldsymbol{y}) + (\nu+1)\sum_{i=1}^{N}\boldsymbol{D}_i^T\frac{\boldsymbol{D}_i\boldsymbol{z}}{(\boldsymbol{D}_i\boldsymbol{z})^2 + \lambda} \tag{3.15}$$

3.5.2 Estimation of $\boldsymbol{\sigma^2}$

Similarly writing out the terms in (3.12) that depend on σ^2, objective function for $\boldsymbol{\sigma^2}$ can be written as following:

$$\begin{aligned} Obj_{\boldsymbol{\sigma^2}}(\boldsymbol{z}, \sigma^2, \lambda, \nu) &= \frac{1}{\sigma^2}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{z} - \boldsymbol{y}) + \frac{1}{2}\log|\Sigma| \\ &= \frac{1}{\sigma^2}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{z} - \boldsymbol{y}) + \frac{1}{2}\log|\sigma^2 I| \end{aligned} \tag{3.16}$$

By taking derivative and solving the equation (3.16), we get a closed form estimate for σ^2:

$$\hat{\sigma}^2 = \frac{1}{M}\left(\sum_{i=1}^{M}(y_i - \boldsymbol{W}_i\hat{\boldsymbol{z}})^2\right) \tag{3.17}$$

3.5.3 Estimation of λ and ν

The terms that depend on λ and ν can be written as follows in equation (3.18) and (3.19):

$$Obj_\lambda(\boldsymbol{z}, \sigma^2, \lambda, \nu) = \frac{N\nu}{2}\log\lambda - \left(\frac{1+\nu}{2}\right)\sum_{i=1}^{N}\log((\boldsymbol{D}_i\boldsymbol{z})^2 + \lambda) \tag{3.18}$$

[1]We made use of the efficient implementation available from http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/

$$Obj_\nu(\boldsymbol{z}, \sigma^2, \lambda, \nu) = N \log \Gamma\left(\frac{\nu+1}{2}\right) + N \log(\lambda)^{\nu/2} - N \log \Gamma\left(\frac{\nu}{2}\right) - N \log(\pi)^{1/2} + \sum_{i=1}^{N} \log\left\{\lambda + (\boldsymbol{D}_i \boldsymbol{z})^2\right\} \tag{3.19}$$

Both of these hyper-parameters need to be positive valued. To ensure our estimates are non-negative, we parametrise the log probability objective (3.18) and (3.19) such as to optimise for the ($\pm$) square root of these parameters. Taking derivatives w.r.t. $\sqrt{\lambda}$ and $\sqrt{\nu}$, we get:

$$\frac{d}{d\sqrt{\lambda}} Obj_\lambda = \sum_{i=1}^{N} \frac{\nu(\boldsymbol{D}_i \boldsymbol{z})^2 - \lambda}{((\boldsymbol{D}_i \boldsymbol{z})^2 + \lambda)\sqrt{\lambda}} \tag{3.20}$$

$$\frac{d}{d\sqrt{\nu}} Obj_{nu} = \left[N \log \lambda - \sum_{i=1}^{N} \log\left((\boldsymbol{D}_i \boldsymbol{z})^2 + \lambda\right) + N\psi\left(\frac{1+\nu}{2}\right) - N\psi\left(\frac{\nu}{2}\right)\right] \sqrt{\nu} \tag{3.21}$$

where $\psi(\cdot)$ is the digamma function. The zeros of these functions give us the estimates of $\pm\sqrt{\lambda}$ and $\pm\sqrt{\nu}$. Although there is no closed-form solution, these can be obtained numerically using any unconstrained nonlinear optimisation method. The square of these estimates give us the estimates of λ and ν respectively.

3.6 The Algorithm

- Initialise the estimate $\hat{\boldsymbol{z}}$, e.g. as some combination of the solution of a Gaussian MRF and random noise.
- Iterate until convergence:
  - Estimate σ^2 using (3.17).
  - Perform iterations to update λ and ν in turn, using (3.20) and (3.21), keeping the current estimate $\hat{\boldsymbol{z}}$ fixed.
  - Perform iterations to update $\hat{\boldsymbol{z}}$ using (3.15)

Note that, the inner loops need not completely converge. It is sufficient to increase, not necessarily to minimise the objective at each intermediate step. However, we observed faster overall convergence by letting the inner iterations make more progress. The reason is probably that the overall objective is complex, with multiple local optima, while the individual updates break it down into simpler objectives in a greedy manner. Our MatLab implementation is available from `http://www.cs.bham.ac.uk/~sxa814/codes/PearsonMRF_code/`

3.7 Experiments and Results

We conducted experiments with both classical super-resolution (SR) matrices where $\boldsymbol{W}$ comprises blur and down-sampling, as well as with random Gaussian compressive sensing (CS) matrices where $\boldsymbol{W}$ has random entries sampled i.i.d. from a standard Gaussian. The latter is of interest in the light of new research in compressed sensing and signal processing [86, 2] directed towards devising hardware that can exploit some good theoretical properties of certain random matrices.

The observation data was generated starting from a ground truth real image via the matrix $\boldsymbol{W}$ and additive noise. Working on synthetic data allows us to compare the recovered image against the ground truth, so that we can measure our recovery performance quantitatively.

3.7.1 Illustrative experiments

We start by demonstrating the working of our algorithm. Figure 3.2 shows an example of under-determined case, where we recover a $[80 \times 70]$ '`cameraman`' image (that is, N=5600 pixels) from its M=4000 randomly compressed measurements and additive noise of $\sigma = 0.5$. We will be concerned with under-determined systems in this paper, however for the sake of completeness, we next show an overdetermined case as well, derived from a classical multi-frame super-resolution task, i.e. the transformation (or measurement) matrix consists of random shifts, Gaussian blur with point spread function set to 0.4 and

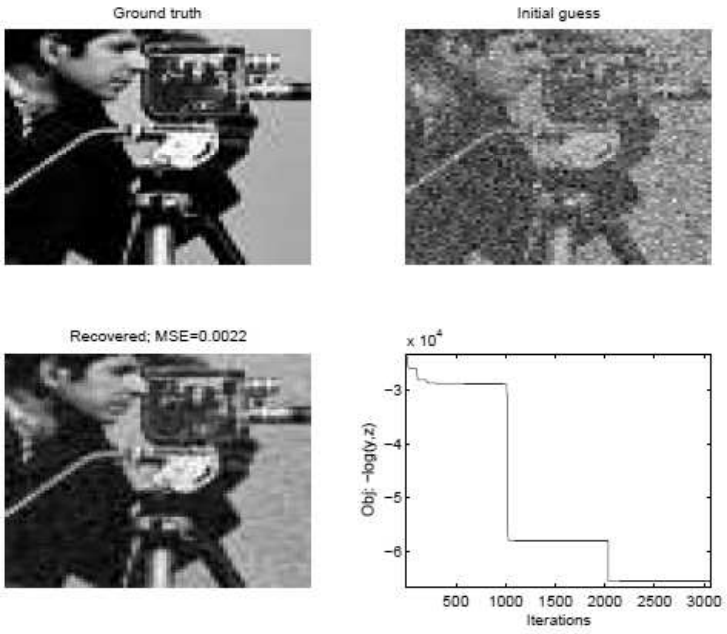


Figure 3.2: Example recovery of '`cameraman`' (5600 pixels) from random linear mapping to 4000 pixels and additive noise with $\sigma = 0.5$.

down-sampling. Here we generated 18 low resolution images with a zoom factor of 3, so the overall system is over-determined in this case. Figure 3.3 shows the ground truth, a straw-man recovery by averaged bi-cubic interpolation from the individual low resolution frames (which we use as an initial guess to seed our algorithm in this experiment), and the obtained recovered image, along with the evolution of the objective over the iterations. It is easy to see from the evolution of the objective function over the iterations (and the quality of recovered image) that having access to more observation frames makes the recovery task much easier.

In the remainder of the chapter we will focus on recovery from a single-frame, i.e. under-determined systems — such as recovering a 5600 pixels [80×70] image from a single $M \leq 4000$-pixel frame. Quite obviously, without the specification of a prior, such a system would have infinitely many solutions, hence under-determined systems are much more reliant on the prior image model. In addition, the observations are subject to Gaussian additive noise, and this makes the recovery problem even harder.

In the case of CS-type $\boldsymbol{W}$, the noise standard deviations that we tested were $\sigma \in \{4 \times 10^{-5}, 0.5, 1, 2\}$. In the case of SR-type $\boldsymbol{W}$ these values were divided by $0.8\sqrt{N}$ to

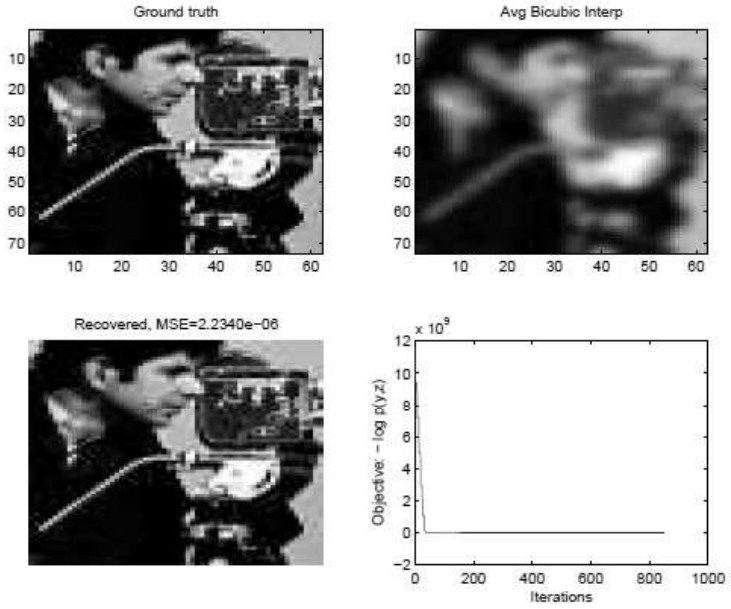


Figure 3.3: Example recovery from multiple (18) low resolution (zoom factor of 3) frames, which together represent an over-determined system.

make the signal-to-noise ratios roughly the same for the two matrix types. This is still relatively high noise, considering that we scaled the pixel intensities in the generating ground truth image to the interval $[-0.5, 0.5]$.

From each low-resolution data set, we then try to recover the ground truth image, and we assess the performance by measuring the mean square error (MSE) between the recovered image $\hat{\boldsymbol{z}}$ and the ground truth $\boldsymbol{z}$ — that is, $MSE = mean((\boldsymbol{z} - \hat{\boldsymbol{z}})^2)$.

Initialisation

Given that we optimise a non-convex objective, the initialisation scheme may impact the solution and the speed of convergence. Empirically we found that using CS-type matrices $\boldsymbol{W}$ the quality of the solution is much less sensitive to initialisation than it is in the case of SR-type matrices. The main issue, for SR-type, is to avoid starting it off from a neighbourhood of a local optimum. Therefore, in the case of under-determined problems we need to avoid using the output of a simpler super-resolution recovery method as an initial guess, as it often turns out to lead to an unwanted local optimum. On the other hand, a completely random initialisation would take longer to converge. Based on these

considerations, in all our experiments we adopted the following scheme. In experiments with SR-type matrices we initialise $\boldsymbol{z}$ with the average between the minimum energy estimate ($\boldsymbol{W}^T\boldsymbol{y}$) and a random guess drawn from standard Gaussians. For CS-type matrices, since we did not experience any local optima issues, we opted to use a ridge regression to produce the initial guess (although other schemes that we tried did not make any noticeable difference). In both cases, we initialised the hyper-parameters with $\nu = 10, \lambda = 1, \sigma^2 = 0.001$.

3.7.2 Assessment of the modelling power of the Pearson-type-VII image-prior

Before diving into the assessment of our full algorithm, we switch off the automated hyper-parameter estimation in this subsection. Here we assess the Pearson type VII based image model by comparing it with state-of-the-art alternative image-priors when each of the competing prior models is supplied their best hyperparameters. For this purpose we select the best hyperparameters for each competing model based on the MSE with the ground truth. This, of course is not feasible in practice since the ground truth is not available, but it provides us information on what each model can achieve at its best. We will then move on to assess our automated hyperparameter estimation procedure against these idealised best results in section 3.7.3. We used CS-type matrices $\boldsymbol{W}$ in this set of experiments. The competing methods are: Gaussian-MRF, a multivariate-Student-t based MRF that we also experimented with, and the Huber-MRF.

Figure 3.4 summarises the results obtained for the '`church`' image[1] against varying noise conditions. Figure 3.5 shows the best recovery for the setting with $\sigma = 0.5$. We see from the figure 3.4 that the Pearson type VII based MRF model can achieve state-of-the-art performance in all noise levels tested, comparable to that of Huber-MRF, while the other priors tested perform worse. However, as already mentioned, for the Huber-MRF, a principled determination of both of its hyper-parameters would not be straightforward.

[1] http://www.robots.ox.ac.uk/∼vgg/research/SR/synthdata.html

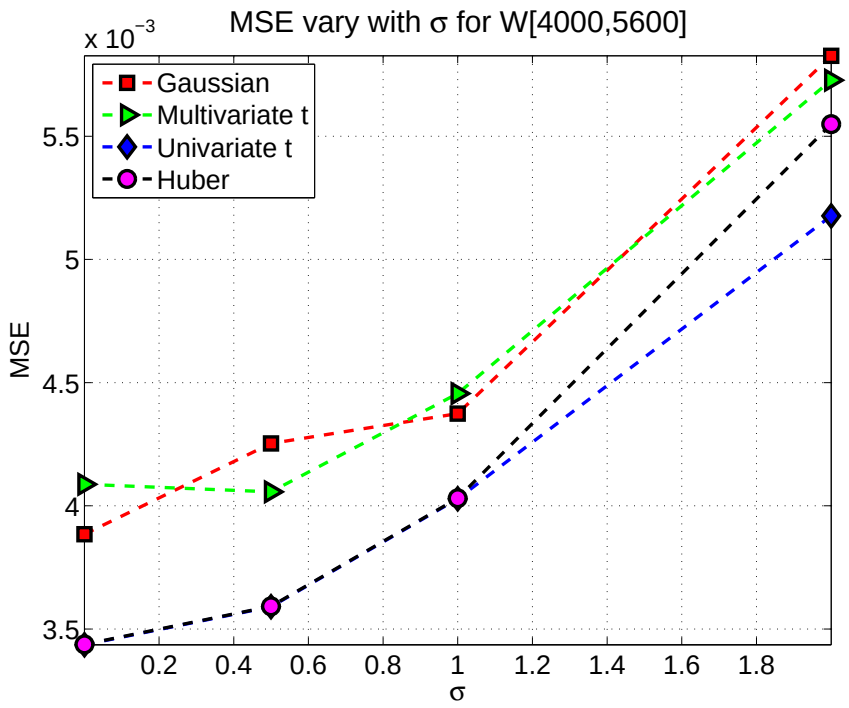


Figure 3.4: Comparative MSE performance for the under-determined system in progressively increasing noise conditions, using the best hyperparameter values (i.e. the value that produces the smallest MSE).

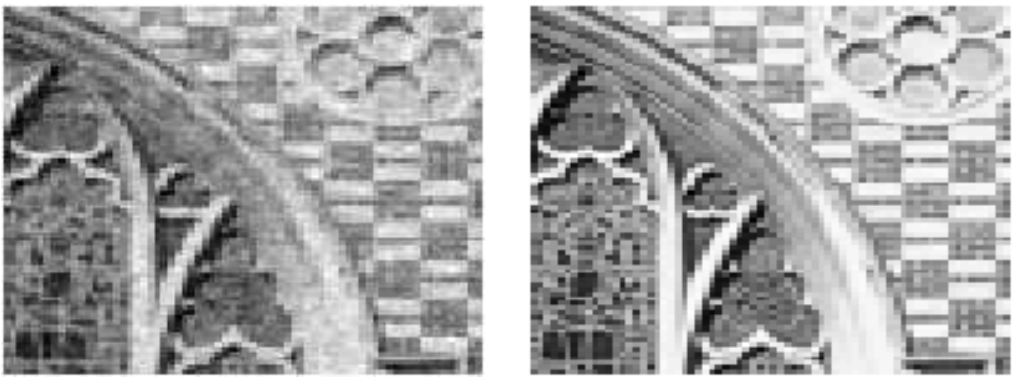


Figure 3.5: On the left plot is the best recovered with manual tuning; and on the right plot is the ground truth.

The next question is then, how does the automated hyper-parameter estimation of our Pearson type VII based MRF prior compare to these hand-picked best results?

3.7.3 Assessment of the automated hyperparameter estimation procedure

Keeping the same experimental conditions set out in the previous section, figure 3.6 shows the MSE achieved by our recovery algorithm that includes automated hyperparameter estimation, superimposed with the best manually picked results (of the same prior) from figure 3.4 for reference. We see that, except for very high levels of noise, the agreement

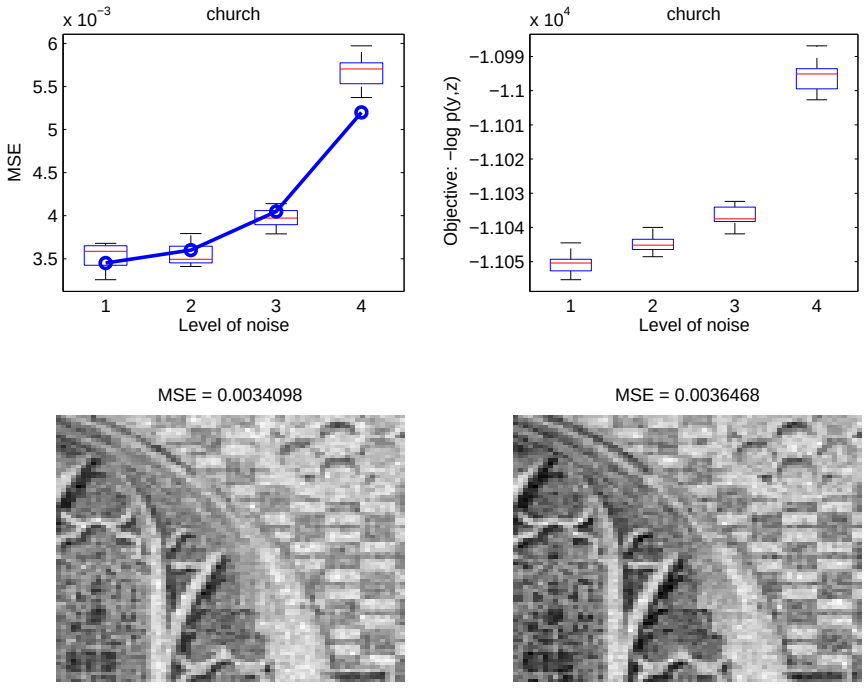


Figure 3.6: Comparing the performance of the fully automated Pearson type VII based MRF approach with the best result found by manual tuning of the hyperparameters. *Top left*: The distribution of MSE; *Top right*: The distribution of the values of the objective function; The boxplots represent 10 independent repeats where in each trial the additive noise and the transform $\boldsymbol{W}$ were randomly drawn anew. *Bottom left*: Best result out of the 10 repeats with $\sigma = 0.5$, picked by lowest MSE. *Bottom right*: Best result out of the 10 repeats, picked by lowest values of the objective function. We see the MSE of the latter is very close to that of the former.

is remarkable. In fact, the MSE at the highest of the noise levels is still comparable with

that of the best manual tuning of Huber MRF. Hence, we can conclude that, in these experiments the Pearson type VII based MRF is preferable as a fully automatic method.

In addition, the good agreement between the MSE and the values taken by the objective function is notable. Note that the calculation of the MSE requires access to the ground truth image, while the objective does not. Hence, the agreement between these two quantities represent further evidence for the appropriateness of our proposed model and automated estimation procedure. In other words, the best (or close to best) results in terms of agreement with the ground truth can be found by accessing the objective function independently of the ground truth. Indeed, the MSE of the recovered image selected solely on the basis of the objective function (MSE=0.0036) is not far off from the best MSE across the 10 repeats (MSE=0.0034).

3.7.4 Comparison with Bayesian Compressed Sensing (BCS)

A recent technique that is also fully automated has been proposed in the field of compressed sensing [2], called Bayesian Compressed Sensing (BCS), which is based on the Automatic Relevance Determination (ARD) principle. It is interesting to compare our results with those of this method, since modelling-wise BCS is somewhat related to our approach in that it targets the solution of an under-determined linear system with the use of a probabilistic model and a prior. The prior they employ is the improper uninformative limit of a Student-t prior. To make the link, we note that the Pearson type VII density subsumes the Student-t if we set λ to $\lambda\nu$. However, the algorithmic solution of BCS differs from ours, and so does the authors [2, 3] choice to use the non-informative limit of the prior. Hence, it is of interest to see the effect of these differences comparatively[1]. Unless stated otherwise, we will use the authors improved version of BCS from [3].

Figure 3.7 shows results obtained with BCS on the same data and experimental setup as we used in figure 3.6. Although the performance seems to be not hugely different, it

[1]We used the authors implementation that is available from `http://people.ee.duke.edu/~lcarin/BCS.html`

is still worse than what we had achieved. From the figure we see the MSE values over 10 repeats are higher, and the best MSE (0.0081) is still worse than the MSE of our choice based solely on our objective function in figure 3.6 (MSE=0.0036). Looking at the

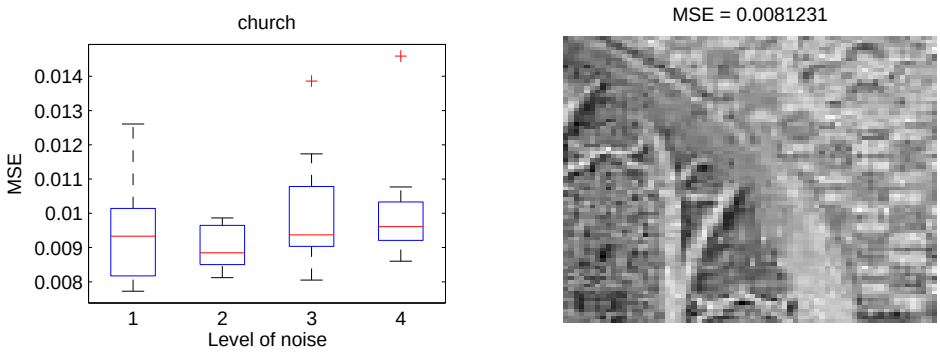


Figure 3.7: Results obtained with the Bayesian Compressed Sensing (BCS) algorithm of [2], to be compared with figures 3.5 and 3.6. Left: The distribution of MSE over 10 independent repeats; Right: The best recovery across all these repeats. Observe the best MSE is still higher than the MSE of the Pearson-VII result picked cf. the best value of the objective function.

recovered image (Fig.3.7) we notice that BCS tends to discard part of the edges in favour of more homogeneous areas, and this degrades performance when there is insufficient data. It should be pointed out that the same over-sparsifying problem associated with the use of the non-informative limit of Student-t priors in under-determined systems has been also reported in [88] in the context of logistic regression-based classification problems. A general theory that explains the behaviour of promoting local strong homogeneity by log-priors that are non-smooth at zero can be found in [89]. It may be interesting to note that our log Pearson prior is actually smooth everywhere, but may exhibit a sharp curvature at zero when ν is small. Hence it is flexible enough to be able to promote local homogeneity without over-emphasising it.

We observed similar results on several different images in our experiments, with both CS-type and SR-type instantiations of $\boldsymbol{W}$. In figures 3.8-3.9 we give results on another image ('`castle`') where we used SR-type $\boldsymbol{W}$ and the same noise conditions as before. The observations made earlier are apparent again. From these results it seems that our algorithm is indeed able to recover a good quality high resolution image even when the

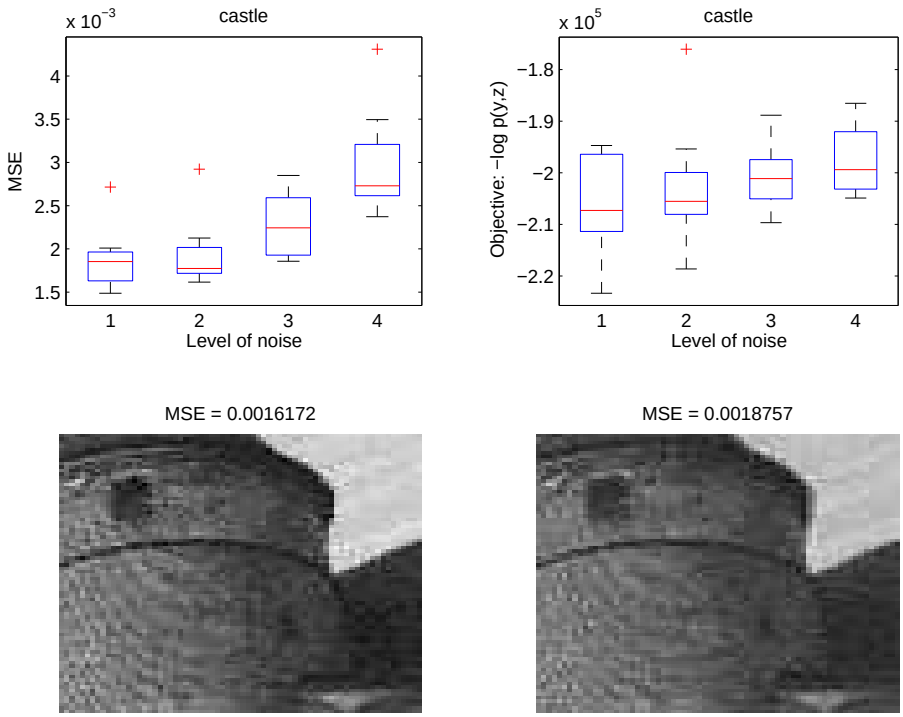


Figure 3.8: *Top left*: The distribution of MSE for our Pearson-VII based approach; *Top right*: The distribution of the objective functions values for our Pearson-VII based approach; The boxplots represent 10 independent repeats. *Bottom left*: Best result with Pearson-VII, out of the 10 repeats, picked by lowest MSE at noise level $\sigma = 0.5$. *Bottom right*: Best result with Pearson-VII, out of the 10 repeats, picked by lowest values of the objective function at noise level $\sigma = 0.5$.

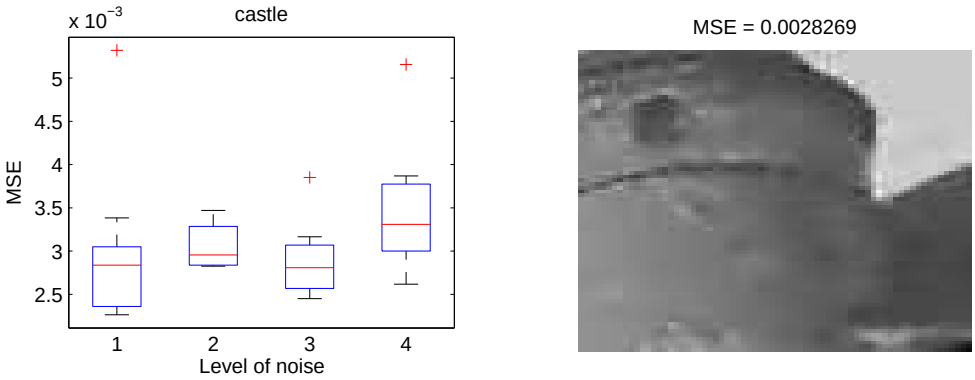


Figure 3.9: *Left*: The distribution of MSE for the BCS approach. *Right*: Best result with BCS, out of the 10 repeats, picked by lowest MSE at noise level $\sigma = 0.5$. The best recovery from BCS (MSE=0.0028) is higher than the pick that only uses the Objective (MSE=0.0018). BCS tends to discard part of the edges in favour of strong local homogeneity.

image contains more texture.

Systematic experimental validation

Next we validate our findings by performing a battery of comparative experiments between our Pearson-type-VII based recovery algorithm and BCS method. Before doing this on image data, we find it instructive to take the one dimensional (1D) sparse spiky signal used as a first test benchmark for compressed sensing algorithms (e.g. in [2] and [3]). It is known that the less sparse the signal is and the less observation measurements we have the more difficult the recovery problem. As in [3] (Fig.2 in [3]), we take signals of length $N = 512$ having 20 non-zero entries of ± 1, random Gaussian compressive transform, we vary the number of observations, and measure the reconstruction error of the two recovery algorithms.

Figure 3.10 shows an example of recovery, where the number of observations are too few for BCS to cope with. In turn, our algorithm manages to recover the signal to a great extent. In figure 3.11, we give the full picture of this comparison for the recovery of the spike signal. As before, BCS refers to the improved version of BCS described in [3], and we also tried the previous version of this method, described in [2], which is referred to as BCS_o in the legend of our figure. We did not consider multi-task settings in this work though.

We see that our Pearson-VII is able to recover the signal from fewer measurements than BCS can. This also means that given the same number of measurements it can recover signals up to a larger number of spikes. This in turn implies in an image-reconstruction context that it can recover more edges, i.e. it can deal with more textured images. Returning to image recovery, we now conduct experiments varying the number of measurements, and fixing the noise level to $\sigma = 0.005$. Figure 3.13 shows the comparative results obtained on four different natural images[1] for both types of $\boldsymbol{W}$ (CS-type, and SR-type). The images are: '`cameraman`' (104×94 pixels), '`castle`', '`face`', and '`flower`'

[1] `http://www.cs.bham.ac.uk/~axk/Sakinah/PearsonMRF_code/`

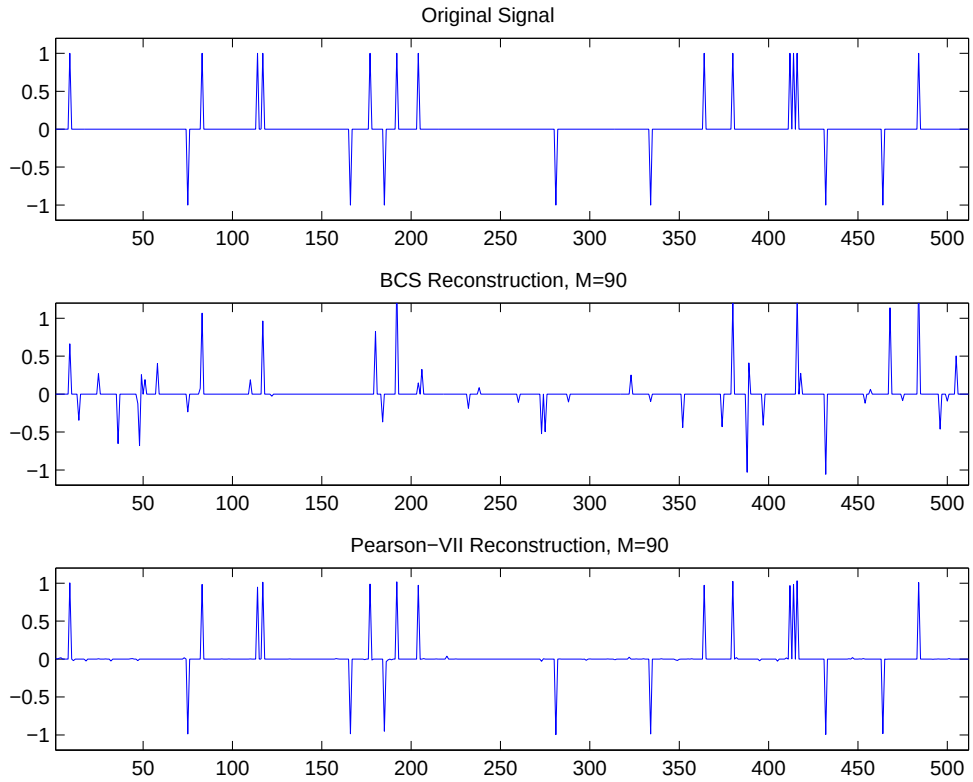


Figure 3.10: Reconstruction of the 'spikes' signal of length $N = 512$ having 20 non-zero entries from only $M = 90$ random compressive measurements. The first two subplots are reproduced from Fig.2 of [3] whereas the last subplot shows our recovery result.

(90×90 pixels each). We vary the number of compressive / low resolution observations down to 300 pixels. From these figures, in each experiment our algorithm achieves a statistically significant improvement over BCS in severely under-determined problem settings.

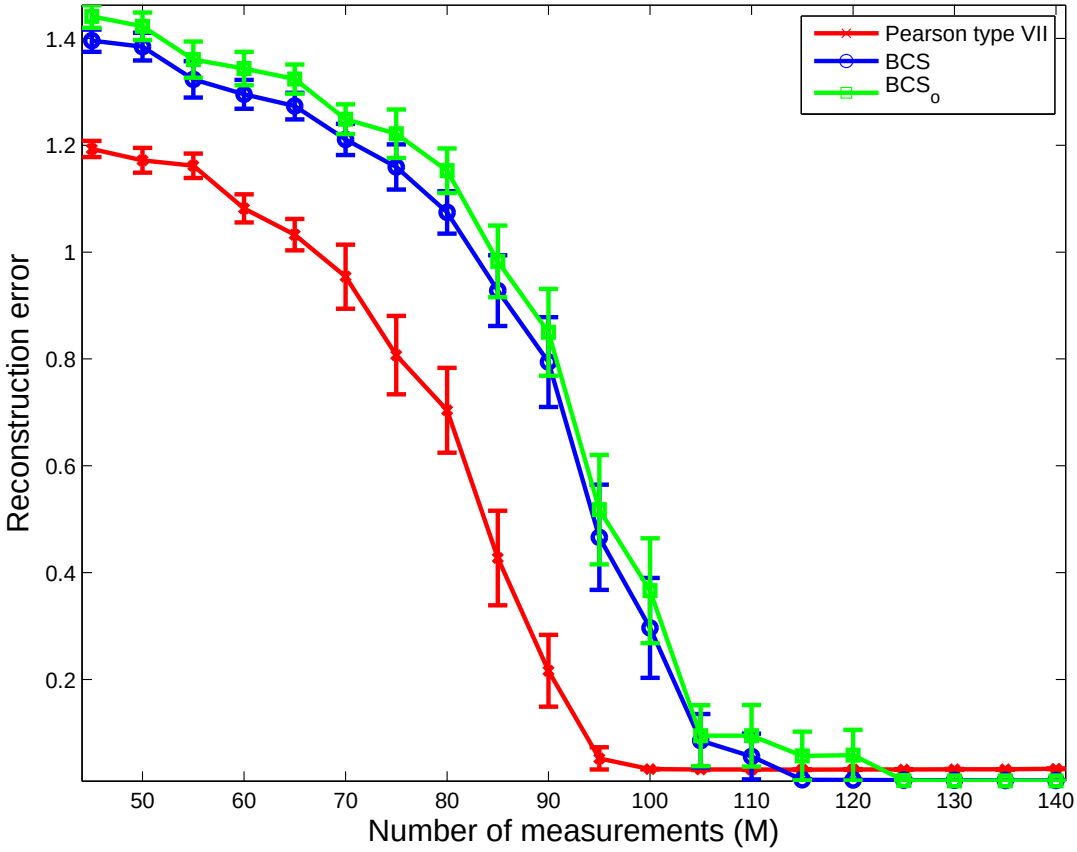


Figure 3.11: Comparison of recovering the 'spikes' signal from its CS measurements and additive noise of $\sigma = 0.005$ (the same setting as in Fig.2 of [3]). The error bars represent one standard error about the mean, from 25 independent repeats. We see that PearsonVII can recover the signal from fewer measurements than BCS.

3.8 Conclusions

In this chapter we formulated a new image-prior based on Pearson type VII densities integrated with a MRF. Our main motivation has been to exploit the heavy-tail property of this density, which indeed seems to be a good way of preserving edges while imposing smoothness. The form of this prior has the additional advantage of allowing us to perform a fully automated hyperparameter estimation. Our recovery algorithm, although very simple to implement, achieves statistically significant improvements over Bayesian Compressed Sensing in under-determined problem settings, and is also able to recover more textured images than BCSg can. Future work inludes multi-task extensions of this approach, where the hyperparameters would be shared accross tasks while other similar-

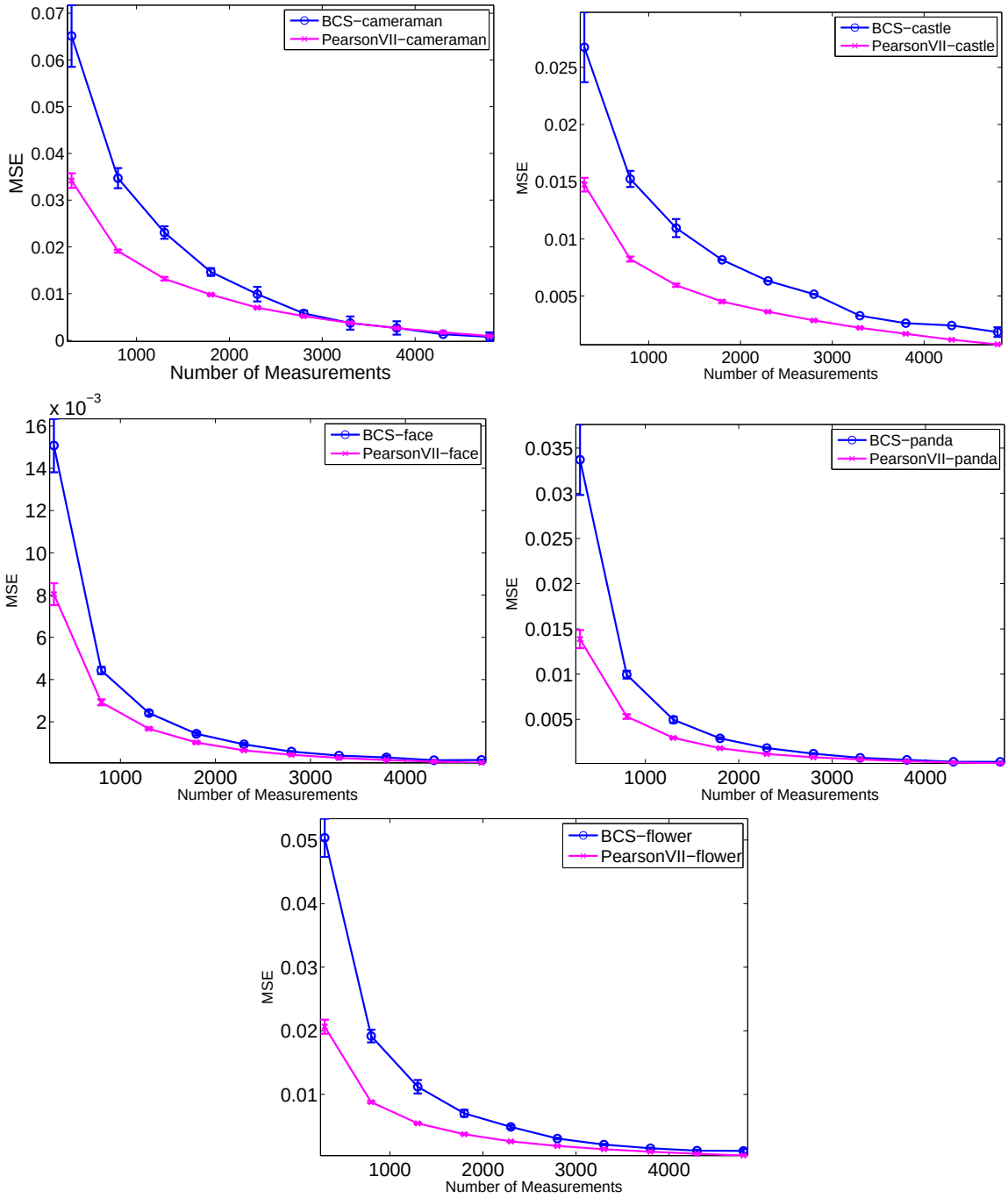


Figure 3.12: Comparative results on five different images, when $\boldsymbol{W}$ is CS-type and the number of observation is varied. The error bars represent one standard deviation from 10 independent repeats. We see the Pearson-based algorithm performs better than BCS in the under-determined regime.

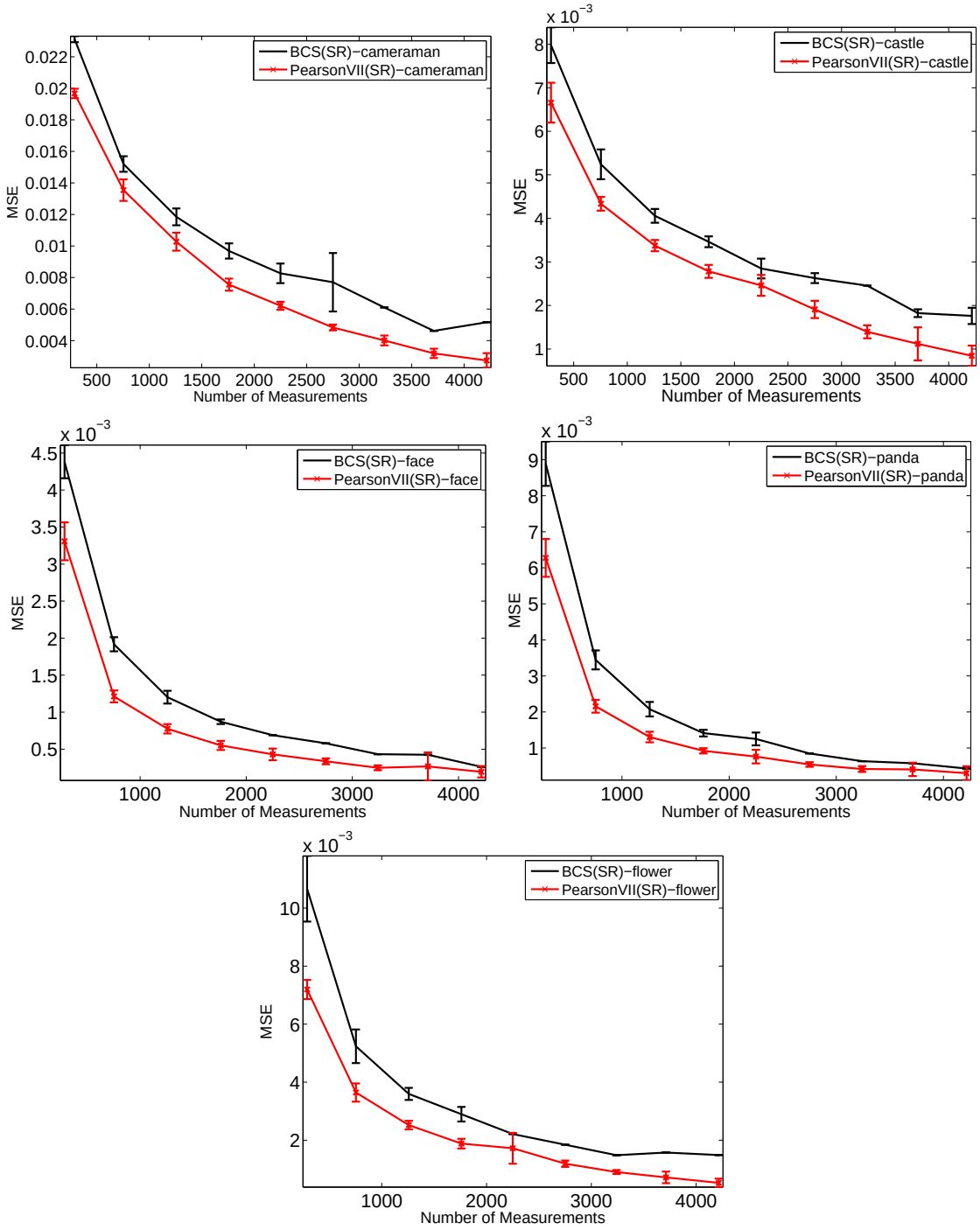


Figure 3.13: Comparative results on five different images, when $\boldsymbol{W}$ is SR-type that consists of blur and down-sampling, and the number of observation is varied. The error bars represent one standard deviation from 10 independent repeats. We see the Pearson-based algorithm performs better than BCS in the under-determined regime.

ities may or may not be present. In the following chapter, the joint probability model as described in this chapter will be used to investigate the alternative hyper-parameters estimation.

CHAPTER 4

INVESTIGATING ALTERNATIVE HYPER-PARAMETER ESTIMATION APPROACHES

In this chapter,[1] we formulate another different version of Pearson type VII image-prior, one that acts on the pixel level and another one that acts on the entire image. Having a single down-sampled and noisy version of low resolution frame, we aim to obtain the high resolution image using Maximum A Posteriori method. We compare the state-of-the-art of image-priors in super-resolution application and we discover that our image prior Pearson-MRF achieves the best performance in terms of quantitative measurement. This chapter also concentrated on recovering the high resolution image and the hyper-parameter using an alternative method. The contribution of this chapter is to show the assessment of the modelling power of the developed image-prior Pearson type VII using several alternative methods when it is compared to the existing image-priors.

4.1 Introduction and Overall Framework

In chapter 3, we proposed a robust density, the univariate version of Pearson type VII formulated as Markov Random Field in image recovery approach. Previuos research [90]

[1] A slightly shorter version of the works presented in this chapter have been accepted for publication in the *Intelligent Automation and Systems Engineering,* pp. 295-307, Lecture Notes in Electrical Engineering 103, (c) Springer, 2011.

based their comparisons on compressive matrices transformation. Due of curiosity, we formulate and examine the multivariate of Pearson type VII and compare it with the state-of-the-art approach using the classical super-resolution technique.

We will use the same observation and joint model as presented in chapter 3. Note that the type of $\boldsymbol{W}$ in this tested case is a product of blurring and down-sampling matrix of size [M×N], usually ill-conditioned matrix that models a linear blur operation and the down-sampling by row and column operator. The blur operation is a linear blur of a 2-dimensional convolution matrix from an averaging filter matrix of 3-by-3. The down-sampling operator discards some of the row and column elements of the matrix while others remain unchanged.

4.1.1 The multivariate Pearson type VII MRF

The probability density function given in equation (4.3) is a multivariate density that is instantiated here on the level of the full image. Meanwhile the previous version in equation (3.1) is parameterised differently but is the same as in equation (4.3) and we can see this as follows. Let denotes $\boldsymbol{C}$ as $(\boldsymbol{D}^T\boldsymbol{D})^{-1}\lambda$ and $m{=}(\nu + N)/2$. Now, look at the right hand side of equation (3.1) apart from the normalising constants:

$$\begin{aligned}[1 + \boldsymbol{u}^T\boldsymbol{C}^{-1}\boldsymbol{u}]^{-m} &= \left(\frac{1}{\lambda}[\lambda + \boldsymbol{u}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{u}]\right)^{-(\frac{\nu+N}{2})} \\ &= \underbrace{\left(\frac{1}{\lambda}\right)^{-(\frac{\nu+N}{2})}}_{\text{constant}}[\lambda + \boldsymbol{u}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{u}]^{-(\frac{\nu+N}{2})} \end{aligned} \tag{4.1}$$

Plugging the the equation in (4.1) into the full right hand side equation of (3.1), we will get:

$$Pr(\boldsymbol{u}) = \underbrace{\frac{\boldsymbol{\Gamma}(\frac{\nu+N}{2})}{\pi^{\frac{N}{2}}\boldsymbol{\Gamma}\left(\frac{\nu+N}{2} - \frac{N}{2}\right)}|(\boldsymbol{D}^T\boldsymbol{D})^{-1}\lambda|^{-\frac{1}{2}}\left(\frac{1}{\lambda}\right)^{-(\frac{\nu+N}{2})}}_{\text{constant}}[\lambda + \boldsymbol{u}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{u}]^{-(\frac{\nu+N}{2})} \tag{4.2}$$

From equation (4.2), we replace $\boldsymbol{u}$ with the variable of interest $\boldsymbol{z}$ and get the multivariate Pearson type VII as in equation (4.3):

$$\begin{aligned} Pr(\boldsymbol{z}) &\propto \left\{\boldsymbol{z}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{z} + \lambda\right\}^{-\left(\frac{\nu+N}{2}\right)} \\ &\propto \left\{\sum_{i=1}^{N}(\boldsymbol{D}_i\boldsymbol{z})^2 + \lambda\right\}^{-\left(\frac{\nu+N}{2}\right)} \end{aligned} \tag{4.3}$$

The univariate version devised in equation (3.7) may be regarded as having independent Pearson-priors on each neighbourhood-feature. Of course, we ought to point out that the neighbourhood features are not independent in reality. However, since each pixel only depends on four others, it may be a reasonable approximation. The version given in equation (4.3), in turn, does not allow such independence interpretation. Conversely, this can has the advantage that the spatial dependencies are not broken up, but more reliably accounted for. On the downside, the heavy-tail behaviour is more advantageous to have on the pixel level, i.e., on the distribution of neighbourhood features. Indeed, it is the distribution of neighbourhood features the one in which the edges from the image creates outliers. In turn, the multivariate Pearson-MRF is a density on images. Hence, its heavy-tail behaviour would be well suited to account for outlying or a typical images. Including both of these versions in our comparison will therefore uncover which of these pros or cons are more important for recovering quality high resolution images.

4.2 Experiments

We present two sets of a single-frame image super resolution experiments illustrating the performance of the hyper-parameters for testing the Pearson prior. We compare the state-of-the-art image priors such as Gaussian [32] and Huber [35]. The LR image is blurred by the unifrom blur matrix of size [3×3], down-sampled by factor 4 and contaminated by standard deviation of Gaussian noise of 0.001, 0.01, 0.05 and 0.1. All images are in size [100×100] and the pixel intensities are scaled to interval [-0.5, 0.5]. The initial guess is

initialized with Gaussian-MRF with σ^2/λ set to 1 and was used as a starting point for the recovery algorithm.

In this section, we present the manual selection and cross validation to address the issue of parameter selection for estimating λ and ν. For the automated estimation (referring to cv), we initialised the initial $\boldsymbol{z}$ with a product of the inverse transformation matrix $\boldsymbol{W}$ and the low resolution $\boldsymbol{y}$. We employed a conjugate gradient type method, which requires the gradient vector of the objectives. Before we proceed on presenting all the alternative methods that we have developed, a summary of the findings based on two types of cross validation methods (section 4.3.1 and 4.3.2), manually tuned (section 4.4) and manual method using a grid search (section 4.5) are presented in table 4.1.

Table 4.1: The table presents a summary of the findings on estimating the hyper-parameters using alternative methods.

Method	**Optimal λ**	**Optimal ν**	**Outcome**
Hold out estimation	0.001	0.101, 0.151, 0.801, 09.51	Good recovery as shown in figure 4.2
k-folds cv	0.951	0.001	Good recovery as shown in figure 4.5
Manually tuned	0.1 to 100 (i.e: 1)	1-10 (i.e: 0.05)	Good recovery as shown in figure 4.9
Grid search based on 50 images (brute-force algorithm)	0.0013	2	Bad recovery as shown in figure 4.13
Grid search based on 20 images (brute-force algorithm)	0.00012	0.8	Bad recovery in figure 4.13

4.3 Cross Validation

This section is devoted to investigating alternative methods for estimating the hyper-parameters. To automate the search, we developed cross validation method. Validation is done by computing the minimum error of the mean square error on the similarity of the observed data $\boldsymbol{y}$, with the model $\boldsymbol{Wz}$. On the other hand, k-folds cross validation was developed for the classical transformation because its component is sparse[1] and this

[1]a matrix populated primarily with zeros

meant that algorithm could be executed faster. To reduce variability, five rounds of cross validation were performed using different folds, and the validation results were averaged over the rounds. Both forms of the algorithms are as follows in Algorithms 2 and 3. Indeed, in the described approach as shown in figures 4.1 and 4.4, the algorithm is less expensive and a more precise search space has been tested.

4.3.1 Hold out estimation

The performance of the image recovery of high resolution depends on how good the selection value of the hyper-parameters in image-prior. In this section, we developed hold out estimation and the algorithm is described as follows and the comparison result with the state-of-the-art methods is displayed in figure 4.3.

Algorithm 2 : Hold out estimation

1: **Goal:** To find optimal values for ν and λ by training a model using the training data set and its minimum error using the validation data set.
2: **Inputs:** training data, validation data, number of k-groups, ν and λ range, variance σ^2
3: **Outputs:** optimal ν, optimal λ, optimal error
4: Randomize and divide the data set into two groups: 5% for validation and the remainder for the training set.
5: **for** $i = 1$ **to** $length(\nu)$ **do**
6: **for** $j = 1$ **to** $length(\lambda)$ **do**
7: Minimize wrt $\boldsymbol{z}$ using training set.
8: Compute performance: mean(($\boldsymbol{y}$(validate)- $\boldsymbol{W}$(validate) $\times$ $\boldsymbol{z}$(training))^2)
9: Record the performance matrix error.
10: **end for**
11: **end for**
12: Find ν and λ that belong to the minimum error.

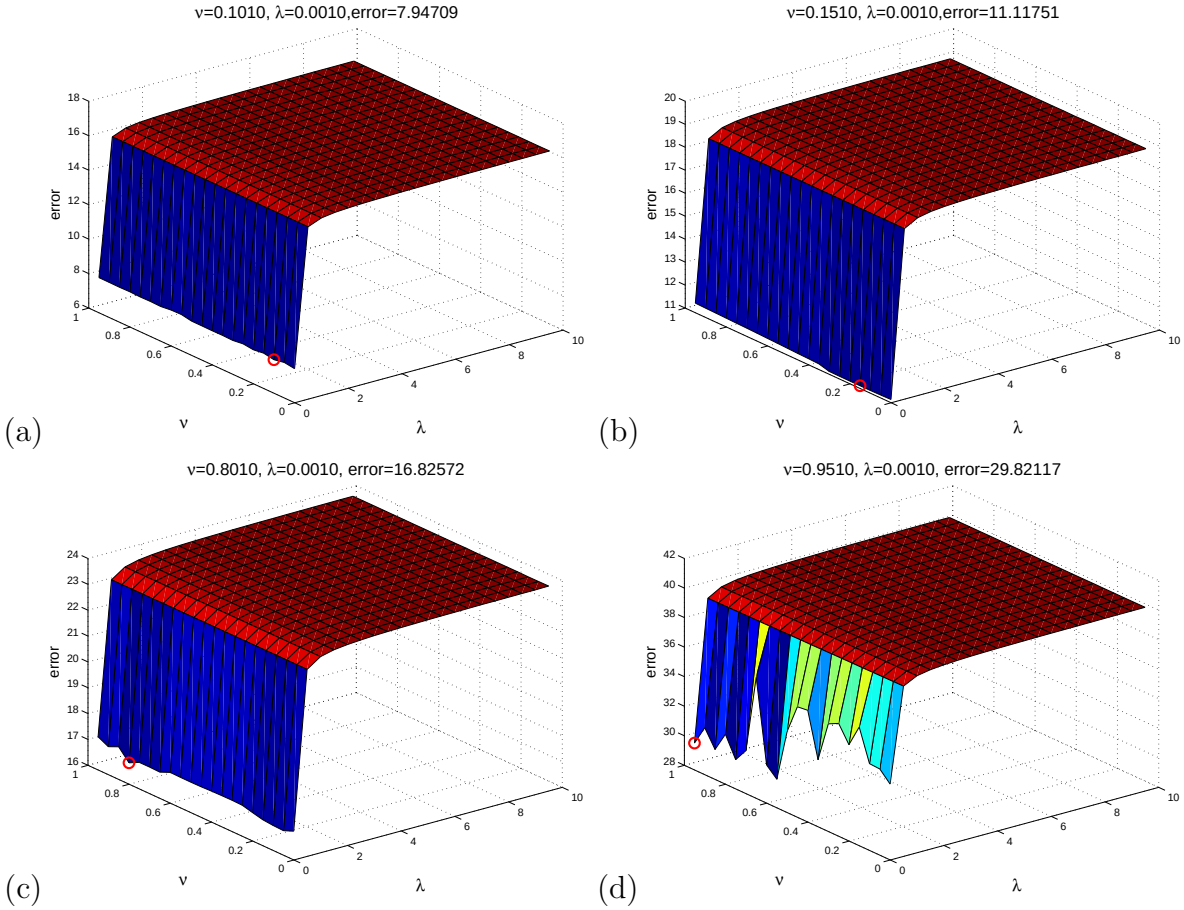


Figure 4.1: Examples of 3-dimensional plot varying ν, λ and its mean squared error was computed at 5% of '`cameraman`' (5600 pixels) from random transformation to 4000 pixels. Additive noise with σ^2: (a) 0.005, (b) 0.01, (c) 0.05 and (d) 0.1 . We demonstrate the search space for ν from range 0.001 to 1 (with the interval 0.05) and λ from range 0.001 to 10 (with the interval 0.5) using 95% data set. This range was chosen based on the best manual selection range that we achieved for the '`cameraman`' image. Optimal values for ν were found best (a) $\nu = 0.101$, (b) ν=0.151, (c) ν=0.801, (d) ν=0.951 and λ remained its optimal for every level of noise, $\lambda = 0.001$. This experiment was performed to automate the hyper-parameters without gaining access to the true image and was able to recover the image well.

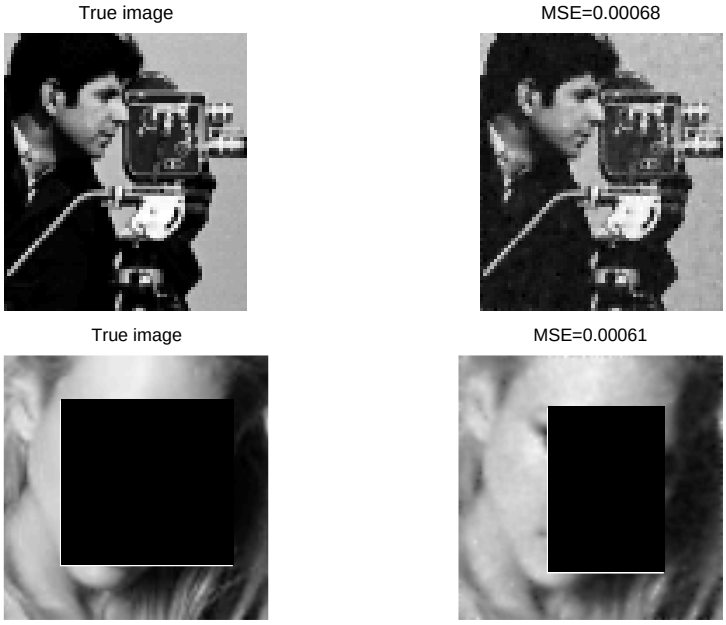


Figure 4.2: Examples image recovery of '`cameraman`' (5600 pixels) from random projection to 4000 pixels in the top subplot and a '`woman's face`' (10000 pixels) from random projection to 3000 pixels in the final subplot. ν and λ are found using hold out estimation and both additive noise, σ=0.05

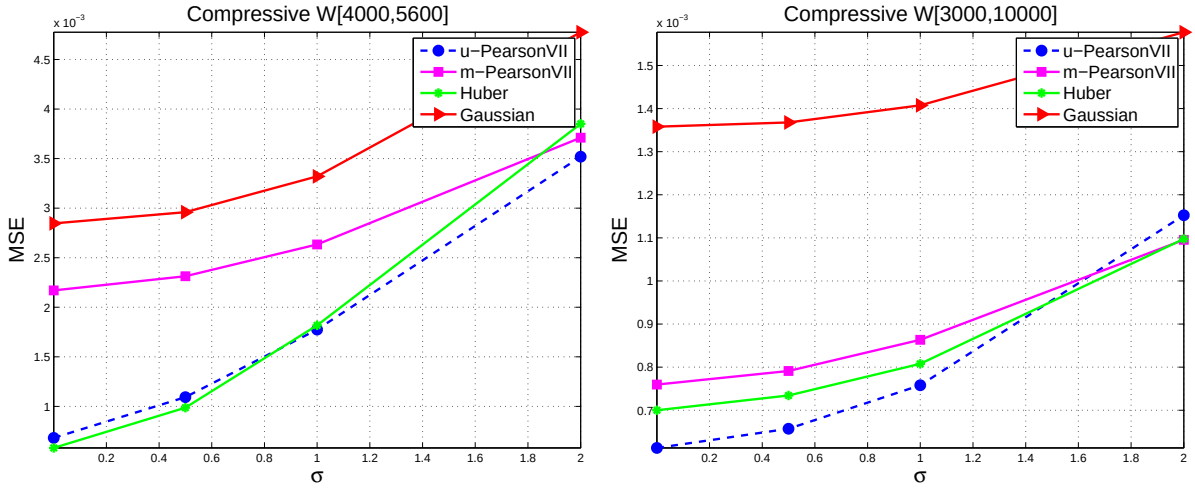


Figure 4.3: Comparative MSE performance for underdetermined system for '`cameraman`' and '`woman's face`', varying four levels of noise using the best values of hyper-parameter for every image-prior found using hold out estimation. These results demonstrate that our proposed prior, u-Pearson type VII is competitive with the state-of-the-art approach on that type of data considered here (i.e: low observations, 4000 and 3000 pixels, distorted data, random transformation).

4.3.2 *k*-fold cross validation

To asses the goodness of the proposed method, Pearson MRF estimation results are compared with image enhancement state-of-the-art methods in [32, 35, 67] using the quantitative measurement, mean square error. Our proposed algorithm for parameter estimation illustrates the performance result over 5-folds cross validation. To compare the quality of the recovered image across the four image-priors (e.g: univariate Pearson MRF, multivariate Pearson MRF, Huber MRF and Gaussian MRF), we used automated estimation of the hyper-parameters. These results are presented in figure 4.6 and we can see that the univariate Pearson type VII based MRF can achieve state-of-the-art performance and give a competitive solution to Huber MRF across the four levels of noise. Finally we also illustrated two set of image recovery in figure 4.5.

Algorithm 3 : k-fold cross validation for estimating ν and λ

1: **Goal:** To find optimal ν and λ by training a model using the training data set and its minimum error using 5-folds cross validation.
2: **Inputs:** training data, validation data, number of k-groups, ν and λ range, variance σ^2
3: **Outputs:** optimal ν, optimal λ, optimal error
4: Randomize and divide the data set into k-groups.
5: **for** $k = 1$ **to** $k - groups$ **do**
6: validate = find(group==k)
7: training = find(group$\tilde{=}$k)
8: **for** $i = 1$ **to** $length(\nu)$ **do**
9: **for** $j = 1$ **to** $length(\lambda)$ **do**
10: Minimize wrt $\boldsymbol{z}$ using training set.
11: Compute the performance found using the k-th set: mean((**y**(validate)-**W**(validate) $\times$ **z**(training))^2)
12: Record the performance matrix error.
13: **end for**
14: **end for**
15: Report the mean error over all k test sets.
16: **end for**
17: Find ν and λ that belongs to the minimum 5-folds error value.

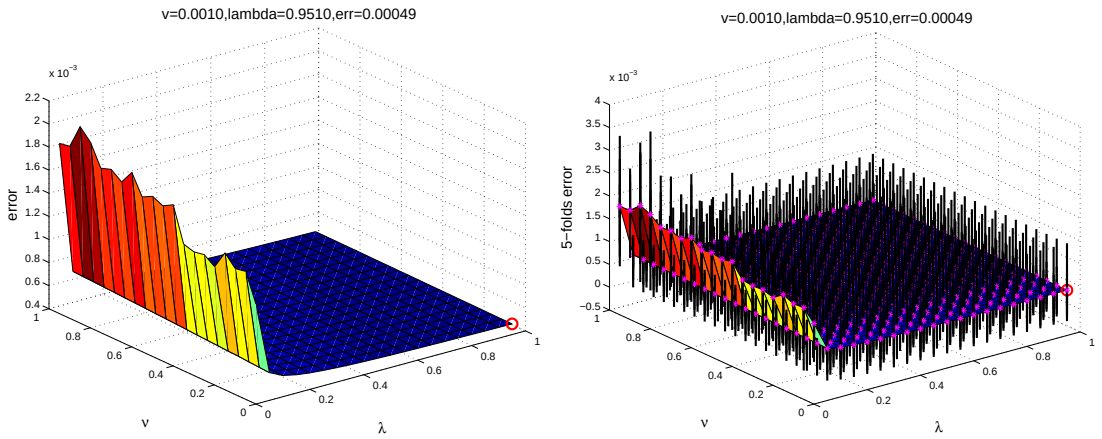


Figure 4.4: Example of mean error over all k test sets (left) and mean and standard deviation over 5-folds repetition (right) for variance= 0.005 using classical transformation matrix in [4] for '`Phantom`' [100×100] image. This plot illustrates the performance of mean square error at 5% data set.

Figure 4.5: Examples image recovery of '`cameraman`' and '`panda`' images (10000 pixels) from blurred and down-sampled to 2601 pixels and additive noise with σ^2 = 1e-3 using univariate Pearson type VII based MRF.

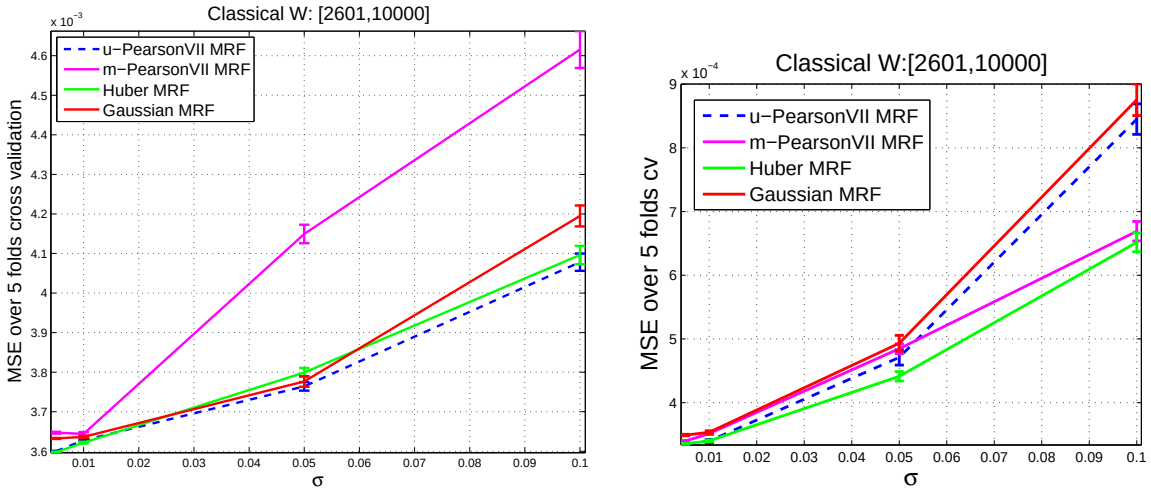


Figure 4.6: Comparative MSE performance for under-determined system for '`cameraman`' and '`woman's face`', varying four levels of noise using the best values of hyper-parameter for every image-prior using 5-folds cross validation technique. The error bars are over 10 independent trials. The experiments were performed using conventional transformation which consists of blurred and down-sampled operators. Pearson prior maintains its good performance on the left figure for every level of noise. However, it does not seem to recover well for the second image, especially for greater noise. The right plot illustrates that this prior is not always best for every data and condition. Nonetheless, our proposed image-prior is still competitive with the state-of-the-art method for smaller noise.

4.4 Manually Tuned

In this section, two types of manual settings are examined. First, we used manual tuning and secondly using a grid search for estimating the hyper-parameters. The manual tuning can be categorized into two different objective functions to find out the optimal values for both parameters. The first objective function finds the lowest mean square error from the search space of ν= 5e-20 - 50000 and λ from 0.001 - 10000. The second objective function finds the maximum score of the log probability density function of Pearson type VII as the optimal value. The performance of the image recovery of high resolution depends on how good selection value of hyper-parameters in image-prior. We also want to compare the outcome from the manual search with the automated version.

4.4.1 Pseudocode

- Goal: The optimal values of hyper-parameters λ and ν are manually tuned to get the best(lowest) mean square error(MSE).
- Search range: ν=5e-20, 5e-18, 5e-15, 5e-13, 5e-10, 5e-8, 5e-7, 5e-5, 5e-3, 5e-2, 5e-1, 1, 10, 100, 1000, 10000 and 50000, λ = 0.001,0.01,0.1,1,10, 10,100,1000 and 10000
- Under-determined problem where $\boldsymbol{W}$ of size [2500,10000] is tested.
- Outputs: ν_{opt} range from 1-10, λ_{opt} range from 0.1 to 100

From the observation using the constructed blur and down-sampling matrix $\boldsymbol{W}$, we found practical range of λ and ν. The results are presented in figure 4.7. Too small λ (0.001) and ν values reduce the effect of prior and the solution approaching the Maximum Likelihood, whilst too big of λ such as 10000 will blur the edges. The overall performance of the recovered image depends strongly on the selection of λ.

We can conclude that ν can be fixed into a practicable range (i.e:1-10) so that the iteration could terminate earlier and the λ is found best from 0.1 to 100. Two set of images ('`cameraman`' and '`panda`' images) are examined to analyse the best performance based on manual tuning. Figure 4.8 shows the variation performance varying several λ and figure 4.9 shows the effect of the bad and good selection of hyper-parameters. From figure 4.7, we also investigate on how another image behaves from that stable range. Figure 4.10 shows those values still practicable on selected range. Besides, the performance of several level of noise is investigated using one of the stable range of ν and the results are presented in figure 4.11.

4.5 Manual Method Using a Grid Search

Manual method using a grid search allows us to find the optimal value on a larger scale. This could be achieved by finding the optimal value using the brute-force algorithm [91],

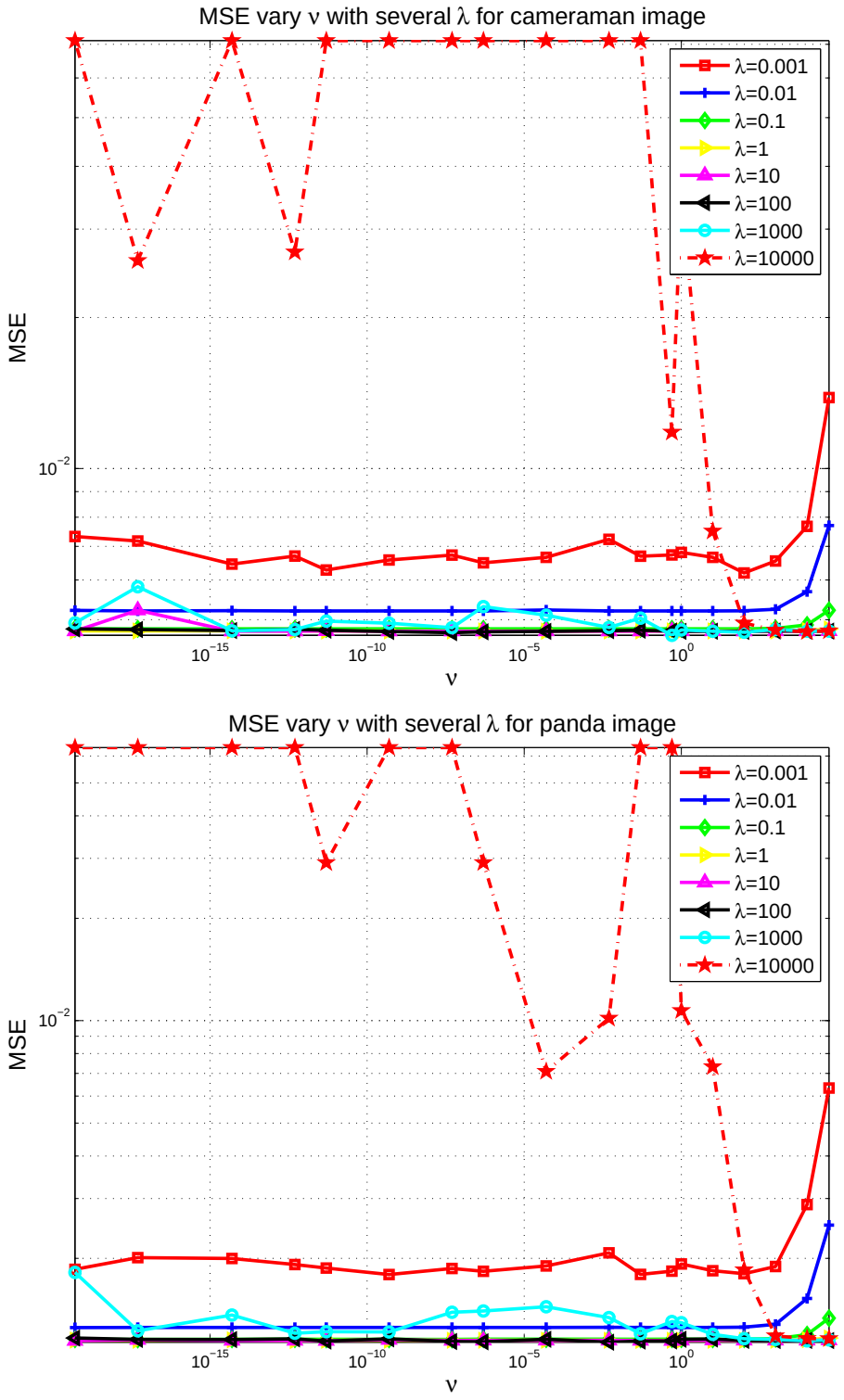


Figure 4.7: Top: Test image of '`cameraman`', bottom: test image of panda are used to inspect the best value of hyper-parameters by computing the MSE performance varying several fixed λ and the algorithm is provided the true noise variance, σ^2=0.001. The range of ν are 5e-20, 5e-18, 5e-15, 5e-13, 5e-10, 5e-8, 5e-7, 5e-5, 5e-3, 5e-2, 5e-1, 1, 10, 100, 1000, 10000 and 50000.

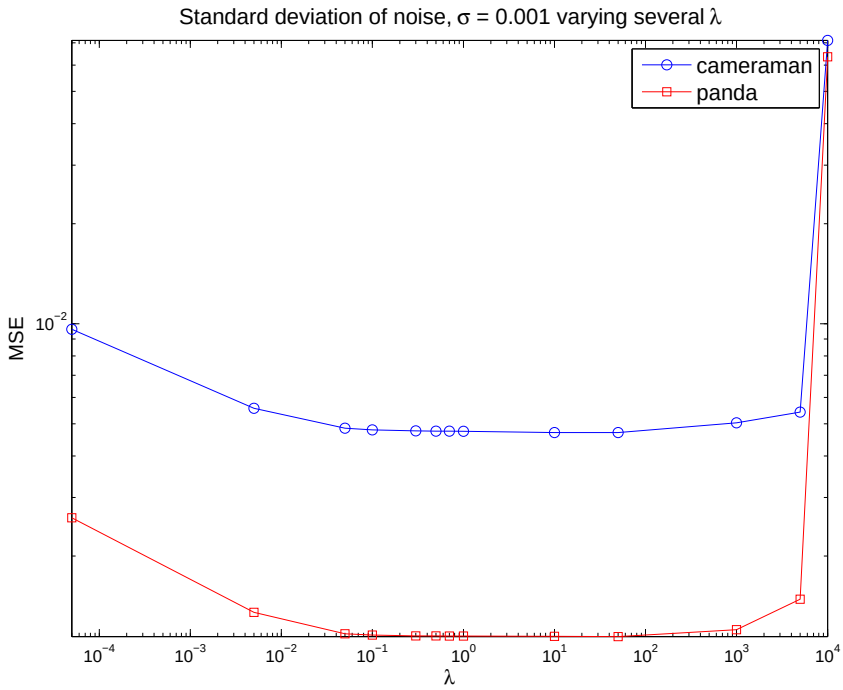


Figure 4.8: MSE measurement varying λ where the ν is fixed to 0.05 and this value is found one of the best from manually search for two set of different images.

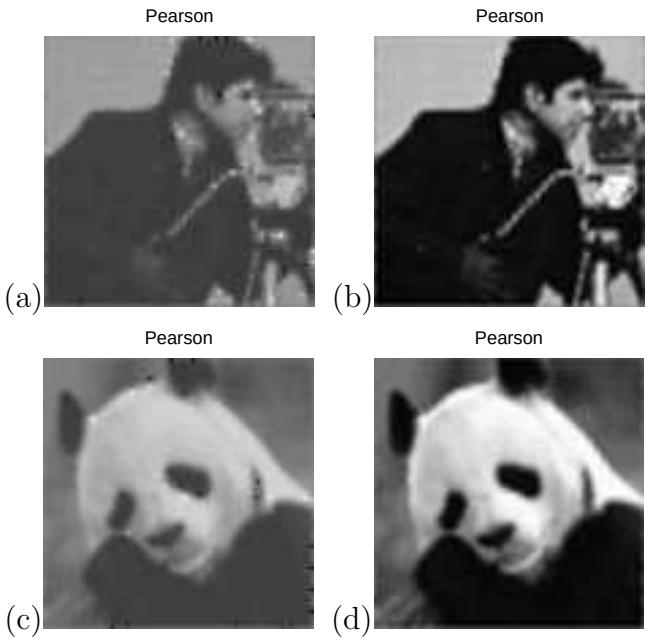


Figure 4.9: Subplot (a) and (c) show examples of bad image recovery when $\lambda = 0.01$ and $nu = 5e - 15$). Subplot (b) and (d) represent good image recovery when λ=1 and ν=0.05 using manual tuning hyper-parameters. The problem is under-determined where $\boldsymbol{W}$[2500,10000].

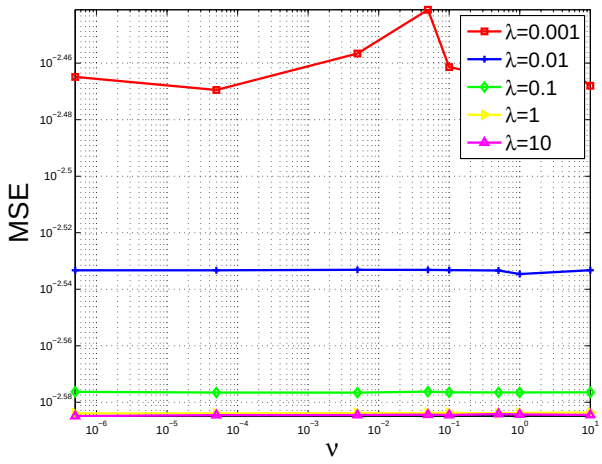


Figure 4.10: Test image of a '`ladybug`' using the stable range of those value of hyper-parameters by computing the MSE performance varying several fixed λ and the algorithm is provided the true noise variance, σ^2=0.001. The range of ν are from 5e-7 to 1.

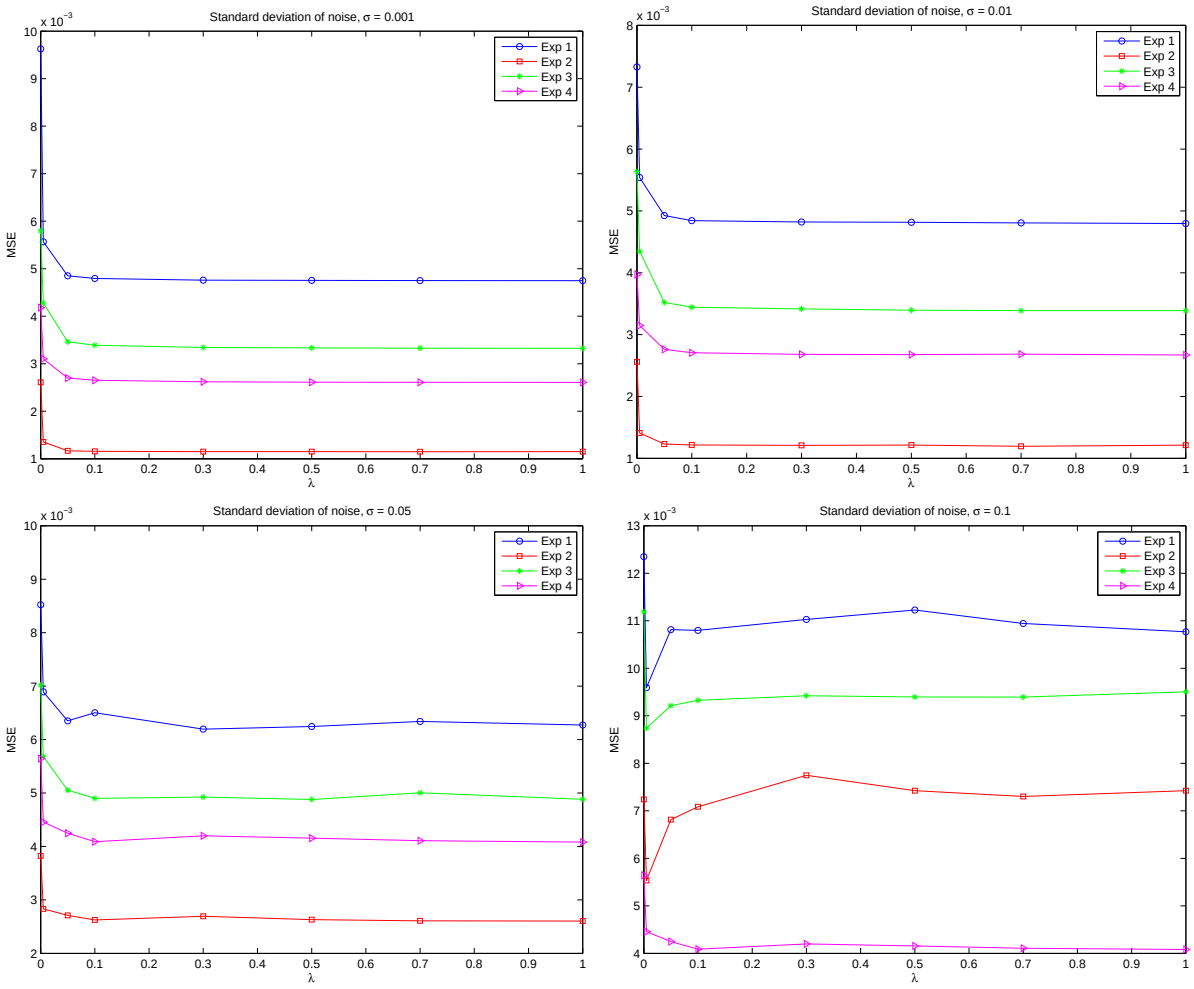


Figure 4.11: Test set on a different level of noise for four type of images varying several fixed λ using one of the optimal value found ($\nu = 0.05$). Different set standard deviation of additive noise, from top left: σ=0.001, σ=0.01 and from bottom left: σ=0.05, σ=0.1

although it requires long computation time. It searches exhaustively in a range of possible values by trying all combinations of ν and λ on a grid search. We consider 100000 possible values in this algorithm. All these combinations will generate in a 1-dimensional density form.

4.5.1 Pseudocode

The pseudocode based on the brute-force algorithm is described as follows:

- Goal: To estimate optimal value of hyper-parameters λ and ν using brute-froce algorithm
- Inputs: vectorised image($\boldsymbol{z}$), number of rows(r) and columns(c) of the image($\boldsymbol{z}$), $\boldsymbol{D}$ matrix, neighbourhood features($\boldsymbol{x}$), ν range, λ range and Pearson type VII density function
- Search range: ν = 0.0001 - 100, λ = 1e-13 - 10
- Outputs: λ_{opt}, ν_{opt} and maximum $score_{pdf}$
- Algorithm:

  1. Construct $\boldsymbol{D}$ matrix according to the image($\boldsymbol{z}$) size.
  2. Compute neighbourhood features, $\boldsymbol{x}$=$\boldsymbol{D} \times \boldsymbol{z}$
  3. Initialise the variables ν and λ range for 100000 possible values.
  4. Initialise $\nu_{opt}, \lambda_{opt}, score_{opt}$.
  5. Compute the sum of Pearson log function.

     for i=1:length(ν)

     for j=1:length(λ)

     $$score_{pdf}(i,j) = sum[\log\left(\Gamma\left(\tfrac{1+\nu(i)}{2}\right)\right) \times \log(\lambda(j))$$
     $$-\left((\tfrac{1+\nu(i)}{2}) \times \log(x^2+\lambda(j))\right) - \log\left(\Gamma\tfrac{\nu(i)}{2}\right) - \tfrac{1}{2}\log(\pi)]$$

6. Obtain the maximum $score_{opt}$

for i=1:length(ν)

for j=1:length(λ)

if $score_{pdf}(i,j) > score_{opt}$;

$\nu_{opt} = \nu$(i);

$\lambda_{opt} = \lambda$(j);

$score_{opt} = score_{pdf}(i,j)$;

end

end

end

7. Obtain the optimal values for ν and λ by finding the maximum score on the grid search of 3-dimensional plot.

4.5.2 Results and discussion

The optimal value is obtained from the probability density function (PDF) which generates the highest score. In other words, the highest score of the PDF is the most probable solution that matches the histogram of $\boldsymbol{Dz}$. We obtained the optimal value for these hyper-parameters by computing the average of 50 natural images with various sizes. The test set images are taken from Matlab and internet database (Google[1]). Unfortunately, the optimal values for ν=2 and λ=0.0013 obtained from the average of 50 natural images did not recover the images well when we employed these values. An inspection has been done and we discovered almost half of the randomly chosen natural images have the peculiar shapes of the neighbourhood feature as shown in figure 4.14.

Once we eliminated the unusual natural images which are distant from the rest of the data, we computed the average for those hyper-parameters again and the results are

[1]http://www.google.com

presented in figure 4.12. Then, we inserted these values (λ=1.2e-4, ν =0.8) into our recovery algorithm again to estimate the high resolution images. However, this modified solution did not recover as good as the manual selection, and we found that it is worse than the average of the 50 images. This is because the optimal λ now is decreasing and from the previous tuning method, we learned that too small λ will reduce the effect of the image-prior.

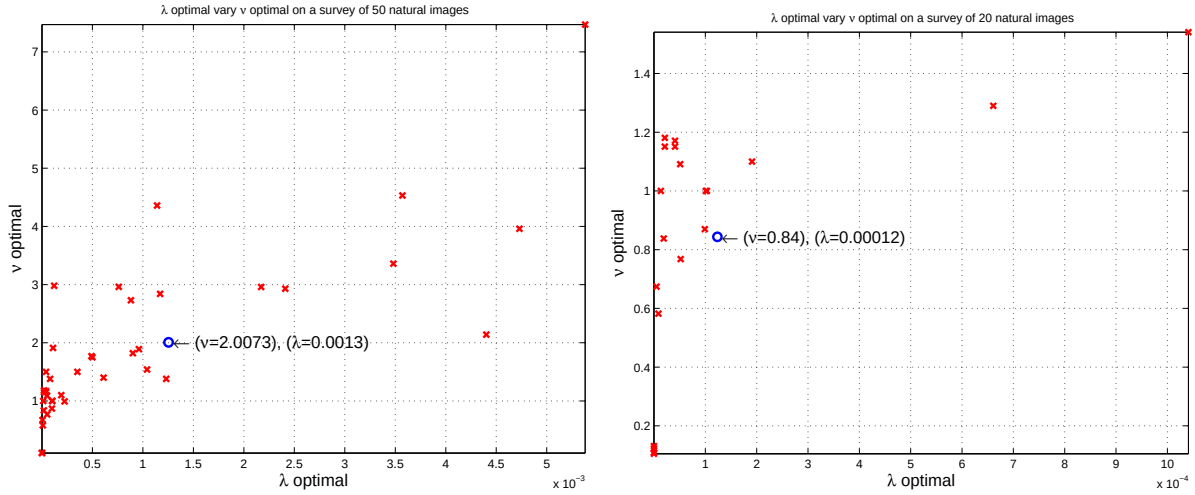


Figure 4.12: On the left, the optimal values for λ and ν obtained from 50 natural images of neighbourhood features. On the right plot, the optimal values for λ and ν obtained from 20 natural images of neighbourhood features. The blue circle indicates the average of the optimal values from a different set of natural images.

When we inspected the reason behind the failure, we noticed that the density itself is a non-convex solution. This function has several local minima and it is possible for the search algorithm to get stuck in the early local minima. To solve this, the first attempt was to repeat the experiments for several times (i.e:20) with the hope that one of the optimisation will reach the global optimum. Unfortunately, the results turned out to be almost the same all the time. Then we continued with the second attempt by carrying out this grid search using the convex density function such as Gaussian and Huber on the '`cameraman`' image. We discovered that optimal result for Gaussian prior is λ=1, and that the value we obtained using the manual selection ($\lambda = 0.8$) is very close and is capable of recovering the images well.

In summary, the average of optimal values from several images or even the optimal

(a)  (b) 

Figure 4.13: Subplot (a) shows the recovered image using the average optimal values based on 50 images and subplot (b) displays the recovered image using the average optimal values based on 20 images. Both results are based on manual method using a grid search. The problem is under-determined where $\boldsymbol{W}$[2500,10000].

values from a specific image still does not produce a better result for the Pearson type VII case. The recovered image is not as good as the results of the algorithm manually tuned because this method does not have access to the ground truth image. However, both optimal values of ν are still found in a good range but both values of λ are under-estimated which makes the image recovery worse.

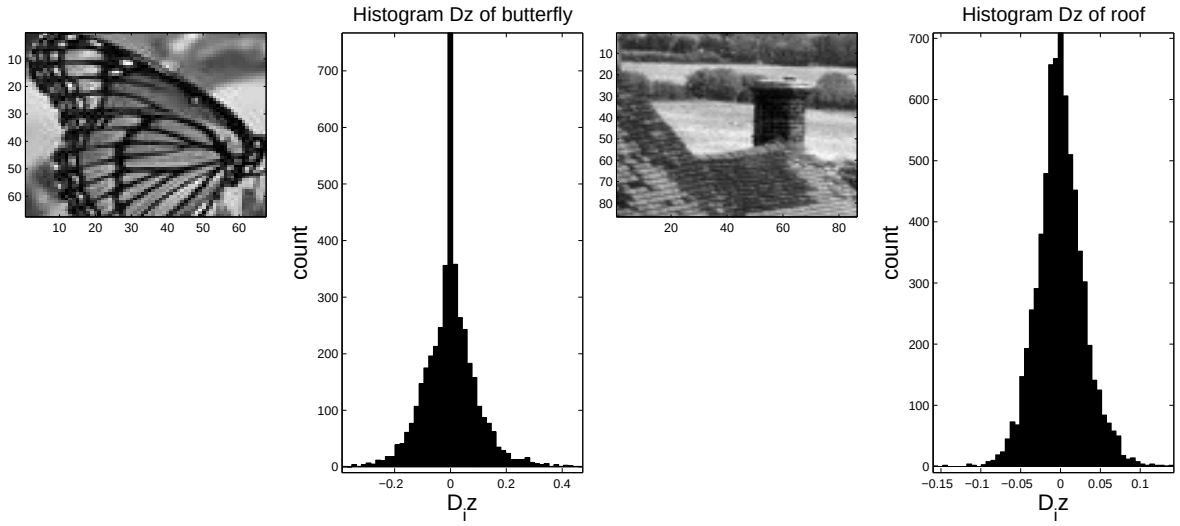


Figure 4.14: Examples of the peculiar histograms of the distribution of neighbourhood features $\boldsymbol{D}_i\boldsymbol{z}, i = 1, \cdots, N$.

4.6 Conclusions

In this chapter, we formulated the multivariate of Pearson MRF image-prior, and conducted a comparative experimental study among state-of-the-art methods of image-prior from a single noisy version of low resolution image. We discovered that the quality of the recovered image performed better on the pixel level when we employed our univariate Pearson MRF in comparison to the performance of the multivariate Pearson image-prior. We demonstrated that our proposed prior, univariate Pearson Type VII MRF is likewise comparable with Huber MRF for all levels of noise and we assessed on four different images. The quality of the recovered image is always consistent although it has several local optima.

We also developed some of the alternative methods and compared it with the optimisation approaches in chapter 3. Our main motivation is to avoid the crucial initialisation of the hyper-parameters in the conjugate gradient method. Secondly, we compare how the alternative methods such as cross validation, hold-out estimation and manual tuning estimation to recover the signals. The alternative way using cross validation obtained a good recovery and is superior in comparison to the existing image-priors such as Huber and Gaussian MRF. Although it is quite simple to implement, the challenges is to provide it with a good search range for the hyper-parameters. Otherwise, the signal recovery will be a time consuming process.

CHAPTER 5

SINGLE-FRAME SIGNAL RECOVERY USING A SIMILARITY-PRIOR

This chapter[1] presents a similarity-prior framework with the aid of extra information. We consider the problem of signal reconstruction from noisy observations in a highly under-determined problem setting. Most of previous work does not consider any specific extra information to recover the signal. Here we address this problem by exploiting the similarity between the signal of interest and a consecutive motionless frame. We incorporate this extra information of similarity that is available into a probabilistic image-prior based on the Pearson type VII Markov Random Field model. Results on both synthetic and real data of MRI images demonstrate the effectiveness of our method in both compressed setting and classical super-resolution experiments.

5.1 Introduction to Similarity-Prior

As mentioned in previous chapters, the main focus of this thesis is to consider the problem of signal reconstruction from noisy observations in a highly under-determined problem setting. In this chapter, we tackle the problem using more specific prior information, namely the similarity to a motionless consecutive frame as the additional input for recovering the

[1]A shorter version of the work presented in this chapter has been accepted for publication in Springer Proceedings in Mathematics for Mathematical Methodologies in Pattern Recognition and Machine Learning, ICPRAM'2012 special issue.

signals of interest in a highly under-determined setting. This has real applications e.g. in medical imaging where such frames are obtained from several scans. Recent work by Vaswani and Lu [92] found the average frame from those scans to be useful for recovery.

In principle, the more information we have about the recovered signal, the better the recovery algorithm is expected to perform. This hypothesis seems to work in [92, 93], however they require us to tune the free parameters of the model manually, and Giraldo *et al.* [93] mentioned that the range of parameter values was not exhaustively tested. Vaswani and Lu [92] also mentioned that they were not able to attain exact reconstruction using fewer measurements than those needed by compressed sensing (CS) for a small image. In contrast, we will demonstrate a good recovery from very few measurements using a probabilistic model that includes an automated estimation of its hyper-parameters.

Related works on sparse reconstruction gained tremendous interest recently and can be found in [2, 6, 94, 95]. The sparser a signal is, in some basis, the fewer random measurements are sufficient for its recovery. Somewhat related, the recent work by Lu and Vaswani [96] exploits partial erroneous information to recover small image sequences. However, previous research does not consider any specific extra information that could be used to accentuate the sparsity, which is our focus.

This chapter is aimed at taking these ideas further through a more principled and more comprehensive treatment. We study the case when the observed frame contains too few measurements, but with an additional motionless consecutive scene in high resolutions is provided as an extra input. This assumption is often realistic in imaging applications. Our aim is to reduce the requirements on the number of measurements by exploiting the additional similarity information. To achieve this, we employ a probabilistic framework, which allows us to estimate all parameters of our model in an automated manner. We conduct extensive experiments that show how our approach not only bypasses the requirement of tuning free parameters but is also superior to a cross validation method in terms of both accuracy and computation time. Results on both synthetic and real data of MRI images demonstrate the effectiveness of our method in both compressed setting

and classical super-resolution experiments.

5.2 Image Recovery Framework with Similarity Information

In this section, the observation model, joint model, similarity-prior and its approximation are presented.

5.2.1 Observation model

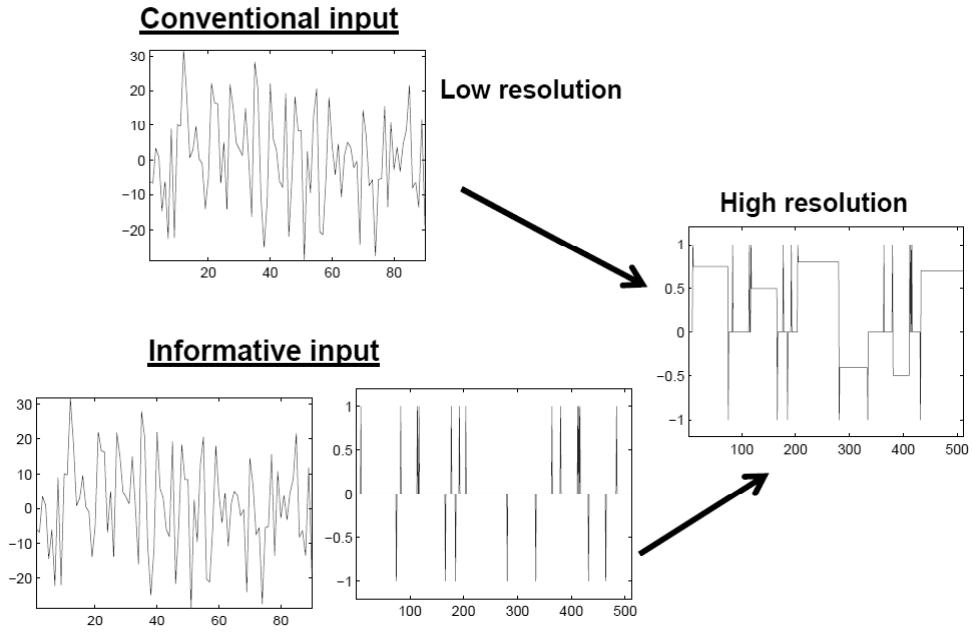
A model is good if it explains the data. The following linear model has been used widely to express the degradation process from the high resolution signal $\boldsymbol{z}$ to a compressed or low resolution noisy signal $\boldsymbol{y}$ [67, 36, 35, 32]:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\eta} \tag{5.1}$$

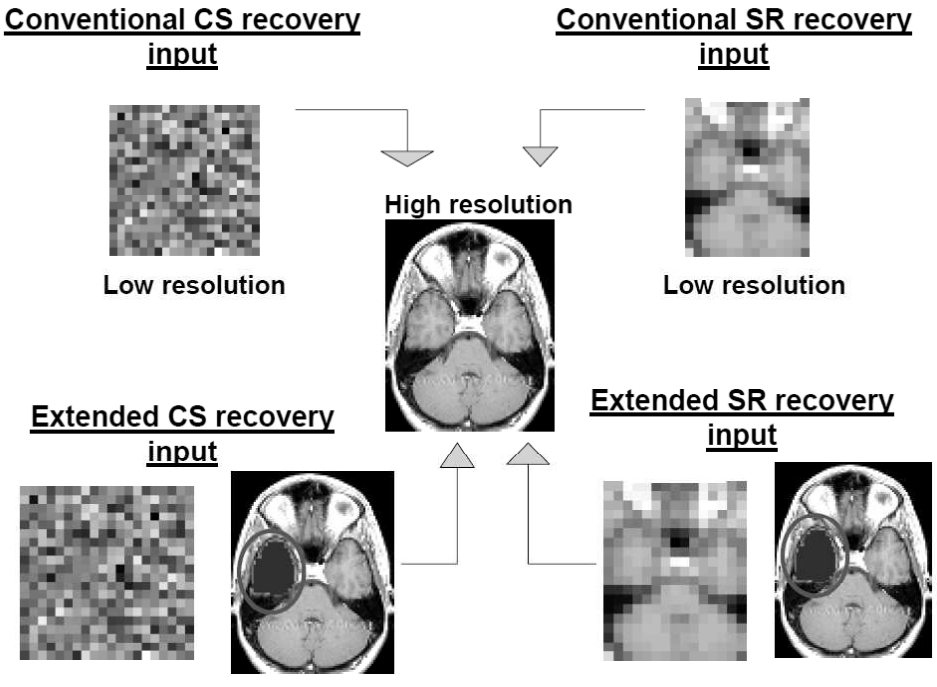
where the high resolution signal denoted by $\boldsymbol{z}$ is an N-dimensional column vector and $\boldsymbol{y}$ is an $M \times 1$ matrix representing the noisy version of the signal, with $M < N$.

5.2.2 The similarity-prior

The Pearson type VII MRF prior presented in [90] is used to the construction of a generic prior for images. $\boldsymbol{D}$ matrix is a $N \times N$ size that encodes the cardinal neighbour relationship. The elements in $\boldsymbol{D}$ matrix is filled by the entries defined in equation (3.5). This $\boldsymbol{D}$ matrix is multiplied by a vector $\boldsymbol{z}$ of size $N \times 1$ (i.e: $\boldsymbol{D}\boldsymbol{z}$). $\boldsymbol{D}$ makes the signal sparse because the intensity difference between each pixel in $\boldsymbol{z}$ with the average of its cardinal neighbours is close to zero. In this chapter, we aim to recover both 1D and 2D signals



(a)



(b)

Figure 5.1: An illustration of a signal recovery process from a noisy version of low resolution for 1D signals in subplot (a) and 2D signals in subplot (b) with the aid of informative input.

using the additional similarity information. We define the entries of $\boldsymbol{D}$, i.e d_{ij} as follows:

$$d_{ij} = \begin{cases} 1 & \text{if } i = j; \\ -1/\mathbb{N} & \text{if i and j are neighbours;} \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathbb{N}$ denotes the number of cardinal neighbours; 4 for images and 2 for 1D signals.

In general, the main characteristic of any natural image is a local-smoothness. This means that the intensities of neighbouring pixels tend to be very similar. Hence, $\boldsymbol{Dz}$ will be sparse. Therefore, we propose an enhanced prior to exploit more information that leads to more sparseness. By employing the given additional information of the consecutive image or signal, we will employ the difference $\boldsymbol{f}$, between the recovered image $\boldsymbol{z}$, and the extra information denoted as $\boldsymbol{s}$. Obviously the more pixels $\boldsymbol{z}$ and $\boldsymbol{s}$ have in common, the more smooth their difference will be. Figure 5.2 shows a few examples of histograms of the neighbourhood features $\boldsymbol{Dz}$ from real images, where the sparsity is entirely the consequence of the local smoothness. We also show the histograms of the new neighbourhood features $\boldsymbol{Df}$ that includes the additional similarity information. We see the latter (e.g. $\boldsymbol{D}_i\boldsymbol{f}$) is a lot sparser than the former (e.g. $\boldsymbol{D}_i\boldsymbol{z}$). Then we can formulate

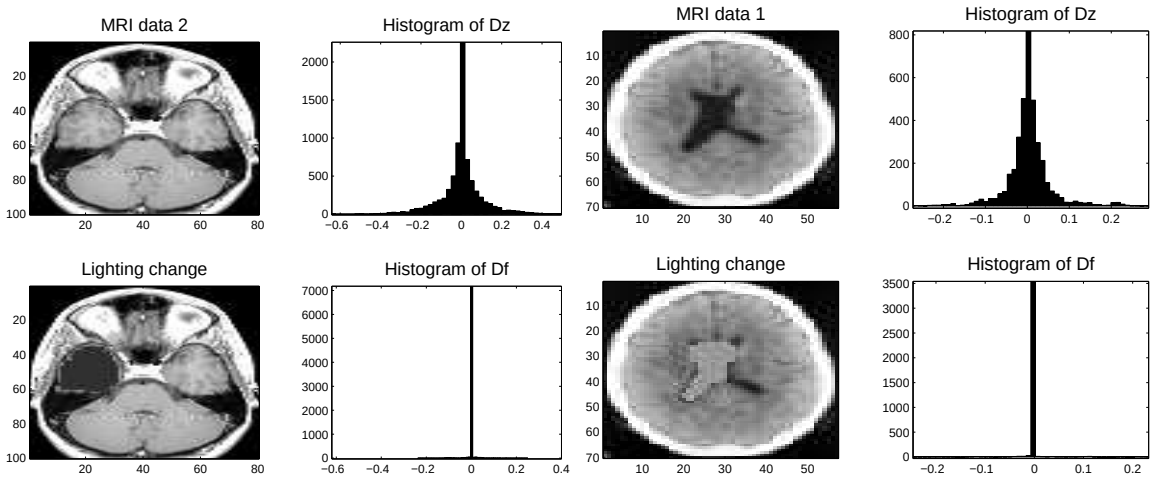


Figure 5.2: Example histograms of the distribution of neighbourhood features $\boldsymbol{D}_i\boldsymbol{z}$ in the top subplot, and $\boldsymbol{D}_i\boldsymbol{f}$ in the last subplot where $i = 1, ..., N$ from a MRI real data.

the i-th feature in a vector form, with the aid of the i-th row of this matrix (denoted $\boldsymbol{D}_i$)

as the following:

$$f_i - \frac{1}{\mathbb{N}} \sum_{j \in \mathbb{N} \text{ neighb}(i)} f_j \;\; = \;\; \sum_{j=1}^{N} d_{ij} f_j = \boldsymbol{D}_i \boldsymbol{f} \tag{5.2}$$

Since our task is to encode the sparse property of signals, therefore this feature is useful when the difference between a pixel of the difference image f and the average of its neighbours is close to zero, almost everywhere except the edges of the dissimilarity areas. Replacing f with the notation $(\boldsymbol{D}_i(\boldsymbol{z} - \boldsymbol{s})$ and plugging this (5.2) into the Pearson-MRF density, we have the following prior that we refer to as a *similarity-prior*:

$$Pr(\boldsymbol{z}) = \frac{1}{Z_{Pr(\lambda,\nu)}} \prod_{i=1}^{N} \{(\boldsymbol{D}_i(\boldsymbol{z} - \boldsymbol{s}))^2 + \lambda\}^{-\frac{1+\nu}{2}} \tag{5.3}$$

where $Z_{Pr(\lambda,\nu)} = \int dz \prod_{i=1}^{N} \{(\boldsymbol{D}_i(\boldsymbol{z} - \boldsymbol{s}))^2 + \lambda\}^{-\frac{1+\nu}{2}}$ is the partition function that makes the whole probability density function integrate to one, and this multivariate integral does not have an analytic form.

5.2.3 Pseudo-likelihood approximation

As in our previous work [90], we employ a pseudo-likelihood approximation to the partition function $Z_{p(\lambda,\nu)}$. Replacing the approximation using the extra information into (5.3), we obtain the following approximate image model:

$$Pr(\boldsymbol{z}|\lambda,\nu) \approx \prod_{i=1}^{N} \frac{\Gamma\left(\frac{1+\nu}{2}\right) \lambda^{\nu/2} \{(\boldsymbol{D}_i(\boldsymbol{z} - \boldsymbol{s}))^2 + \lambda\}^{-\frac{1+\nu}{2}}}{\Gamma(\frac{\nu}{2})\sqrt{\pi}} \tag{5.4}$$

We shall employ this to infer $\boldsymbol{z}$ simultaneously with estimating our hyper-parameters λ, ν and σ.

5.2.4 Joint model

The entire model is the joint model of the observations $\boldsymbol{y}$ and the unknowns $\boldsymbol{z}$.

$$Pr(\boldsymbol{y}, \boldsymbol{z}, f|\boldsymbol{W}, \sigma^2, \lambda, \nu) = Pr(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{W}, \sigma^2)Pr(\boldsymbol{z}|f, \lambda, \nu) \tag{5.5}$$

where the first factor is the observation model and the second factor is the image-prior model with its free parameters defined as λ and ν.

5.3 MAP Estimation

We will employ the joint probability (5.5) as the objective to be maximised. Maximising this w.r.t. $\boldsymbol{z}$ is also equivalent to finding the most probable image $\hat{\boldsymbol{z}}$, i.e. the maximum a posteriori (MAP) estimate, since (5.5) is proportional to the posterior $Pr(\boldsymbol{z}—\boldsymbol{y})$.

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}}\{-\log[Pr(\boldsymbol{y}|\boldsymbol{z})] - \log[Pr(\boldsymbol{z})]\} \tag{5.6}$$

Namely, the most probable high resolution signal is the one for which the negative log of the joint probability model takes its minimum value. Hence, our problem can be solved through minimisation. The expression for the negative log of the joint probability model will then be defined as our minimisation objective and also called as the error-objective. It can be written as:

$$Obj(\boldsymbol{z}, \sigma^2, \lambda, \nu) = -\log[Pr(\boldsymbol{y}|\boldsymbol{z}, \sigma^2)] - \log[Pr(\boldsymbol{z}|f, \lambda, \nu)] \tag{5.7}$$

Equation (5.7) may be decomposed into two terms: the first one that contains all the entries that involve $\boldsymbol{z}$ and the second one contains the terms that do not — i.e. Obj($\boldsymbol{z}$,σ^2, λ, ν) = $Obj_z(\boldsymbol{z})$ + $Obj_{(\lambda,\nu)}(\lambda, \nu)$.

5.3.1 Estimating the most probable $\boldsymbol{z}$

The observation model is also called the likelihood model because it expresses how likely it is that a given $\boldsymbol{z}$ produced the observed $\boldsymbol{y}$ through the transformation $\boldsymbol{W}$. Hence we have for the first term in (5.5):

$$Pr(\boldsymbol{y}|\boldsymbol{z}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{z})^T(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{z})\right\} \tag{5.8}$$

By plugging in the term for the observation model and the prior into (5.7), we obtain the objective function. The terms of the objective (5.7) that depend on $\boldsymbol{z}$ are as follows:

$$Obj_z(\boldsymbol{z}) = \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{z})^2 + \frac{\nu+1}{2}\sum_{i=1}^{N}\log\{(D_i(\boldsymbol{z}-\boldsymbol{s}))^2+\lambda\} \tag{5.9}$$

The most probable estimate is the $\hat{\boldsymbol{z}}$ that has the highest probability in the model. It is equivalently the one that achieves the lowest error. Recap, our model has two factors which depend on the likelihood (or also known as the observation model), and the image-prior that assists the signal recovery. Thus, our error models both the *mismatch* of the predicted model $\boldsymbol{W}\boldsymbol{z}$ with the observed data $\boldsymbol{y}$ and the *determinant* for allowing the free parameters to control the smoothness and the edges encoded in the image-prior. The objective is differentiable; therefore any non-linear optimiser could be practical to optimise the term (5.9) w.r.t. $\boldsymbol{z}$. The gradient of the negative log likelihood term is given by:

$$\nabla(z)Obj_z = \frac{1}{\sigma^2}\boldsymbol{W}'(\boldsymbol{W}\boldsymbol{z}-\boldsymbol{y}) + (\nu+1)\sum_{i=1}^{N} D_i^T\frac{D_i(\boldsymbol{z}-\boldsymbol{s})}{(D_i(\boldsymbol{z}-\boldsymbol{s}))^2+\lambda} \tag{5.10}$$

5.3.2 Estimation of σ^2, λ and ν

Writing out the terms in (5.7) that depend on σ^2, we obtain a closed form for estimating the σ^2.

$$\sigma^2 = \frac{1}{M}\left(\sum_{i=1}^{M}(y_i - \boldsymbol{W}_i\boldsymbol{z})^2\right) \tag{5.11}$$

Terms that depend on λ and ν are given by:

$$Obj_{(\lambda,\nu)} = N\log\Gamma\left(\frac{1+\nu}{2}\right) - N\log\Gamma\left(\frac{\nu}{2}\right) + \frac{N\nu}{2}\log\lambda - \frac{1+\nu}{2}\sum_{i=1}^{N}\log((\boldsymbol{D}_i(\boldsymbol{z}-\boldsymbol{s}))^2 + \lambda) \tag{5.12}$$

Again, both of these hyper-parameters need to be positive values. To ensure our estimates are actually positive, we parameterise the log probability objective (5.12) such that we optimise the +/- square root of these parameters. Taking derivatives w.r.t $\sqrt{\lambda}$ and $\sqrt{\nu}$, we obtain:

$$\frac{d\log p(\boldsymbol{z})}{d\sqrt{\lambda}} = \sum_{i=1}^{N}\frac{\nu(\boldsymbol{D}_i(\boldsymbol{z}-\boldsymbol{s}))^2 - \lambda}{((\boldsymbol{D}_i(\boldsymbol{z}-\boldsymbol{s}))^2 + \lambda)\sqrt{\lambda}} \tag{5.13}$$

$$\frac{d\log p(\boldsymbol{z})}{d\sqrt{\nu}} = \left[N\log\lambda - \sum_{i=1}^{N}\log((\boldsymbol{D}_i(\boldsymbol{z}-\boldsymbol{s}))^2 + \lambda) + N\psi\left(\frac{1+\nu}{2}\right) - N\psi\frac{\nu}{2}\right]\sqrt{\nu} \tag{5.14}$$

where $\psi(.)$ is the digamma function. The zeros of these functions give us the estimates of $\pm\sqrt{\lambda}$ and $\pm\sqrt{\nu}$. Although there is no closed-form solution, these can be obtained numerically using any unconstrained non-linear optimisation method[1], which requires the gradient vector of the objectives.

[1]We made use of the efficient implementation available from http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/

5.3.3 Recovery algorithm

Our algorithm described in Algorithm 4 implements the equations given in the previous section. At each iteration of the algorithm, two smaller gradient descent problems have to be solved; namely one for λ, ν and one for $\boldsymbol{z}$. However, our experiment in 5.4 suggests that it is not necessary to estimate the minimum with high accuracy. We noticed that the inner loops do not require the entire convergence. It is sufficient to increase but not necessarily minimise the objective at each intermediate step.

Algorithm 4 : Recovery algorithm

1: Initialise the estimates $\boldsymbol{z}$
2: iterate until convergence: **do**
3: estimate σ^2 using equation (5.11)
4: iteratively update λ and ν in turn using definiton
5: (5.13) and (5.14), with the current estimate $\boldsymbol{z}$.
6: iterate to update $\boldsymbol{z}$ using equation (5.10)
7: **end**

5.4 Experiments and Discussion

We devise the following two hypotheses to investigate the role of the new prior and we test those using synthetic 1D and 2D signals and real MRI signals:

1. The quality of the recovered signal using the additional information is no worse than the one without the extra information provided, that the extra information is *useful*. This is when the number of zero entries in the new form of the neighbourhood feature, i.e $\boldsymbol{Df}$ is larger than the number of zero entries in $\boldsymbol{Dz}$, that is the generic feature that has not been given the extra similarity information.

2. The fewer the edges in $\boldsymbol{f}$ (that is, the non-zeros in $\boldsymbol{Df}$), the fewer measurements are sufficient for enabling a successful recovery.

We should mention the construction of the measurement matrix $\boldsymbol{W}$ from CS-type $\boldsymbol{W}$ is a random Gaussian matrix ($M \times N$) with $i.i.d$ entries. The SR-type $\boldsymbol{W}$ is a deterministic

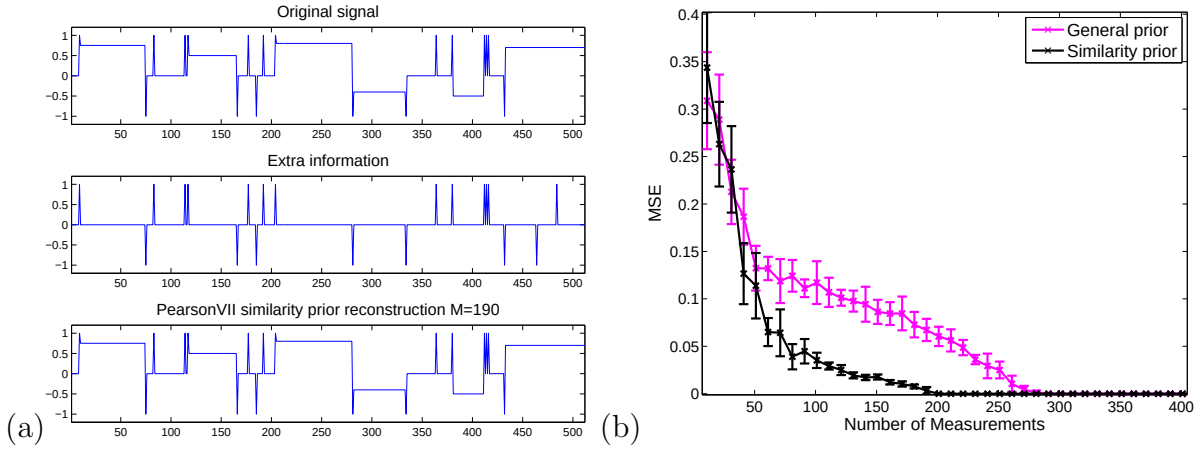


Figure 5.3: (a) The original spike signal; the extra similarity information; and an example of recovered signal from 190 measurements. (b) Comparing the MSE performance of 1D spike signal recovery with and without the extra information. The error bars are over 10 independent trials and the level of noise was σ=8e-5.

transformation that blurs and down-samples the image[1].

5.4.1 Illustrative 1D experiments

In this section, we implement our recovery algorithm on the 1D data, derived from a spike signal[2] of size 512×1 as shown in figure 5.3(a). We proceed by plugging the extra signal into our image-prior and varying the number of measurements using randomly generated measurement matrices $\boldsymbol{W}$ with *i.i.d* Gaussian entries as in CS. The recovery results are summarised in figure 5.3(b). We see our enhanced prior is capable of achieving a good recovery and has a lower mean square error (MSE) than the one without extra information. We also examine the MSE performance as a function of the number of zero entries in the relevant feature vectors (i.e. $\boldsymbol{Df}$ in our case). Figure 5.4 shows MSE results when varying the number of zero entries by constructing variations on the signals. We see when the recovery algorithm received sufficient measurements, for example when M=250 in Figure 5.3, the role of the proposed *similarity prior* gradually reduces. In other words, this *similarity-prior* is *useful* in massively under-determined problems and provided that

[1]Code to generate the SR-type matrices can be found from http://www.robots.ox.ac.uk/∼elle/SRcode/ index.html

[2]Data is taken from http://people.ee.duke.edu/∼lcarin/ BCS.html

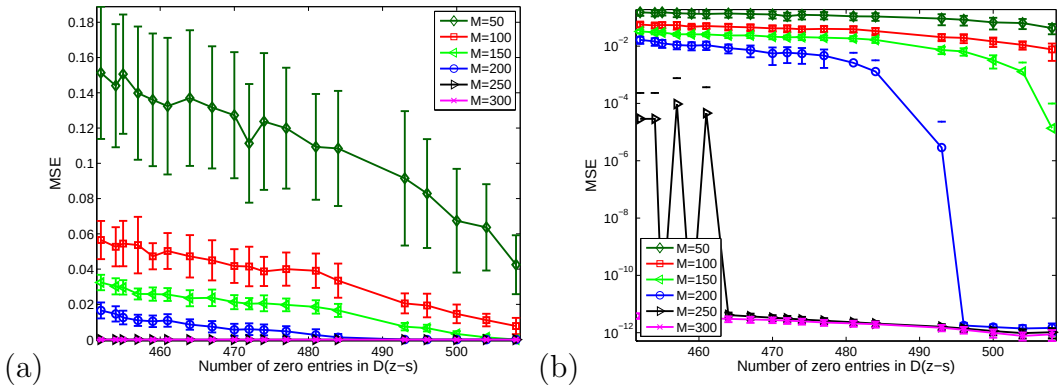


Figure 5.4: (a) Linear scale. (b) Log scale. MSE performance of 1D spike signal using the extra information. The number of zero entries in $\boldsymbol{D}(\boldsymbol{z}\text{-}\boldsymbol{s})$ is varied. The error bars represent one standard error about the mean, from 50 independent trials. The level of noise was σ=8e-5.

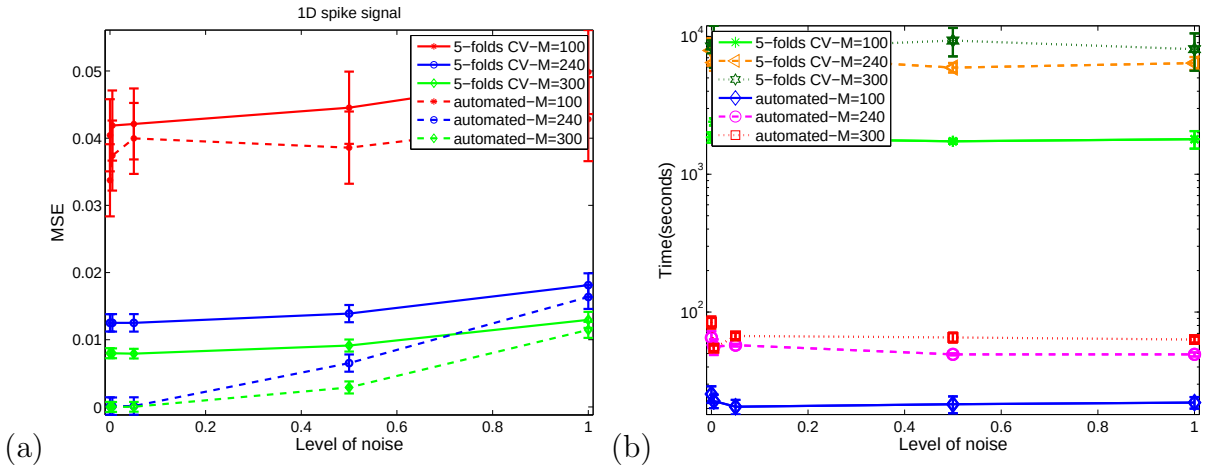
the given extra information has the characteristics described previously in section 5.4.

A widely used alternative way to set hyper-parameters is cross-validation. It is therefore of interest to see how the automated estimation of the hyper-parameters of our Pearson type VII based MRF compares to a cross-validation procedure. We address this by looking at two aspects: MSE performance, and central processing unit (CPU) time. We use the same spike signal for this purpose. For our comparison, we have chosen 5-folds cross validation method for estimating the hyper-parameters λ and ν and the noise variance is assumed to be known for this method. A sensible search range is pursued to avoid a long execution time as we are aware that this method can be extremely time-consuming if the search space is too large.

Figure 5.5 shows the MSE performance and the associated values for the four levels of noise using the CS-type $\boldsymbol{W}$. It is interesting to see that our fully automated parameter estimation turns out to be superior to 5-folds cross validation and it has fast convergence and less execution time.

5.4.2 2D experiments

Following the thorough understanding gained in the previous section regarding the situation *when* the extra information is helpful on the spike signal test cases, we conducted



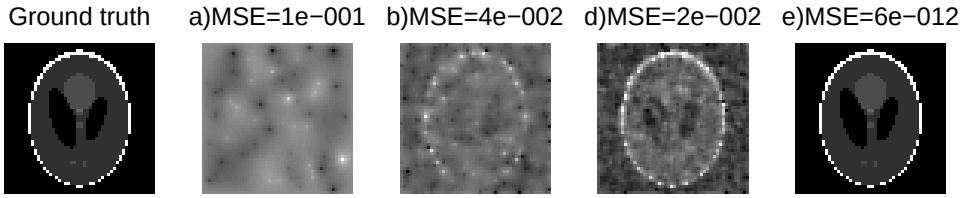
(a) (b)

Figure 5.5: (a) Comparing the MSE performance of the fully automated Pearson type VII based MRF approach with the 5-folds cross validation, tested with four levels of noise (σ= 0.005, 0.05, 0.5, 1). (b) CPU time performance against the same four levels of noise. We see that our automated estimation and recovery is significantly faster than the 5-folds cross validation method. The error bars are over 10 repeated trials for each level of noise. Three sets of measurements (M=100, 240, 300) have been tested for this accuracy comparison.

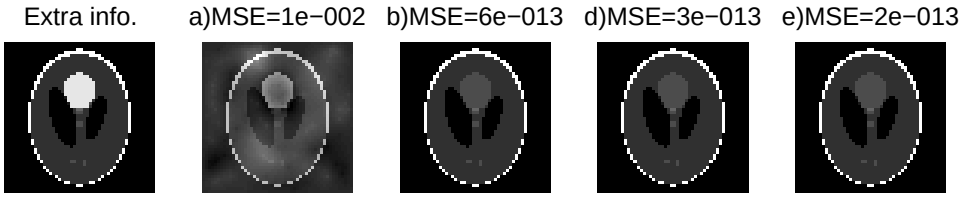
experiments with both compressive sensing (CS) matrices where $\boldsymbol{W}$ contains random entries and also the classical super-resolution matrices where $\boldsymbol{W}$ consists of blur and down-sampling. In this set of experiments, we consider a motionless scene as the extra information. More precisely, the extra information that we employ in our similarity-prior consists of a change in the lighting of some area in the image.

We start by conducting the recovery algorithm on a synthetic data of size [50×50]. The noise variance σ tested in all experiments are set to a smaller range in order to tally the general noise in real data. Figures 5.6 and 5.7 show examples of vastly under-determined problems using the extra information for recovery in comparison with the previous prior devised in [90].

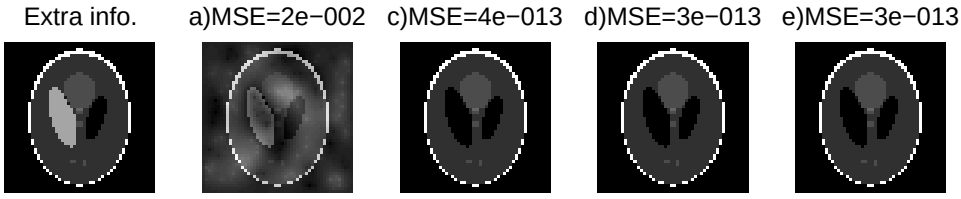
The MSE performance results are given in figure 5.8, and we see the MSE drops rapidly with as the measurement size is increased. Figure 5.9 shows examples of recovered images from this process. We observe that the quality of the recovered image increases rapidly for all 5 levels of noise tested. This contrasts with the recovery results from the general prior, which needs a lot more measurements to perform well. From these findings, the degree



Sample image recovery without extra information.

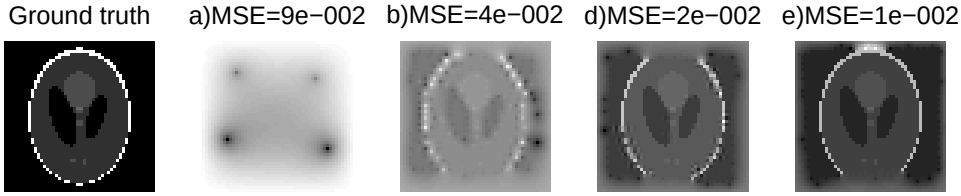


Sample image recovery using extra information of lighting change 1.



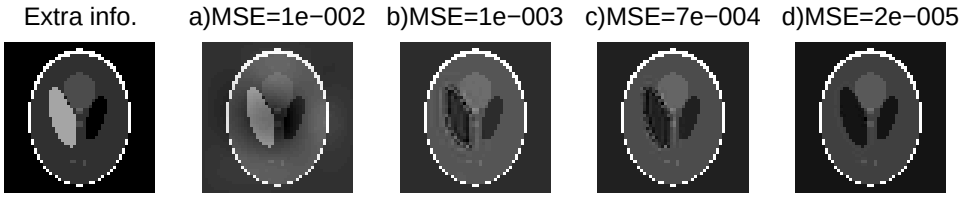
Sample image recovery using extra information of lighting change 2.

Figure 5.6: Examples recovery of 2D synthetic data of size [50×50] in the case of using SR-type $\boldsymbol{W}$, and given two slightly different light changes as extra similarity information. The number of measurements (M) are: a) M=60, b) 460, c) 510, d) 960, e) 1310. The additive noise level was σ=8e-5.



Sample image recovery without extra information.

Sample image recovery using extra information of lighting change 1.



Sample image recovery using extra information of lighting change 2.

Figure 5.7: Examples recovery of 2D synthetic data of size [50×50] in the case of using SR-type $\boldsymbol{W}$, and given two slightly different light changes as extra similarity information. The number of measurements (M) are: a) M=9, b) 441, c) 784, d) 1296, e) 1849. The additive noise level was σ=8e-7.

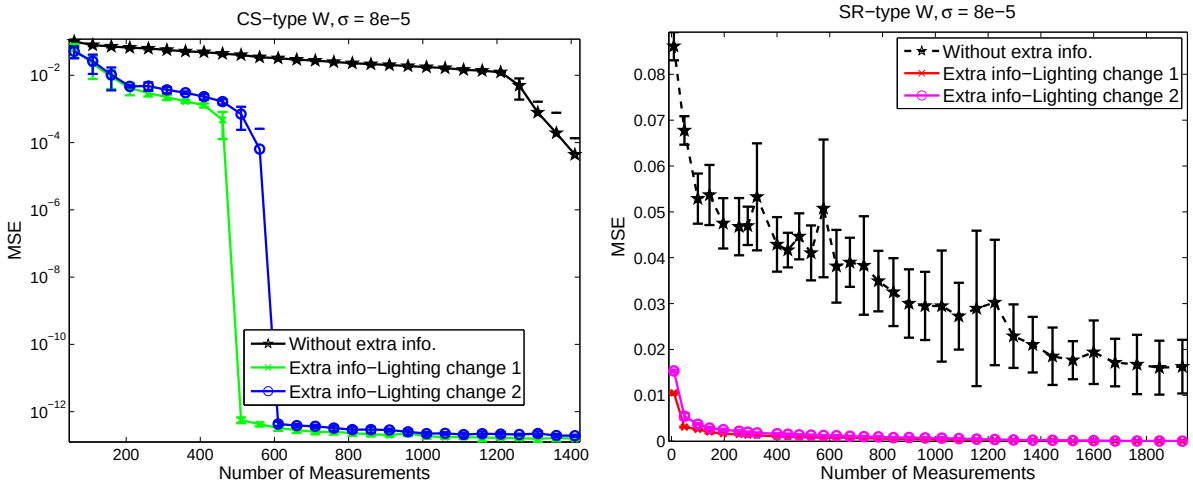


Figure 5.8: MSE performance of synthetic data [50×50] in comparison with the two types of extra information. Here, both types of $\boldsymbol{W}$ were tested and the noise standard deviation was σ=8e-5.

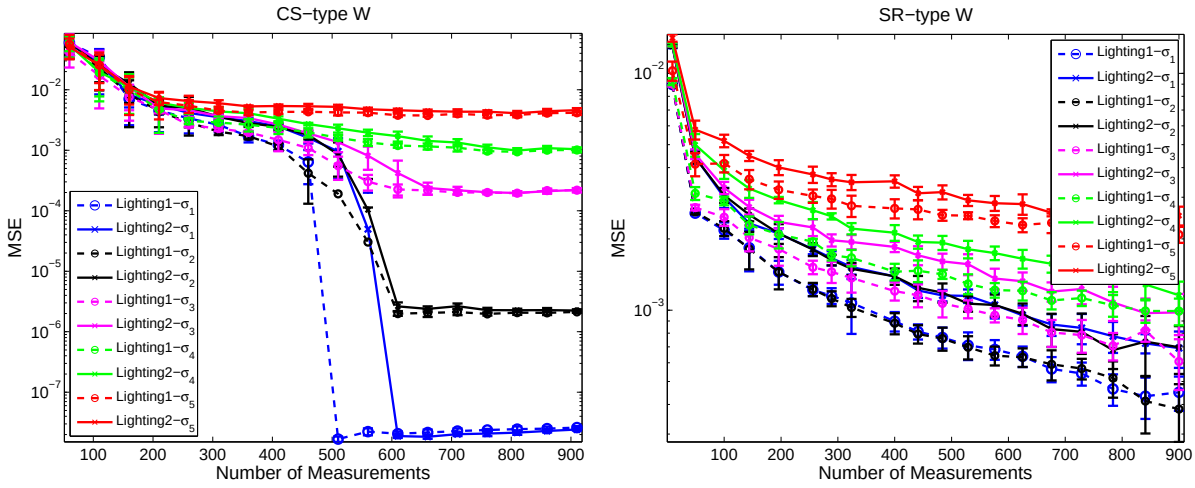


Figure 5.9: Recovery of a [50x50] size image from random measurements (top) and blurred and down-sampled measurement (bottom). The MSE is shown on log scale against varying the number of measurements, in 5 different levels of noise conditions. The noise levels were as follows. Top: $\sigma \in \{\sigma_1$=0.005, σ_2=0.05, σ_3=0.5, σ_4=1, σ_5=2$\}$; Bottom: $\{\sigma_1$=8e-5, σ_2=8e-4, σ_3=8e-3, σ_4=0.016, σ_4=0.032$\}$ — that is the previous noise levels were divided by 0.8 $\sqrt{N}$ to make the signal-to-noise ratios roughly the same for the two measurement matrix types.

of similarity of the available extra information offers a significant impact on the recovery from insufficient measurements. We find that without informative extra information the recovery algorithm does not perform well with such few measurements. The recovered signal and the MSE using the artificial *Phantom* data in figures 5.6 and 5.8 demonstrate that the fewer the edges in the difference image $\boldsymbol{f}$ the better the recovery, or the smaller the number of measurements needed for a good recovery. This result validates our second hypothesis.

In the remainder of the experiments, we will now focus on image recovery using real image data of MRI. We obtained this data from the Matlab image database and we created the additional similarity information from it by changing the lighting of an area on the image. Next we validate our second hypothesis on a variety of MRI images and its lighting changes. The recovery results for both types of $\boldsymbol{W}$ are presented in figures 5.11 and 5.12. The MSE performance for the CS-type $\boldsymbol{W}$ is shown in figure 5.10. Interestingly, we observe that the log scale in that figure is in more direct correspondence with our visual perception rather than using the standard linear scale, and this will be seen by the

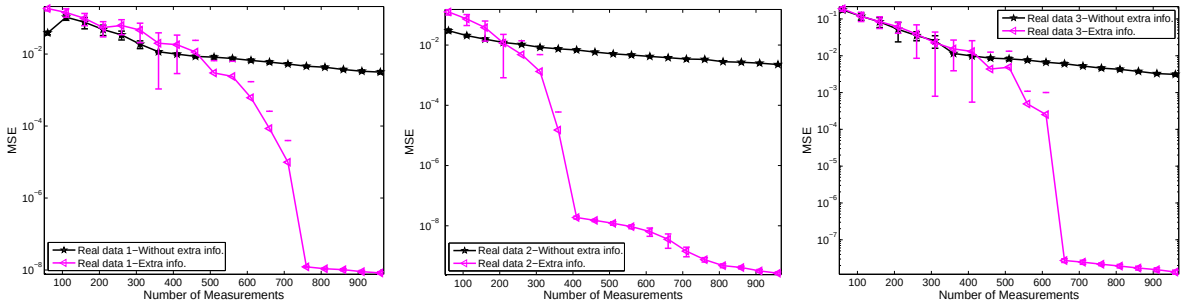
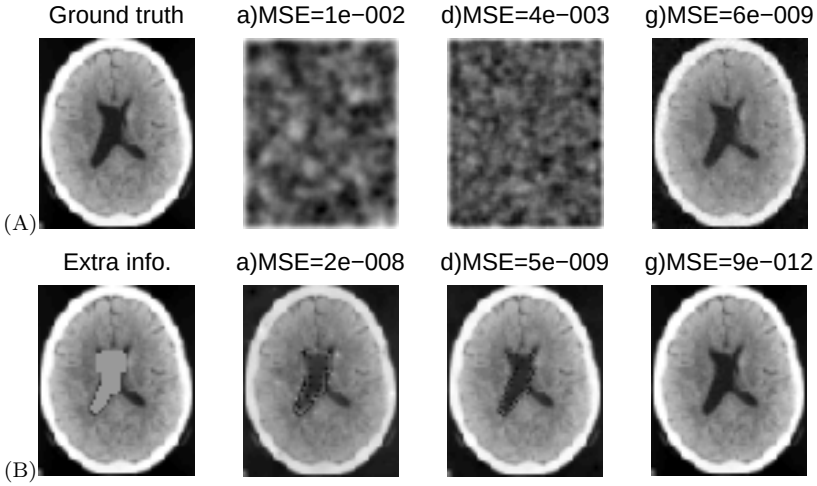


Figure 5.10: From left: MSE performance of real MRI images of size [70×57], [70×57] and [100×80] in comparison with three types of extra information on the three different sets of data. CS-type $\boldsymbol{W}$ was used and the noise standard deviation was σ=8e-5.

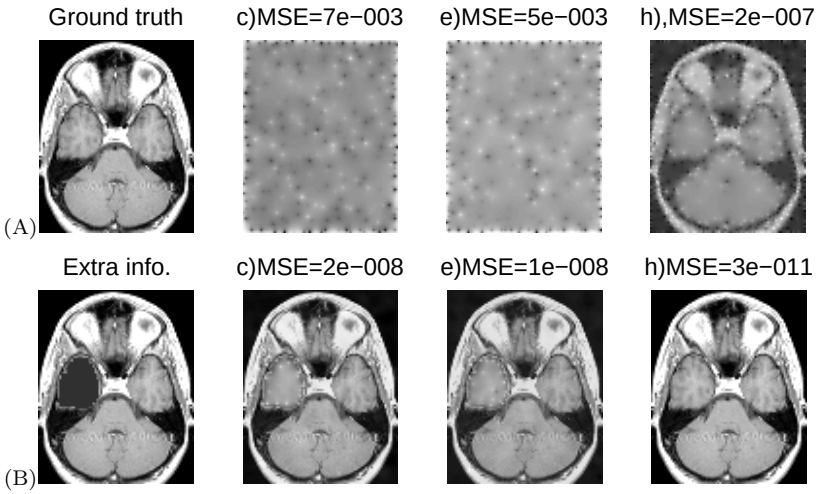
comparison in figures 5.11 and 5.12.

We observed that more than 6000 measurements are required for a good recovery without the extra information in this example. However, from these results we see that our similarity-prior achieves high quality recovery from an order of magnitude less measurements. The recovered images are presented in figures 5.11 and 5.12 for a visual comparison. Finally, we also show a running example of our automated parameter estimation algorithm in figure 5.13 for completeness. As one would expect, the speed of convergence varies with the difficulty of the problem.

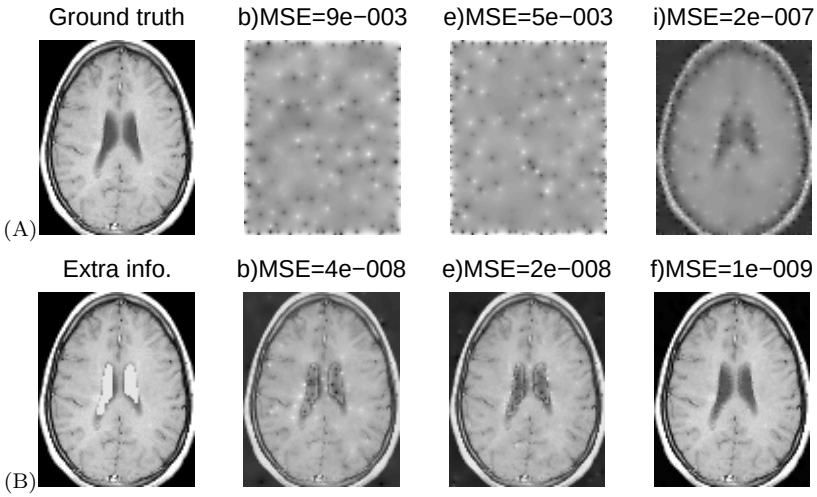
In closing, we should comment on the possibility of using the other types of extra information for signal recovery. Throughout this paper we exploited the similarity created by a lighting change. Depending on the application domain, one might consider a small shift or rotation instead. However, we have seen that the key for the extra information to be useful in our similarity prior is that the difference image must have fewer edges than the original image. This is not the case with shifts or rotations. Therefore to make such extra information useful we would need to include an image registration model into the prior. This is subject to future work.



Sample image recovery of size [70×57] (A)without extra information and (B)using the extra information.

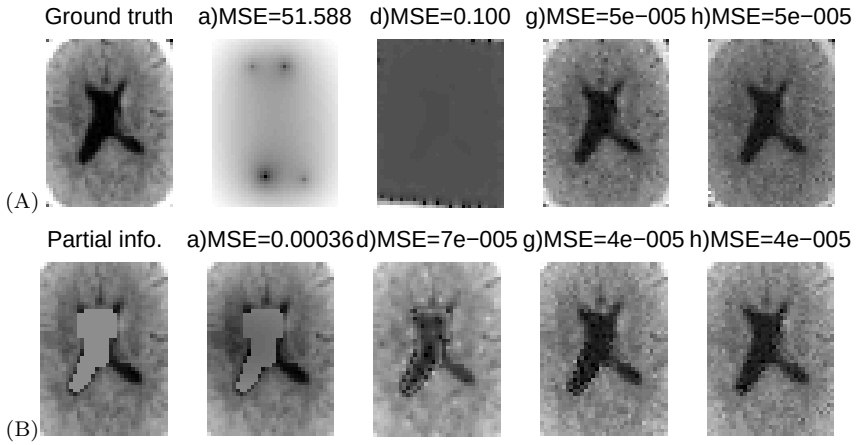


Sample image recovery of size [100×80] (A)without extra information and (B)using the extra information.

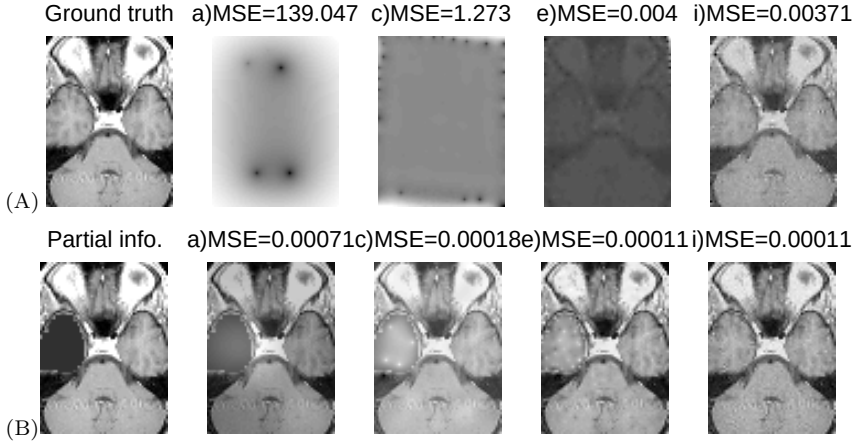


Sample image recovery of size [100×80] (A)without extra information and (B)using the extra information.

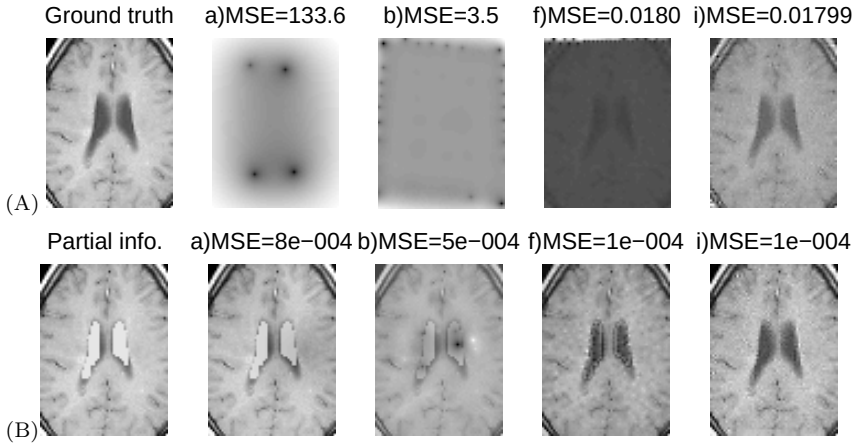
Figure 5.11: Examples of MRI image recovery in the case CS-type $\boldsymbol{W}$, given a motionless consecutive frame with some contrast changes. The number of measurements (M) were: a) M=310, b) 460, c) 560, d) 610, e) 760, f) 1310, g) 3010, h) 5610 i) 7610 and additive noise with σ = 8e-5.



Sample image recovery of size [70×57] (A)without extra information and (B)using the extra information.



Sample image recovery of size [100×80] (A)without extra information and (B)using the extra information.



Sample image recovery of size [100×80] (A)without extra information and (B)using the extra information.

Figure 5.12: Examples of MRI image recovery in the case of SR-type $\boldsymbol{W}$, given a motionless consecutive frame with some contrast changes. The number of measurements (M) were: a)M=6, b) 99, c) 154, d) 396, e) 918, f) 1462, g) 1505, h) 2000, i) 4234. The additive noise is σ=8e-5.

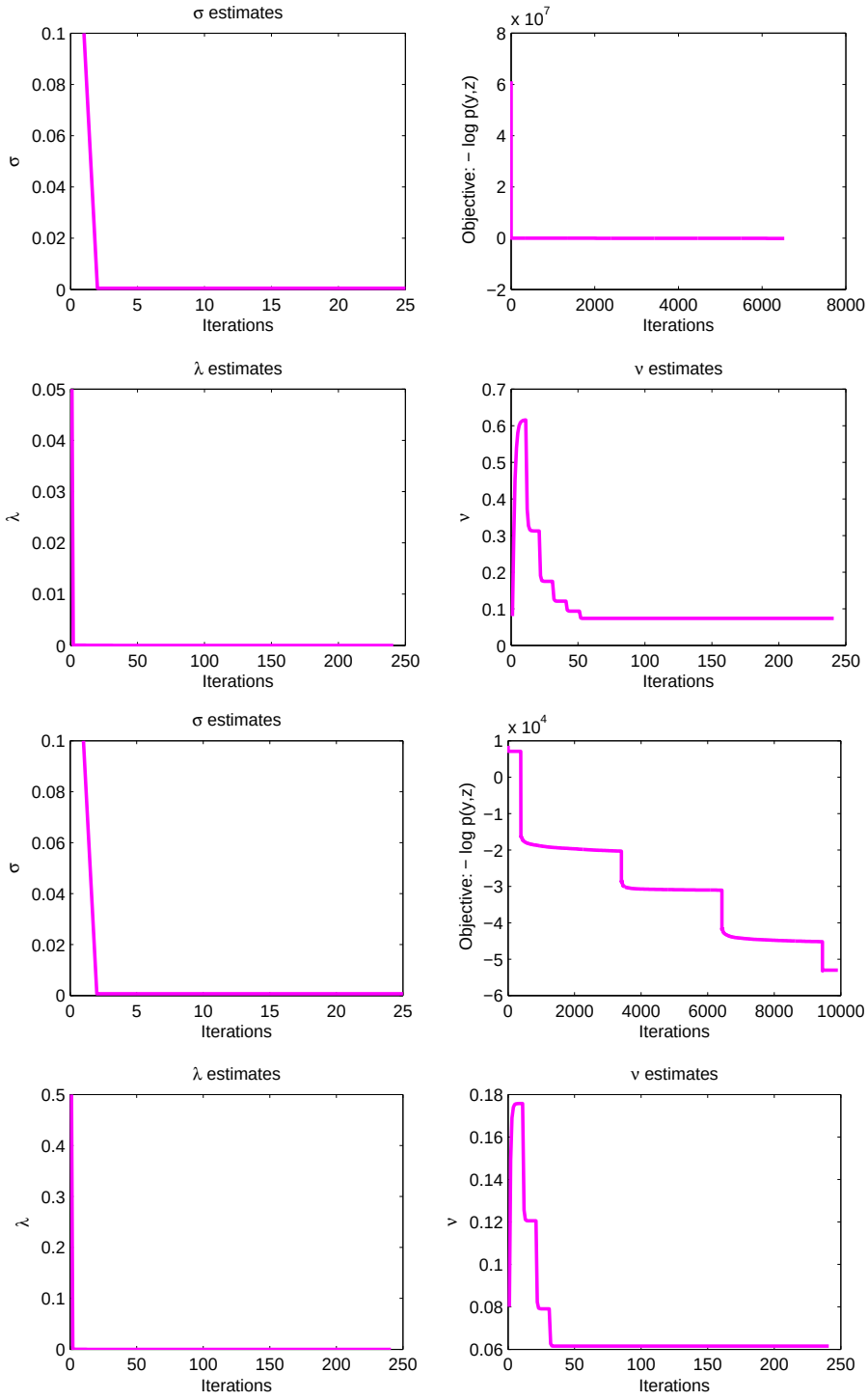


Figure 5.13: Examples evolution of the hyper-parameter updates (σ, λ, ν) and objective function versus the number of iterations of the optimisation algorithm while recovering a 2D signal: from the left, random measurements; and from the right, a blurred and down-sampled low resolution frame. In both experiments, the noise level is σ=8e-5.

5.5 Conclusions

In this chapter, we have formulated and employed a similarity-prior based Pearson type VII Markov Random Field to include the similarity information between the scene of interest and a consecutive scene that has a lighting change. This prior enables us to recover the high resolution scene of interest from fewer measurements than a general-purpose prior would, and this can be applied, e.g. in medical imaging applications. Also in this chapter, we found out using quantitative measurements that our automated parameter estimation is superior to the 5-folds cross validation method, with respect to 1D signals recovery and computational speed. In the next chapter, we consider several tasks to be employed into our novel image-prior algorithm.

CHAPTER 6

MULTI-TASK RECOVERY WITHOUT CONTENT SIMILARITY

In this chapter[1], multi-task recovery is an extended version from the single task recovery. Sharing of hyper-parameters is often useful for multi-task problems as a means of encoding some notion of task similarity. This chapter presents a multi-task approach for signal recovery by sharing higher-level hyper-parameters which do not relate directly to the actual content of the signals of interest but only to their statistical characteristics. Our approach leads to a very simple model and algorithm that can be used to simultaneously recover multiple natural images with unrelated content. We investigate the advantages of this approach in relation to state-of-the-art multi-task compressed sensing and we discuss our findings. Section 6.2 describes the multi-task recovery framework and section 6.3 presents the quantitative measurements and the visual results in comparison with the existing work in multi-task Bayesian Compressive Sensing [3]. Further investigation on the relatedness between the tasks and length of the recovered signals are presented. Finally, the last section concludes the contribution of this chapter.

[1]Part of the work presented in this chapter has been accepted for publications in print in *Proc. 21st of International Conference on Pattern Recognition* (ICPR'2012), Tsukuba, Japan, 11-15 November 2012, IEEE Computer Press, pp. 2246-2249.

6.1 Introduction to Multi-task Recovery

Multi-task signal recovery aims to perform several single-frame recovery tasks simultaneously by exploiting some form of similarity between the tasks. A recent paper tackles this complex problem by an approach termed as Multi-Task Bayesian Compressed Sensing (MT-BCS) [3]. In this approach the similarity of tasks is defined as a percentage of overlapping content — i.e. the positions of edges or smooth regions should have a non-negligible overlap. By its construction, MT-BCS is able to exploit this definition of similarity to recover multiple signals simultaneously in a single run more efficiently than multiple runs of a single-task recovery method [2] would.

Here we propose and investigate a complementary approach in which we seek to exploit a much weaker notion of similarity that is unrelated to the actual content but only depends on the statistical characteristics of the signals to be recovered. We achieve this by building the model of MT-BCS to a further level and sharing higher level hyper-parameters in the resulting model. This turns out to yield a very simple model in terms of its model and experimental design. It has fewer hyper-parameters in which the edge-content related parameters are integrated out and the remaining shared higher-level hyper-parameters can be estimated automatically in a similar manner to what we have tackled previously [97, 90]. The next section describes our multi-task recovery approach and its relation to MT-BCS.

6.2 Multi-task Recovery Framework

Consider K different (though related) recovery tasks. We will denote by $\boldsymbol{z}^{(k)}$ the k-th high resolution signal (scene) of length N that we aim to recover. The observed low resolution (or compressed) signal $\boldsymbol{y}^{(k)}$ has length $M < N$ and is described by the following forward

model:

$$\boldsymbol{y}^{(k)} = \boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)} + \boldsymbol{\eta}^{(k)} \ \forall \ k{=}1,...,K \tag{6.1}$$

where $\boldsymbol{\eta}$ is a mean-zero i.i.d. additive Gaussian noise with variance $\sigma^2 I$. From eq. (6.1), we can write the likelihood as:

$$p(\boldsymbol{y}^{(k)}|\boldsymbol{z}^{(k)}, \boldsymbol{W}^{(k)}, \sigma^2) = \mathcal{N}(\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)}, \sigma^2) \tag{6.2}$$

and in order to infer $\boldsymbol{z}^{(k)}, k = 1, ..., K$, we need to specify a model on these, which we do in the next subsection.

6.2.1 Prior for multiple signals

The gist of multi-task recovery is to exploit similarities between the multiple tasks in order to gain efficiency against performing the tasks individually. There are many ways to define similarity though, and this is a crucial aspect of designing a suitable prior. Before proceeding we define the notation $\boldsymbol{\theta}^{(k)} = \boldsymbol{D}\boldsymbol{z}^{(k)}$ where $\boldsymbol{D}$ could be a wavelet transform as in [3], or another linear transform that makes the representation of $\boldsymbol{z}^{(k)}$ sparse. In particular, we used a simple linear transform from pixel brightness values into neighbourhood-features by taking the difference between pixel brightness and the average of its four cardinal neighbours (see e.g. [90]). With this latter choice of course the components of $\boldsymbol{\theta}^{(k)}$ are not completely statistically independent, however a pseudo-likelihood approximation (as in [90]) makes it possible to treat them as if they were. The transform $\boldsymbol{D}$ is invertible, so estimating $\boldsymbol{\theta}^{(k)}$ is equivalent to estimating $\boldsymbol{z}^{(k)}$, which allows us to simplify the exposition and make the link between the multi-task image-prior of [3] and ours in the sequel.

Hyper-parameter sharing in [3]

Previous work by Ji *et al.* [3] posited the following Gaussian scale-mixture as a multi-task image-prior:

$$p(\boldsymbol{\theta}^{(k)}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{N}(\theta_i^{(k)}|0, \alpha_i^{-1}) \tag{6.3}$$

$$p(\alpha_i|c,d) = \mathcal{G}a(\alpha_i|c,d) \tag{6.4}$$

where $\boldsymbol{\alpha}$ are hyper-parameters shared across the tasks. They then propose to let $c = d \to 0$, which corresponds to a fat-tail uninformative improper prior. The estimates of $\boldsymbol{\alpha}$ are then obtained by the so-called Type II Maximum Likelihood approach:

$$\boldsymbol{\alpha} = \arg\max_{\boldsymbol{\alpha}} \sum_{k=1}^{K} \log \int d\boldsymbol{\theta}^{(k)} p(\boldsymbol{y}^{(k)}|\boldsymbol{\theta}^{(k)}) p(\boldsymbol{\theta}^{(k)}|\boldsymbol{\alpha}) \tag{6.5}$$

Now, since the components of $\boldsymbol{\alpha}$ are inverse variances of the (zero-mean) pixel neighbourhood features, a large entry in this hyper-parameter vector means a nearly zero variance i.e. a locally smooth region, whereas a small entry signifies a large departure from smoothness i.e. a spike or an edge. Sharing of this parameter vector across all the recovery tasks therefore defines a very strong and very specific kind of similarity: the positions of edges and smooth regions must have a considerable overlap. Hence, whenever we know a-priori that the high resolution images that we try to recover are similar to each other in this sense then we can expect that the method in [3] is best placed to exploit it. However, when the notion of similarity defined above is not satisfied, e.g. the images have independent content, then we conjecture that a weaker, higher level similarity of the natural image statistics could be exploited instead. This is what we investigate next.

Higher-level hyper-parameter sharing

We make two important changes to the model in [3]. First, we will not share the inverse-variances of $\boldsymbol{\theta}$ because we want to relax the definition that the extent of overlap in the

positions of edges and smooth regions is what defines similarity. Secondly, we build the model further: Instead of letting hyper-parameters of the Gamma hyper-prior to zero, we will share these among the tasks and estimate them from all the data of the multiple recovery tasks. In addition, we make the model more flexible by introducing a width parameter λ. Summing up, our model is the following:

$$p(\theta^{(k)}|\alpha^{(k)}) = \prod_{i=1}^{N} \mathcal{N}(\theta_i^{(k)}|0, \lambda/\alpha_i^{(k)}) \tag{6.6}$$

$$p(\alpha_i^{(k)}|\nu) = \mathcal{G}a(\alpha_i^{(k)}|\nu/2, 1/2) \tag{6.7}$$

To estimate the remaining high-level hyper-parameters ν and λ we will use a type-II Maximum Likelihood (ML) on the prior term alone[1], and this will yield a simple and computationally convenient algorithm. That is, we take:

$$\{\nu, \lambda\} = \arg\max_{\nu,\lambda} \sum_{k=1}^{K} \log \int d\alpha^{(k)} p(\theta^{(k)}|\alpha^{(k)}, \lambda) p(\alpha^{(k)}|\nu) \tag{6.8}$$

The reason is, the integral in eq.(6.8) is analytically tractable and yields a product of Pearson type VII densities:

$$\int d\alpha^{(k)} p(\theta^{(k)}|\alpha^{(k)}, \lambda) p(\alpha^{(k)}|\nu) = ...$$

$$\prod_{i=1}^{N} \frac{1}{Z(\nu, \lambda)} [(\theta_i^{(k)})^2 + \lambda]^{-\frac{1+\nu}{2}} =: p(\boldsymbol{\theta}^{(k)}|\lambda, \nu) \tag{6.9}$$

where $Z(\nu, \lambda) = \frac{\Gamma(\nu/2)\sqrt{\pi}}{\Gamma\left(\frac{1+\nu}{2}\right)\lambda^{\nu/2}}$

[1]Although a direct extension of the estimation approach in the previous section i.e. an evidence maximisation in the sense of a type-III ML would be interesting to investigate as well, our approach fits with the MAP estimation that we do for finding the most probable images $\boldsymbol{z}^{(k)}$, and we found it to work well in practice as we shall see in the experimental section.

6.2.2 The joint model and parameter estimation

Putting everything together, our joint model for K recovery tasks are defined by:

$$\begin{aligned} & p(\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(K)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K)} | \boldsymbol{W}^{(1)}, ..., \boldsymbol{W}^{(K)}, \sigma^2, \lambda, \nu) \\ & = \prod_{k=1}^{K} p(\boldsymbol{y}^{(k)} | \boldsymbol{\theta}^{(k)}, \boldsymbol{W}^{(k)}, \sigma^2) p(\boldsymbol{\theta}^{(k)} | \lambda, \nu) \end{aligned} \tag{6.10}$$

where we assume that all the tasks are independent to each other.

The negative log of this joint probability will be our objective function that we minimise to get the MAP estimates of all $\boldsymbol{\theta}^{(k)}, k = 1, ..., K$ and ML estimates of λ, ν and σ^2. Note that we have now integrated out the full set of hyper-parameters $\boldsymbol{\alpha}$ (that appeared in [3]) and these do not need to be estimated at all in our approach.

Part of the assignment is to find the minimum value as possible between the observed data, $\boldsymbol{y}$ and the error model, $\boldsymbol{W}\boldsymbol{z}$. Therefore, the likelihood model for K tasks:

$$\prod_{k=1}^{K} p(\boldsymbol{y}^{(k)}, \boldsymbol{\theta}^{(k)} | \boldsymbol{W}^{(k)}, \sigma^2, \lambda, \nu) \propto \prod_{k=1}^{K} \exp \left\{ \frac{1}{2\sigma^2} (\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)} \boldsymbol{\theta}^{(k)})^2 \right\} \tag{6.11}$$

As the signal recovery requires a prior-knowledge, independent Pearson type-VII image-priors for K the tasks will be employed as following:

$$\begin{aligned} p\left(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K)}\right) &= \prod_{k=1}^{K} p\left(\boldsymbol{\theta}^{(k)} | \lambda, \nu\right) \\ p\left(\boldsymbol{\theta}^{(k)}\right) &= Pearson \end{aligned} \tag{6.12}$$

As we already mentioned, this, and our sharing of only ν and λ means a weaker and higher level notion of task similarity than that of [3]) — essentially we only assume similarity of the statistics of $\boldsymbol{\theta}^{(k)}$ and allow the content of the target signals to be different. We carried out the minimisation of the above objective using conjugate gradients in much the same way as described in our previous works [90]. Therefore, the negative log of the joint probability model for multi-task recovery now can be defined as the minimisation

objective:

$$Obj(\boldsymbol{\theta}^{(1)},...,\boldsymbol{\theta}^{(K)},\sigma^2,\lambda,\nu) = \sum_{k=1}^{K}\Big\{-\log[p(\boldsymbol{y}^{(k)}|\boldsymbol{\theta}^{(k)},\sigma^2)] - \log[p(\boldsymbol{\theta}^{(k)}|\lambda,\nu)]\Big\} \tag{6.13}$$

Writing out the second term of equation (6.13) yields the details of the negative log prior as written in (6.14).

$$\begin{aligned}
-\log p(\boldsymbol{\theta}^{(k)}|\lambda,\nu) = \sum_{k=1}^{K}\Bigg\{ & N\log\Gamma\left(\frac{1+\nu}{2}\right) + N\log(\lambda)^{\nu/2} - N\log\Gamma\left(\frac{\nu}{2}\right) \\
& -N\log(\pi)^{1/2} - \sum_{i=1}^{N}\log\left((\boldsymbol{\theta}_i^{(k)})^2+\lambda\right)^{-\frac{1+\nu}{2}}\Bigg\} \\
= \sum_{k=1}^{K}\Bigg\{ & N\log\Gamma\left(\frac{1+\nu}{2}\right) + \frac{N\nu}{2}\log\lambda - N\log\Gamma\left(\frac{\nu}{2}\right) - \frac{N}{2}\log(\pi) \\
& -\frac{1+\nu}{2}\sum_{i=1}^{N}\log\left((\boldsymbol{\theta}_i^{(k)})^2+\lambda\right)\Bigg\}
\end{aligned} \tag{6.14}$$

By taking the log of equation (6.11) for the observation model and the possible prior (6.14), we now obtain the definition form of this objective function. The terms of the objective (6.13) that depend on $\boldsymbol{\theta}^{(K)}$ can be written as follows:

$$Obj_\theta(\boldsymbol{\theta}^{(1)},...,\boldsymbol{\theta}^{(K)}) = \sum_{k=1}^{K}\left\{\frac{1}{2\sigma^2}(\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)}\boldsymbol{\theta}^{(k)})^2 + \frac{\nu+1}{2}\sum_{i=1}^{N}\log[(\theta_i^{(k)})^2+\lambda]\right\} \tag{6.15}$$

The gradient for the MT recovery of the negative log likelihood term is given by:

$$\begin{aligned}
& \nabla \text{w.r.t } \theta\ Obj_\theta(\boldsymbol{\theta}^{(1)},...,\boldsymbol{\theta}^{(K)}) \\
& = \sum_{k=1}^{K}\left\{\frac{1}{\sigma^2}\boldsymbol{W}'^{(k)}(\boldsymbol{W}^{(k)}\boldsymbol{\theta}^{(k)} - \boldsymbol{y}^{(k)}) + (\nu+1)\sum_{i=1}^{N}\theta_i^T\frac{\theta_i^{(k)}}{(\theta_i^{(k)})^2+\lambda}\right\}
\end{aligned} \tag{6.16}$$

Writing out the terms in (6.13) that depend on σ^2, are as following (6.17). Full derivation

can be found in Appendix B.0.1.

$$
\begin{aligned}
Obj(\sigma^2) &= \sum_{k=1}^{K} \left\{ \frac{1}{2\sigma^2} \left(\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)} \boldsymbol{\theta}^{(k)} \right)^2 + M \log \pi + \frac{1}{2} \log[\Sigma] \right\} \\
\frac{\partial Obj(\sigma^2)}{\partial \left(\frac{1}{\sigma^2}\right)} &= \sum_{k=1}^{K} \left\{ \frac{1}{2} \left(\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)} \theta^{(k)} \right)^2 + 0 + \frac{1}{2} \frac{\partial [\log \Sigma]}{\partial \left(\frac{1}{\sigma^2}\right)} \right\} \text{ where } \Sigma = \sigma^2 I \\
&= \sum_{k=1}^{K} \left\{ -\frac{M}{2} \sigma^2 + \frac{1}{2} \left(\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)} \boldsymbol{\theta}^{(k)} \right)^2 \right\} \qquad (6.17)
\end{aligned}
$$

Finally, writing out equation (6.17) and equating to zero to solve, yields closed form estimation for σ^2 in (6.18) and the term that depends on σ^2 for multi-task recovery is written by:

$$
\sigma^2 = \sum_{k=1}^{K} \left\{ \frac{1}{M} \left(\boldsymbol{y}^{(k)} - \boldsymbol{W}^{(k)} \boldsymbol{\theta}^{(k)} \right)^2 \right\} \qquad (6.18)
$$

From equation (6.14), terms that depend on λ are:

$$
Obj_{(\lambda)} = \sum_{k=1}^{K} \left\{ \frac{N\nu}{2} \log \lambda - \frac{1+\nu}{2} \sum_{i=1}^{N} \log[(\theta_i^{(k)})^2 + \lambda] \right\} \qquad (6.19)
$$

Both of these hyper-parameters need to be positive valued. To ensure our estimates are actually positive, we parameterise the log probability objective (6.19) and (6.21) such as to optimise for the $\pm$ square root of these parameters. Taking derivatives w.r.t $\sqrt{\lambda}$ and $\sqrt{\nu}$, we obtain equations (6.20) and (6.22). The details derivation of λ is presented in Appendix B.0.2. We derive the derivative of $\sqrt{\lambda}$ as in (6.20):

$$
\frac{\partial Obj(\lambda)}{\partial \sqrt{\lambda}} = \sum_{k=1}^{K} \left\{ \sum_{i=1}^{N} \frac{\nu((\theta_i^{(k)})^2 - \lambda}{((\theta_i^{(k)})^2 + \lambda)\sqrt{\lambda}} \right\} \qquad (6.20)
$$

From equation (6.14), terms that depend on ν are defined by:

$$Obj_{(\nu)} = \sum_{k=1}^{K} \left\{ N \log \Gamma\left(\frac{1+\nu}{2}\right) - N \log \Gamma\left(\frac{\nu}{2}\right) + \frac{N\nu}{2} \log \lambda \right. \\ \left. - \frac{1+\nu}{2} \sum_{i=1}^{N} \log[(\theta_i^{(k)})^2 + \lambda] \right\} \tag{6.21}$$

Finally, we get the derivative of $\sqrt{\nu}$ as in (6.22). Each term derivation is derived in Appendix B.0.3.

$$\frac{\partial Obj(\nu)}{\partial \sqrt{\nu}} = \\ \sum_{k=1}^{K} \left\{ \left(N\psi\left(\frac{1+\nu}{2}\right) - N\psi\frac{\nu}{2} + N \log \lambda - \sum_{i=1}^{N} \log\left((\theta_i^{(k)})^2 + \lambda\right) \right) \sqrt{\nu} \right\} \tag{6.22}$$

6.3 Experiments

We investigate three research questions as follows: (i) To what extent our definition of relatedness can be exploited for multi-task recovery? (ii) How does the existing work in MT-BCS [3] perform on data that only has our weaker notion of relatedness? (iii) What do we lose by exploiting only our weaker notion of similarity when the data really has the stronger one exactly as defined in MT-BCS [3]?

6.3.1 Results and discussion

(i) To gain insight into our first question, we conduct experiments to compare the performance of multiple runs of a single-task recovery algorithm with the performance of one run of our multi-task recovery method. In both methods we use the Pearson type VII image model, however the single-task approach estimates the hyper-parameters ν, λ and the noise variance σ^2 separately for each task whereas the multi-task approach uses all the data to estimate these.

From our experiments we found that multiple runs of single task recovery already performs very well in terms of means square error (MSE). Nevertheless, the multi-task approach works in a single run and from our experiments it performs no worse for a class of signals (e.g. natural images have similar statistics even when they have different content), and it may even yield a slight improvement in the quality of recovery since it has more data to estimate these hyper-parameters. Figure 6.1, shows the MSE results

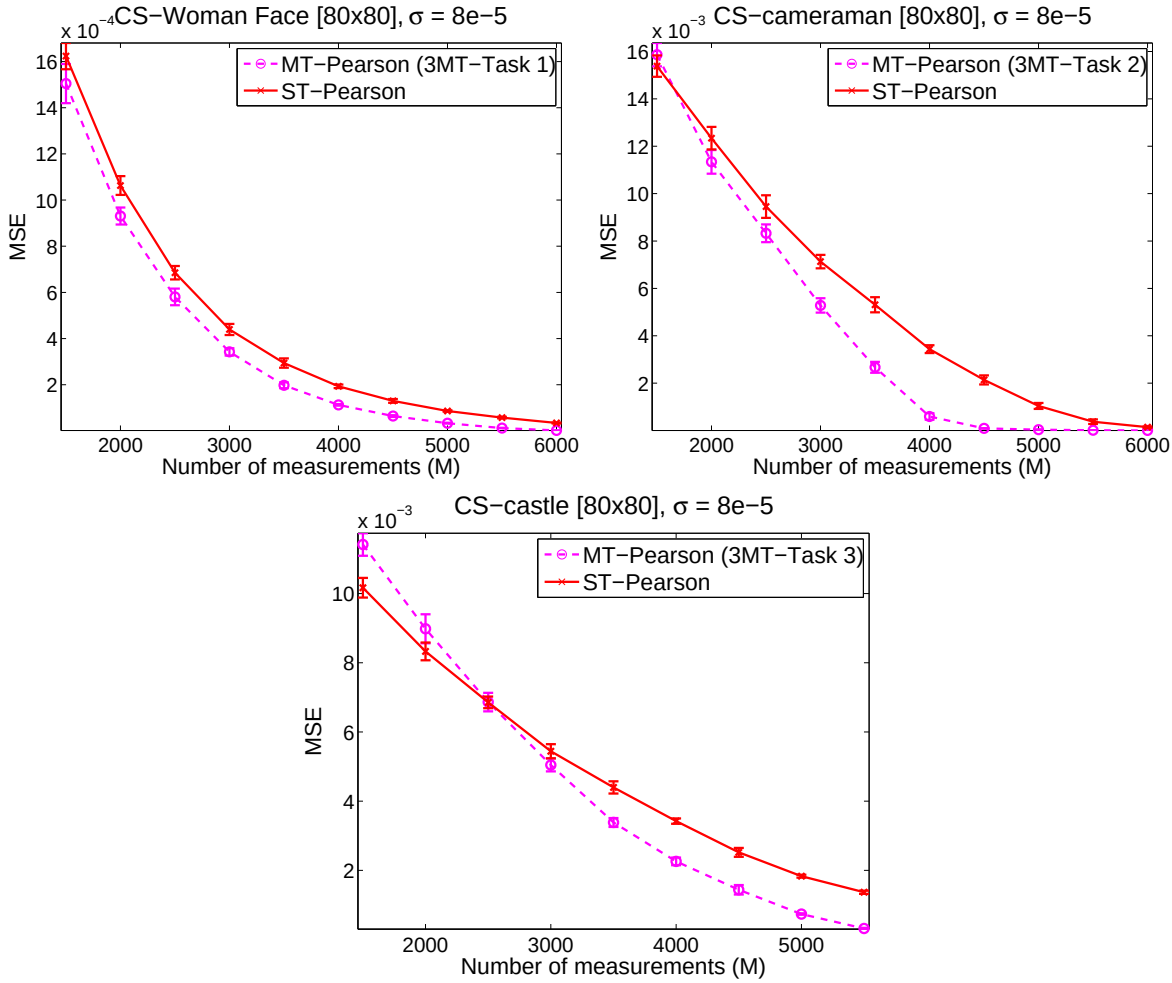


Figure 6.1: Comparing three separate runs of single-task (ST)-Pearson based recovery against one run of multi-task (MT)-Pearson based recovery. The task is to recover three different high resolution images from only one randomly compressed and noisy frame of each. The noise standard deviation was $\sigma = 8 \times 10^{-5}$.

of three single-task recoveries versus one multi-task recovery of the same target images — natural images of size $[80 \times 80]$ pixels each, which have no overlapping content other than their naturally similar image statistics: '`woman face`', '`cameraman`', and '`castle`'.

We varied the number of measurements (extent of compression), and we worked with $\boldsymbol{W}^{(k)}$ randomly generated matrices with i.i.d. standard Gaussian entries. We see the multi-task approach is able to get good recovery in a single run and it needs slightly less measurements for good recovery in this example.

(ii) Next, we compare our multi-task approach presented in the earlier section against M-BCS [3] on data that has no overlapping content but exhibits only our weaker notion of similarity. To perform a systematic study, we first use synthetic 1D spikes signals modified from [3]. We try to recover two signals simultaneously, each having length 512, of which 20 entries are spikes (+1 or -1) and the rest of entries are zero. However, contrary to [3] the positions of these spikes are generated randomly for both signals, with no planned overlap in their positions. Figure 6.2 shows an example of the data, as well as the results of an extensive comparison when the number of measurements available is varied. Clearly, our MT-Pearson approach that only shares high level hyper-parameters performs significantly better in this problem setting. It achieves lower MSE and needs less measurement to recover the high resolution signals. MT-BCS looses out because it expects a content-wise overlap, which is not present in the true signals in this setup.

To further validate this conclusion, figure 6.3 shows multi-task comparison results on image recovery experiments where the task is to recover pairs of natural images simultaneously. Again, we see that our MT-Pearson approach outperforms MT-BCS, and this is because these images have similar statistics but no overlap in their content.

(iii) Finally, we test our approach in scenarios that do have content overlap of the kind that is hard-wired into MT-BCS. We use exactly the same 1D spike signals and use exactly the same experimental setup as [3], and also employ their experimental protocol: That is, the task is to recover two spike signals simultaneously when they have 25%, 50% or 75% of their spikes in the same positions, and the noise level is set to 0.005. By the design of MT-BCS, the larger the percentage of overlap the better MT-BCS will perform, whereas our MT-Pearson does not depend on any content-wise overlap but only on higher level statistical similarity. The upper plot of Figure 6.4 shows the results of MT-BCS

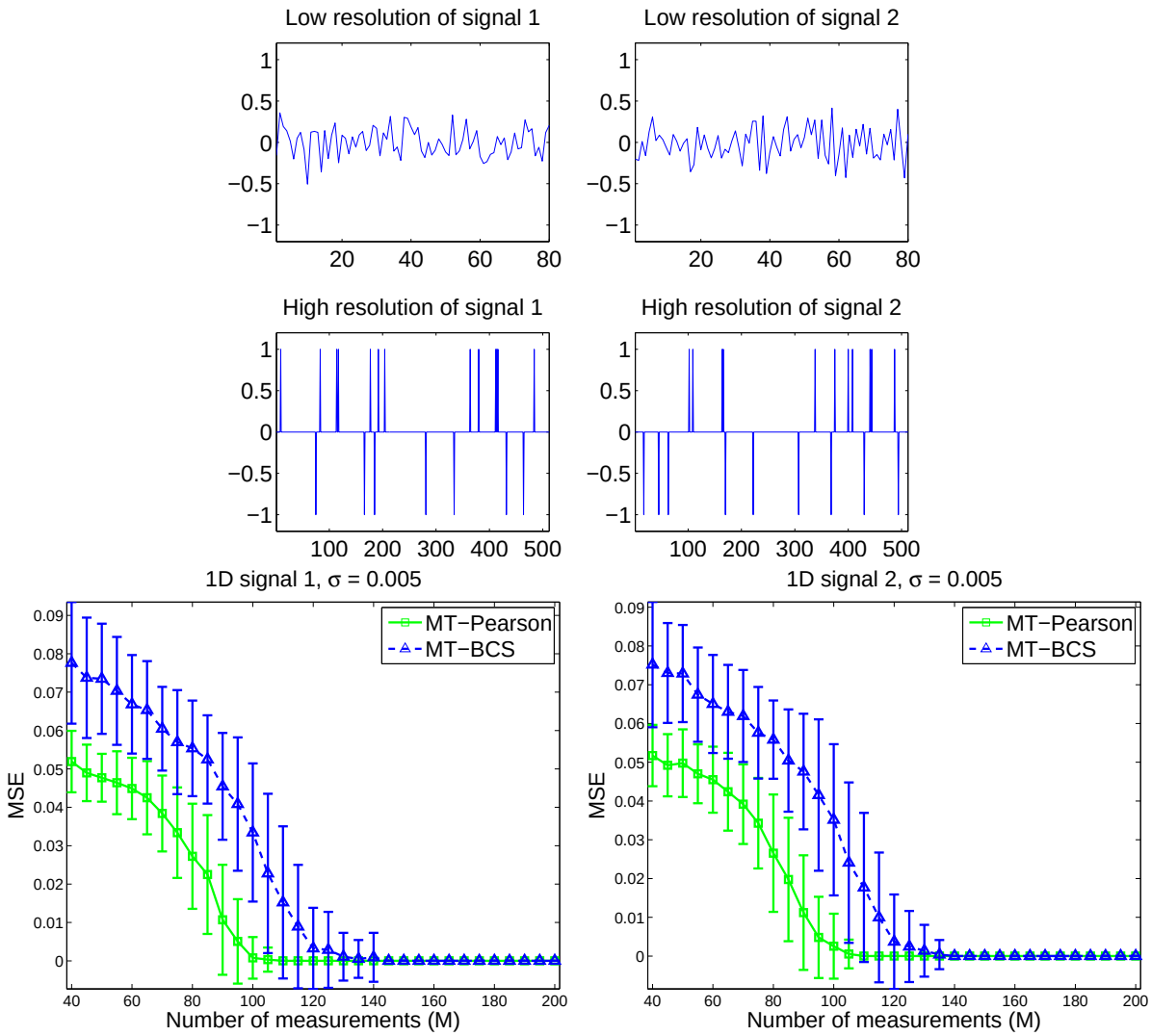


Figure 6.2: First 4 plots: Examples of input measurements and high resolution of 1D signals to be recovered. Last plot: Comparison of our MT-Pearson approach against MT-BCS [3] on recovering two spike signals simultaneously.

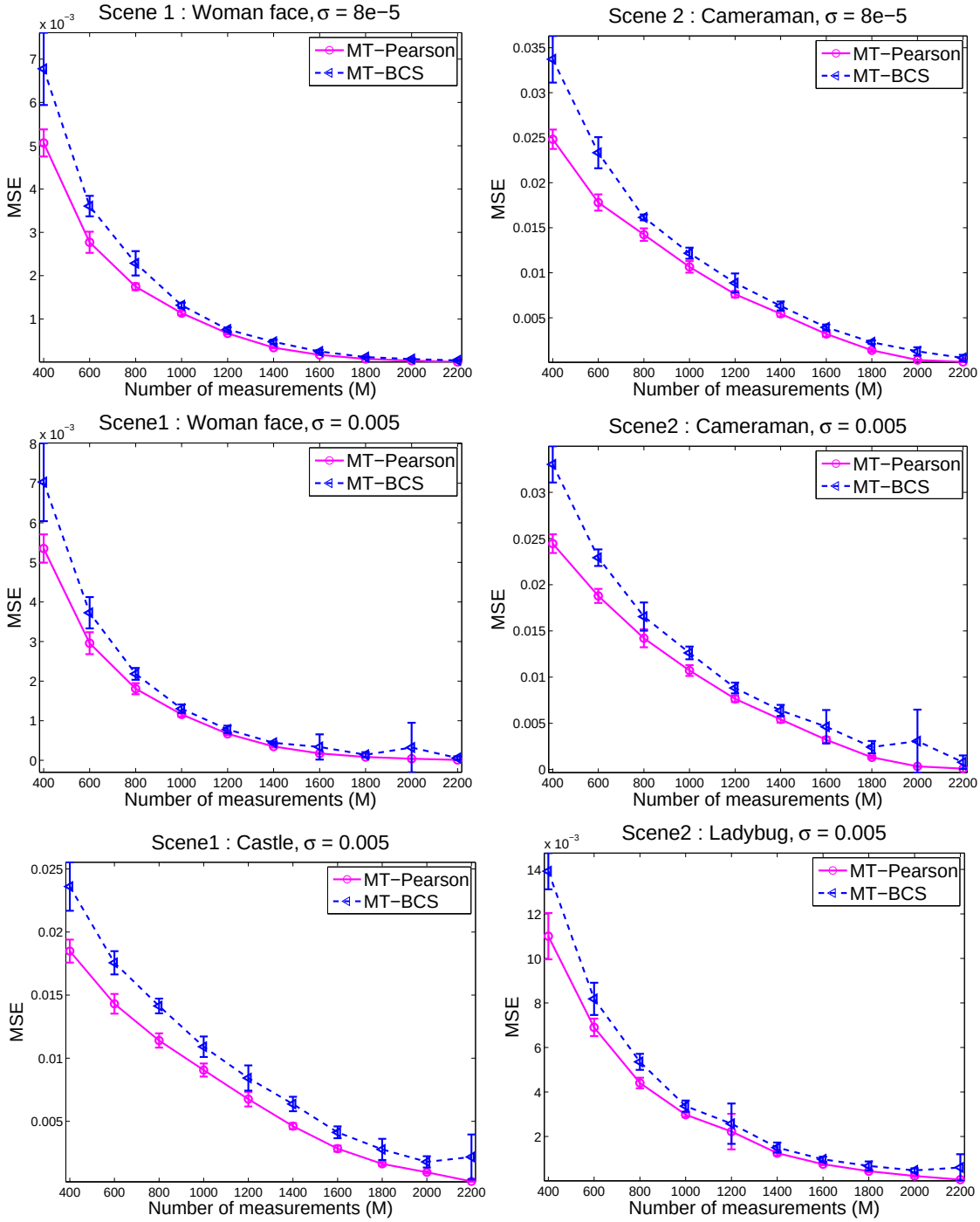


Figure 6.3: Plot represent three sets of experiments simultaneously recovering pairs of natural scenes of size $[50 \times 50]$.

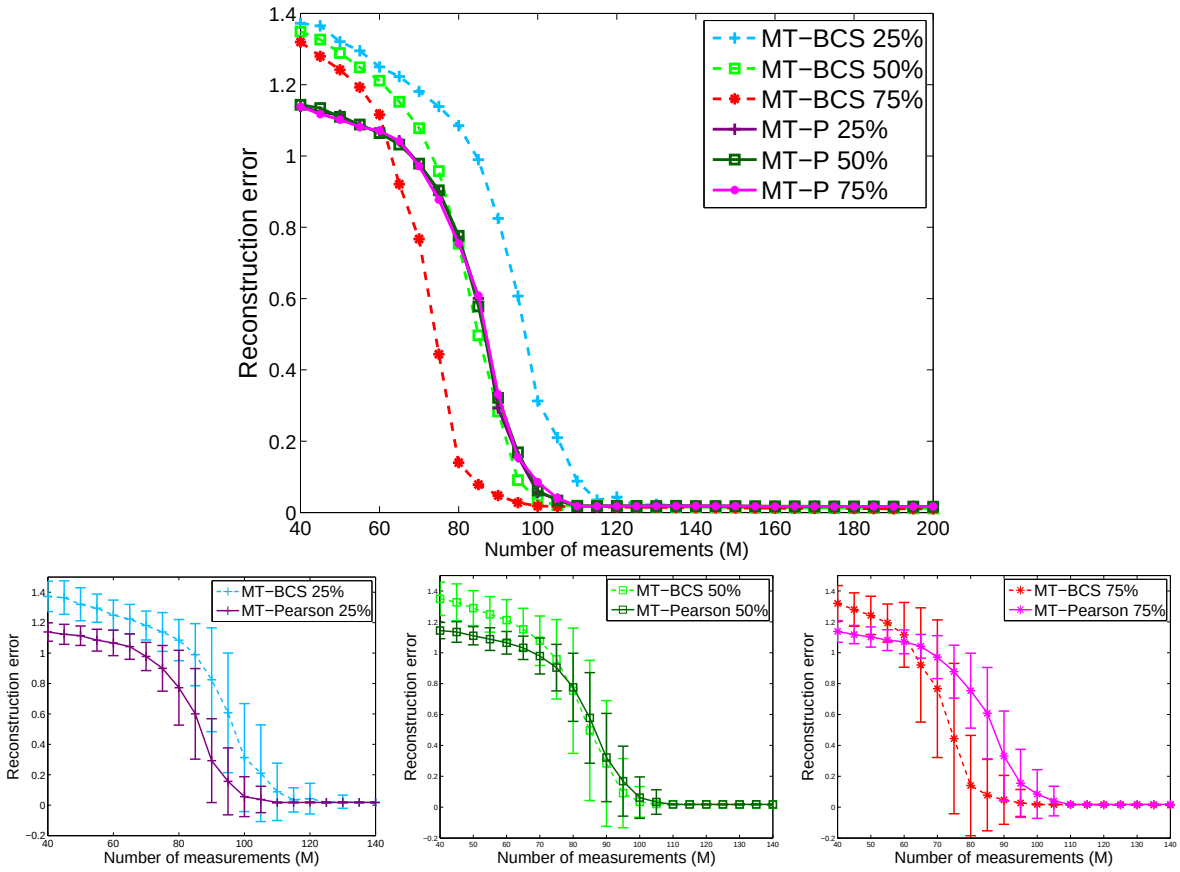


Figure 6.4: *Upper plot*: Reconstruction errors of MT-Pearson and MT-BCS [3], as a function of the number of compressive measurements. *Lower plots*: The variance of reconstruction errors for 25%, 50% and 75% similarity over 100 independent runs.

superimposed with our MT-Pearson. Interestingly, we see that our MT-Pearson is only outperformed by MT-BCS in 75% spike-overlap conditions. It comes out statistically equal to MT-BCS in the 50% overlap setting and it is significantly superior to MT-BCS in settings that have less content-wise overlap. The lower plots of Figure 6.4 detail all pairwise comparisons separately with error bars shown for completeness.

6.3.2 Further investigations on relatedness

We investigate to what extent our definitions of relatedness can be exploited for a multi-task recovery. Previous work of MT-BCS in [3], relatedness is defined as having a fraction of the non-zeros (e.g; edges on 2D) data, spikes on 1D data) in exactly at the same positions. MT-BCS built on this assumption and hence it is able to exploit this kind of relatedness. In order to investigate this relatedness, a hypothesis is devised. Signals that do not satisfy the above definition of relatedness (e.g., spike signals that have their non-zeros in different positions, images that differ in their content) might still be related on the higher level. For instance, the relatedness can also be defined by having similar distributions on the neighbourhood feature with similar levels of sparsity.

The main characteristic of any natural image is a local-smoothness [90, 97], which means that intensities of neighbouring pixels tend to be very similar. Hence, the neighbourhood feature will be sparse. The distributions of the neighbourhood feature in the first image will look similar with the distribution of the second image in term of the histogram shape although the entire content would be different. The same thing goes to 1D signal case. Two different signals are given to be recovered where some of the spikes in 1D are generated randomly and some of the spikes remain at the same positions to retain the relatedness of the previous spike. We assumed that having a similar shape of the distribution may lead to a good recovery on sharing the hyper-parameters than estimating it individually. Therefore, sharing the hyper-parameters (α, ν in [3]) of the image-prior and the noise level sigma is preferable.

Validating the hypothesis can be obtained by comparing the recovery between single

task Pearson algorithm and the multi-task Pearson on 1D spike signals for noise level, σ=0.005. The original signals are illustrated in figure 6.5. In the first investigation, two signals recovery are obtained with the aid of estimating the hyper-parameters individually. For the second finding, two recoveries are obtained simultaneously in a single run by sharing the hyper-parameters. The quantitative results are over 100 independent trials for each measurement presented in figure 6.5 and a visual comparison between single task and multi-task recovery is displayed in figure 6.6. Based on the quantitative results and

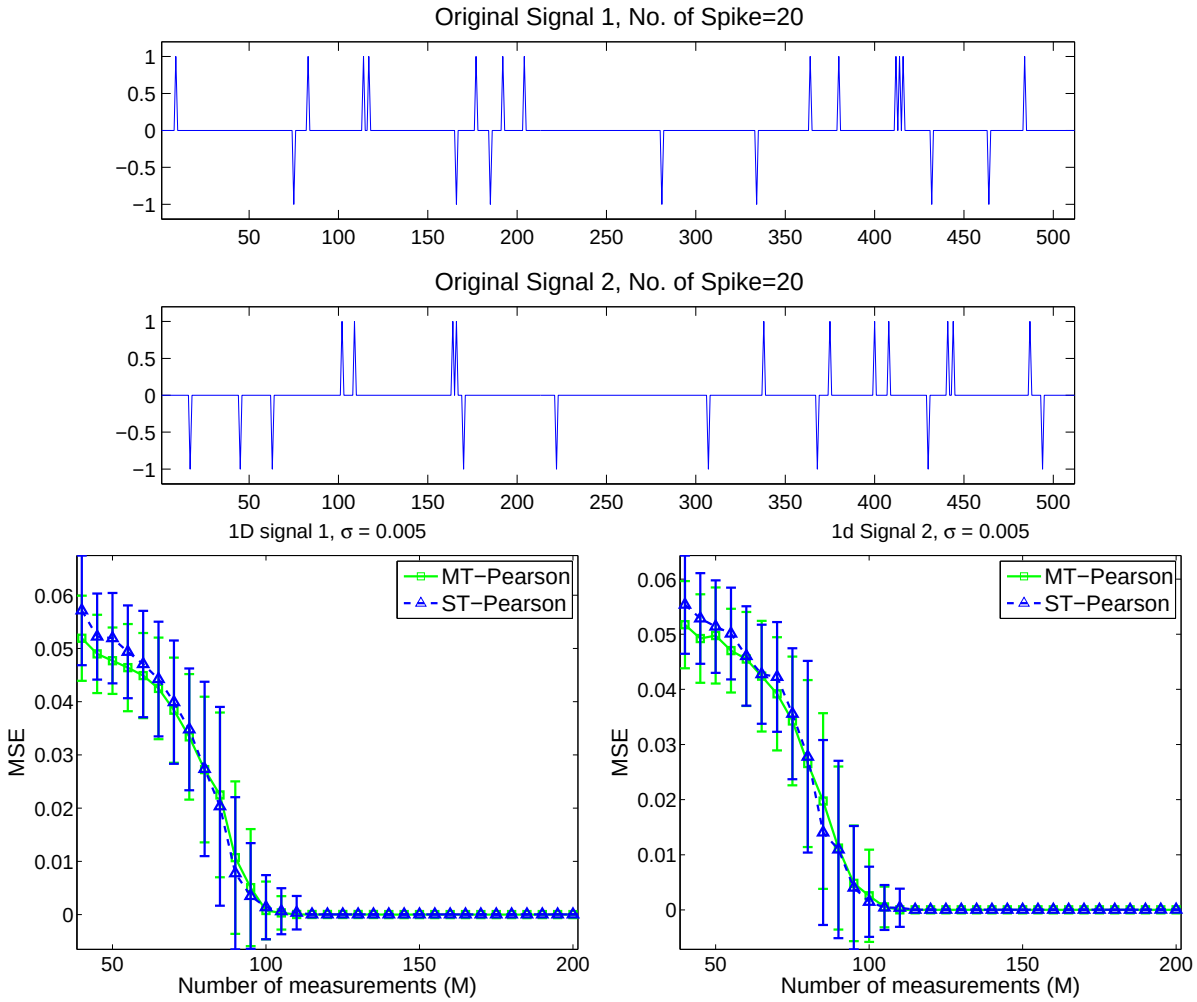
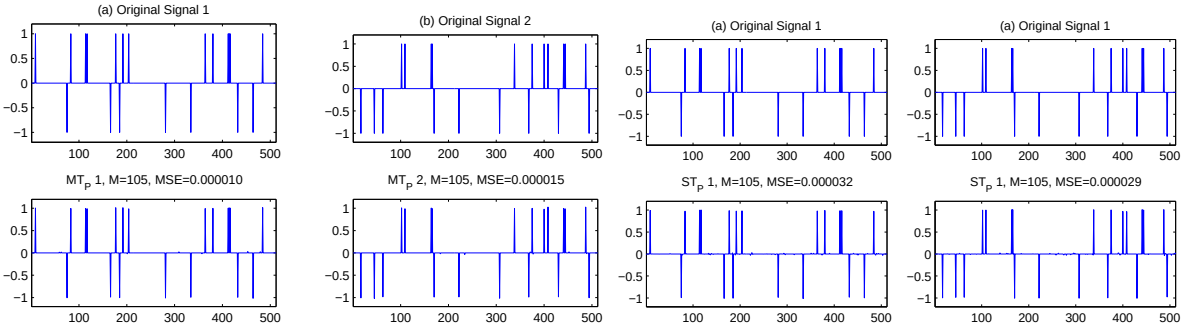
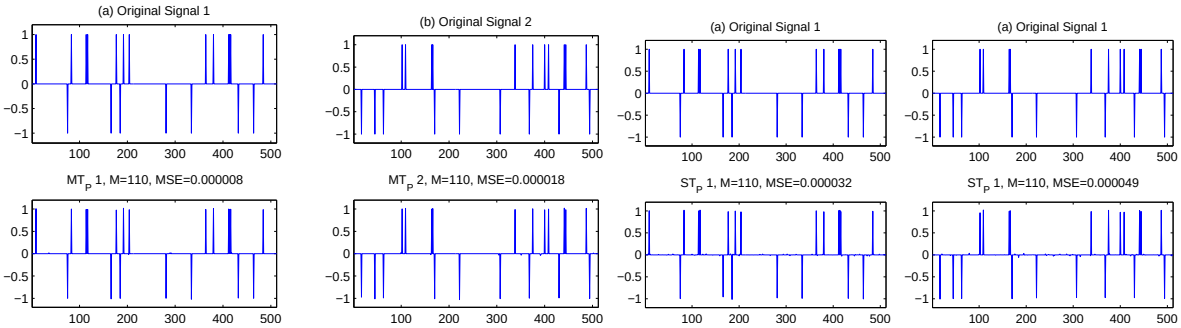


Figure 6.5: Original signal of the 1D spikes of length $N = 512$. The two original signals have random spikes. Comparing the ST-Pearson with the MT-Pearson algorithm using signal 1 and 2 when the number of spike (T) is set to 20.

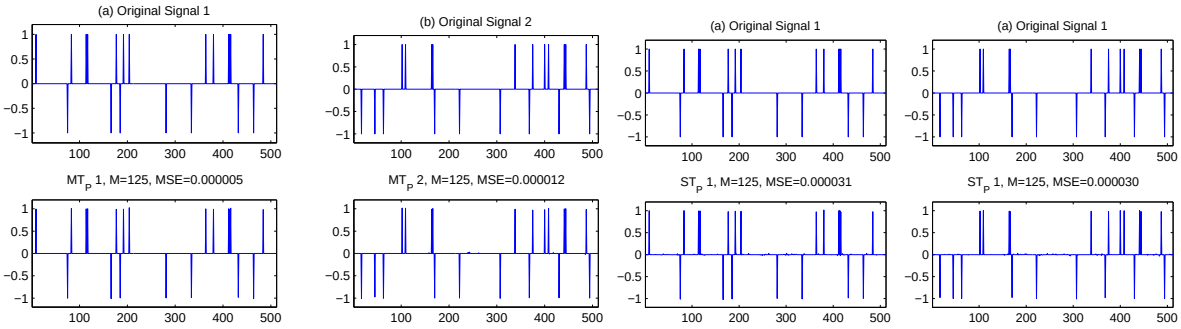
the visual comparison, the differences on the mean square error are too small. Therefore, a statistical test is performed to compare over the two related samples and to assess whether



MT-Pearson recovery is the first two from left and ST-Pearson recovery is the first two from right. The particular measurement is M=105.



MT-Pearson recovery is the first two from left and ST-Pearson recovery is the first two from right. The particular measurement is M=110.



The first two from left is the MT-Pearson recovery and the first two from right is using ST-Pearson recovery. The particular measurement is M=110.

Figure 6.6: Visual comparison for Figure 6.5.

their population mean ranks differ. Rank sum tests are used and presented in Table 6.1 and 6.2 from M=105 until M=200 for both methods, ST-Pearson and MT-Pearson for 1D signal 1. We perform rank sum test [98] that returns the result of the hypothesis test, performed at the 5% significance level, in $\boldsymbol{H}$ as shown in Table 6.1 and 6.2. The null hypothesis for the test is that the median are equal for both methods. As we see from the table, the test rejects the null hypothesis. Therefore, the two methods are not statistically equal. We conclude that MT-Pearson method is slightly better than the ST-Pearson in highly insufficient measurements for both recover signals.

Table 6.1: Rank sum test over 100 independent trials shows the probability ($\boldsymbol{P}$) of observing the given result for each method. $\boldsymbol{H}$=0 indicates that the null hypothesis cannot be rejected at the 5% level and when $\boldsymbol{H}$=1 indicates that the null hypothesis can be rejected at the 5% level.

Number of measurements M	**Method 1 ST-Pearson (MSE mean)**	**Method 2 MT-Pearson (MSE mean)**	$\boldsymbol{P}$	$\boldsymbol{H}$
105	6.3773e-004	3.2724e-004	5.3229e-033	1
110	3.5300e-004	1.3274e-005	2.9772e-034	1
115	3.6792e-005	1.3090e-005	2.7208e-034	1
120	3.5273e-005	1.3218e-005	3.2577e-034	1
125	3.5499e-005	1.2924e-005	4.9545e-034	1
130	3.5326e-005	1.3463e-005	3.8992e-034	1
135	3.5119e-005	1.3638e-005	2.8036e-034	1
140	3.5128e-005	1.2561e-005	3.8997e-034	1
145	3.5178e-005	1.2326e-005	2.8037e-034	1
150	3.5225e-005	1.2247e-005	2.5621e-034	1
155	3.6230e-005	1.1940e-005	2.7208e-034	1
160	3.7221e-005	1.1955e-005	2.7208e-034	1
165	3.7688e-005	1.2290e-005	2.8037e-034	1
170	3.8084e-005	1.1741e-005	2.8037e-034	1
175	3.8250e-005	1.1563e-005	2.5621e-034	1
180	3.8705e-005	1.1248e-005	2.5621e-034	1
185	4.1924e-005	1.1911e-005	2.5608e-034	1
190	4.2315e-005	1.0976e-005	2.5621e-034	1
195	4.4065e-005	1.1360e-005	2.5621e-034	1
200	4.5187e-005	1.0675e-005	2.5621e-034	1

Rank sum test from M=105 until M=200 for both methods, ST-Pearson and MT-Pearson for 1D signal is presented in table 6.2. The noise level conducted in this experiment is 0.005.

We then proceed to test on 2D signal where the natural images were chosen randomly

Table 6.2: Rank sum test over 100 independent trials shows the probability ($\boldsymbol{P}$) of observing the given result for each method. $\boldsymbol{H}$=0 indicates that the null hypothesis cannot be rejected at the 5% level and when $\boldsymbol{H}$=1 indicates that the null hypothesis can be rejected at the 5% level.

Number of measurements M	**Method 1 ST-Pearson (MSE mean)**	**Method 2 MT-Pearson (MSE mean)**	$\boldsymbol{P}$	$\boldsymbol{H}$
105	4.4751e-004	5.3848e-004	8.7440e-032	1
110	3.8571e-004	1.3141e-005	7.0919e-034	1
115	3.6544e-005	1.3182e-005	2.7208e-034	1
120	3.5165e-005	1.3584e-005	1.4927e-033	1
125	3.5143e-005	1.3843e-005	6.8834e-034	1
130	3.5143e-005	1.3843e-005	6.8834e-034	1
135	3.5394e-005	1.3624e-005	6.6809e-034	1
140	3.5705e-005	1.3348e-005	4.6672e-034	1
145	3.3885e-005	1.3073e-005	3.2577e-034	1
150	3.5664e-005	1.2665e-005	5.3229e-033	1
155	3.6080e-005	1.1643e-005	1.0091e-033	1
160	3.6576e-005	1.2182e-005	2.9772e-034	1
165	3.7659e-005	1.2238e-005	4.6672e-034	1
170	3.9024e-005	1.2012e-005	8.4817e-034	1
175	3.8426e-005	1.1771e-005	3.3569e-034	1
180	4.0533e-005	1.1951e-005	2.7208e-034	1
185	4.0420e-005	1.1416e-005	2.5621e-034	1
190	4.0894e-005	1.0851e-005	2.5621e-034	1
195	4.4254e-005	1.1563e-005	2.5621e-034	1
200	4.5510e-005	1.1768e-005	2.5617e-034	1

for noise level, σ=0.005. Experiments are presented in figures 6.7, 6.8 and 6.9 on comparing the ST-Pearson with the MT-Pearson algorithm. From the results, we can see that most of the time when using the ST-Pearson achieved a good recovery than using the MT-Pearson as illustrated in figures 6.7-6.9 in highly under-determined setting. However, we also observed that by using the MT-Pearson, a better recovery can be obtained than the ST-Pearson in certain test cases. It would be caused by accidental recovery. Majority of the performances in figures 1D and 2D signals are matched where ST-Pearson result is always better than the multi-task recovery method.

As a conclusion, both 1D and 2D signals required individual estimating parameters using a single-task recovery algorithm when no information of the spike or edges are shared. We also suggest that the multi-task recovery on sharing the hyper-parameters is more useful for 1D spike signals compared to the 2D signals. It is because in 2D signals, the second task recovery is entirely different and makes the problem more difficult to be solved by sharing the hyper-parameters. Our hypothesis states that having a similar distribution can lead to a better recovery when sharing the hyper-parameters on 1D signal and it is invalid for 2D signals. This is because in 1D signal, the dissimilarity is represented by a few spikes and made the MT-Pearson recovery achieve a slightly better performance in terms of quantitative measurement.

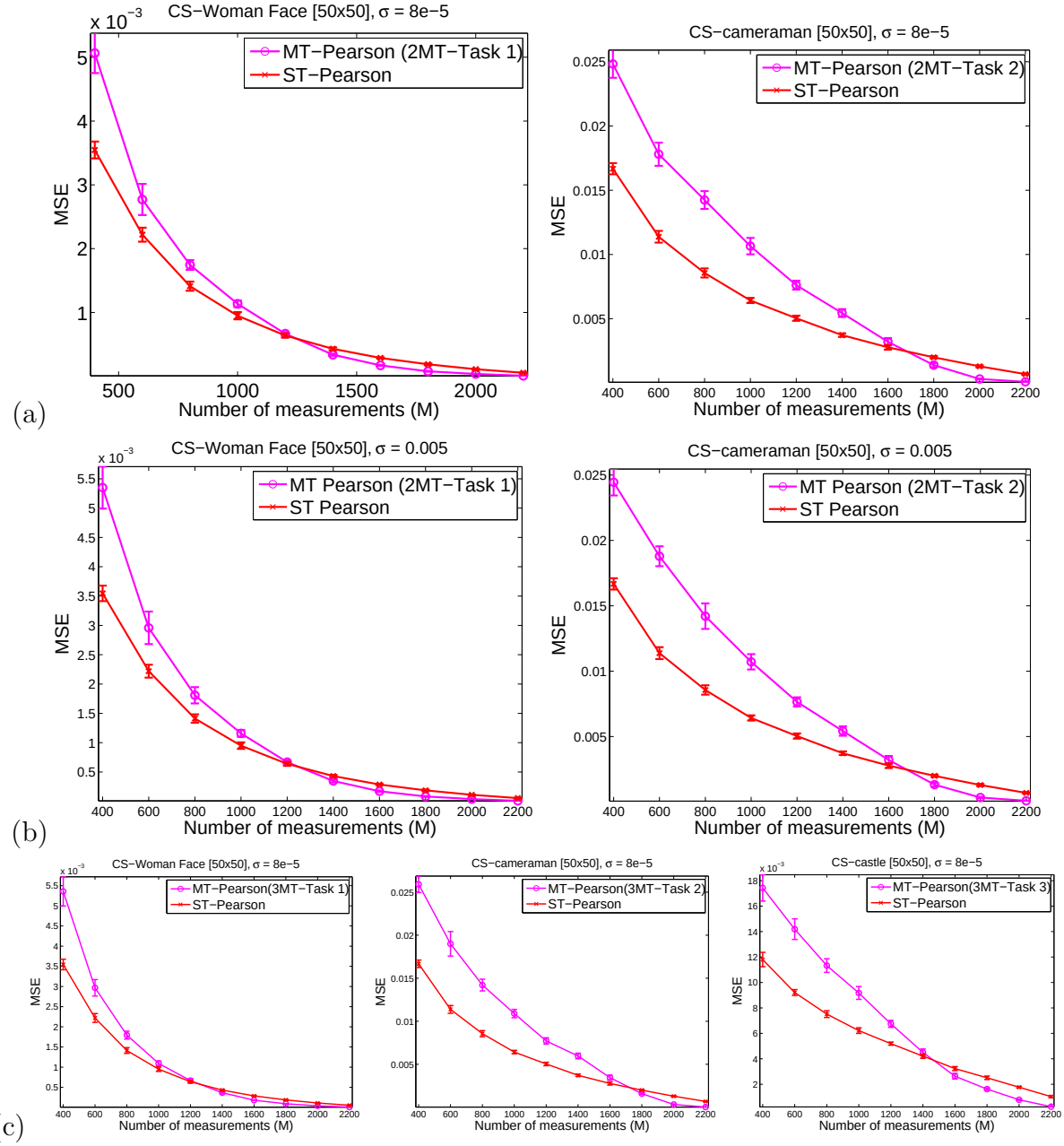


Figure 6.7: Experiments (a) and (b) recover two scenes simultaneously. (c) Recover three scenes simultaneously. Error bars are over 10 independent trials.

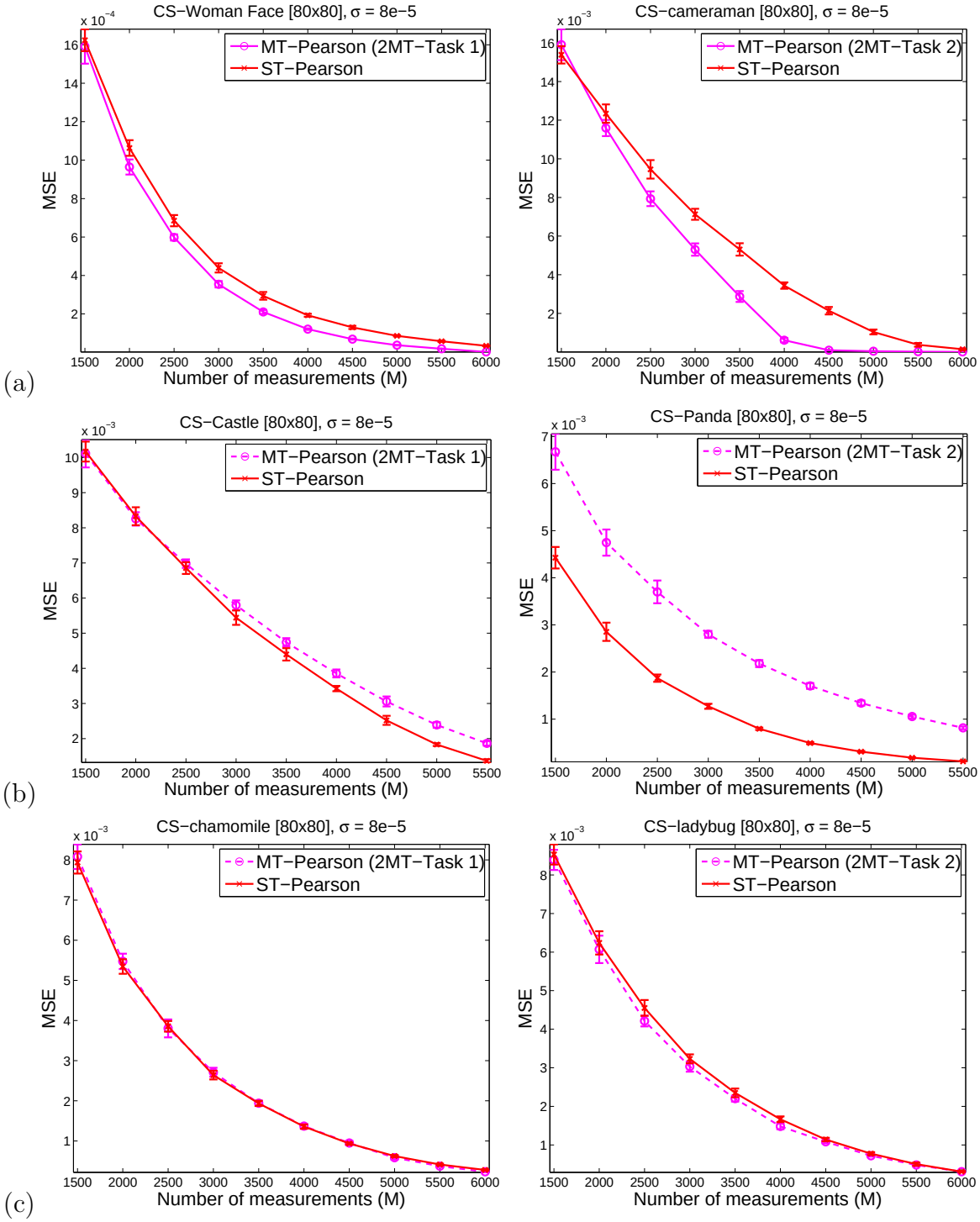


Figure 6.8: Experiment on recovering the two scenes of size [80×80] simultaneously on a three different test cases of 2D images. The level of noise is 8e-5. From three different test cases on recovering the two scenes, test case 1(a) shows the capability of the MT-Pearson is always better than the ST-Pearson. In test case 2 (b), the ST-Pearson performs better than MT-Pearson especially on recovering the second scene. However, test case 3(c) shows almost similar performance for both methods and seems that MT-Pearson achieves a slightly better performance on recovering the second scene. We conclude that test case 1(a) is a rare case and we classified it as an outlier.

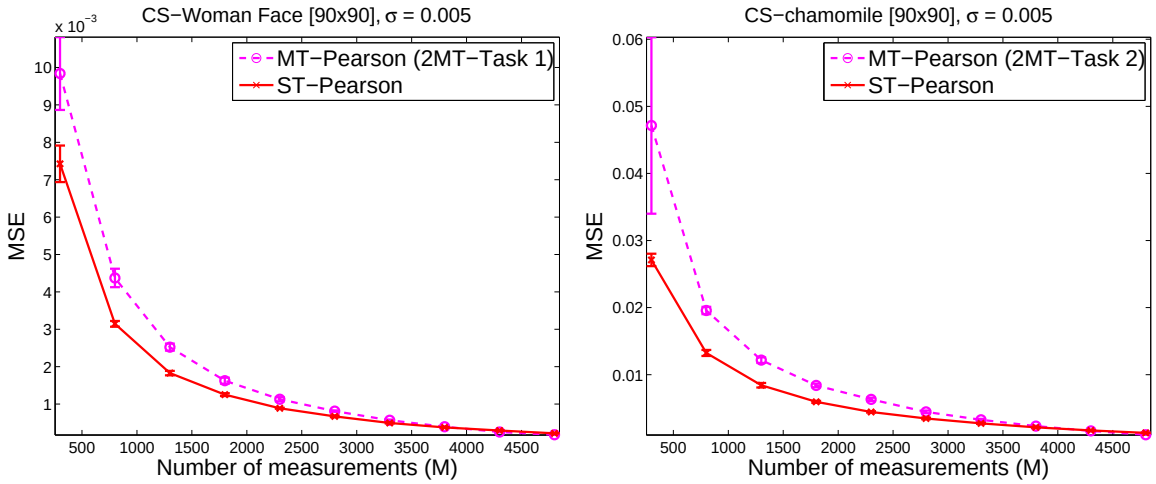


Figure 6.9: Experiment on recovering two scenes simultaneously. The noise level, σ is 0.005.

6.3.3 Further investigations on the length (N)

Next, we investigated the effects on the dimensionality N that was used to recover the high dimensional of 1D signals. We devised experiments where the proportion number of the spike for each test case measurement tested was given equally. Five types of measurements (N=100, 200, 300, 400 and 512) are conducted and results are presented in figure 6.10. From the outcome, the 1D signal does not show any significant result when comparing the single task with the multi-task recovery for all tested measurements. We notice a significant result in 2D signals case as illustrated in figure 6.1 where it shows that the multi-task recovery algorithm performs better than the single task recovery for three different tasks. However, this is not the case as illustrated in figure 6.9 when two tasks are recovered simultaneously. These could be different images that may contain extra texture decrease the performance of the MT-Pearson when sharing the hyper-parameters. For 1D signals performance as shown in figure 6.10, we observe both methods (ST and MT-Pearson) behave equally. We can conclude that sharing hyper-parameters are more useful for the 2D signals compared to the 1D signals with a condition that those different images exhibit a similar pattern on its histogram of the neighbourhood feature.

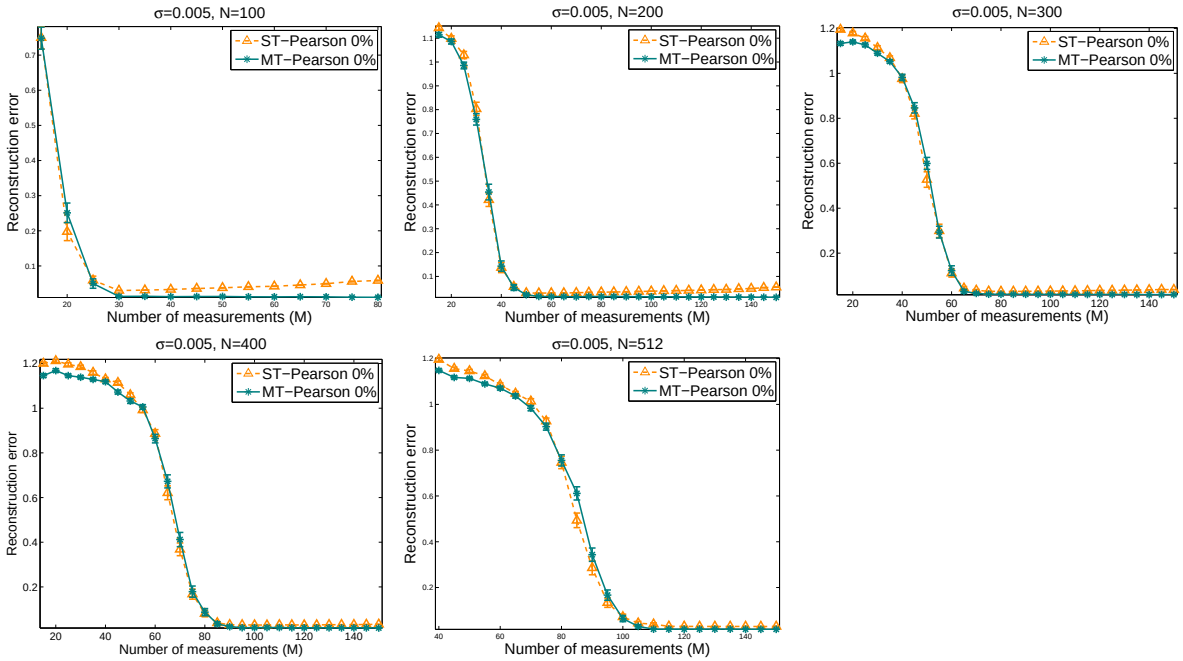


Figure 6.10: The results are averaged over one standard error from 100 independent trials.

6.4 Conclusions

We presented a new approach to multi-task signal recovery where the target signals need not have any overlap in their content but only share their higher level statistical characteristics. This can be used for simultaneous recovery of sets of natural images in a single run. We compared our MT-Pearson approach with multi-task BCS, which is the state-of-the-art for multi-task signal recovery and we highlighted the settings in which our approach is advantageous.

In this chapter, we also investigated the relatedness by devising a hypothesis where signals that do not satisfy the relatedness definition might still be related on the higher level (e.g., spike signals that have their non-zeros in different positions or images that differ in their content). We conclude that signals with non-similarity recovered effectively with individual estimation of the hyper-parameters. Nevertheless in some cases where three different images were recovered simultaneously by sharing the hyper-parameters in the image-prior, a significant finding shows that multi-task Pearson type VII is superior compared to the single task Pearson recovery.

CHAPTER 7

SUMMARY AND CONCLUSIONS

This chapter presents a summary of the thesis and outlines some potential future directions.

7.1 Concluding Remarks

This thesis provides the following achievements:

- We formulated a new image-prior based on Pearson type VII densities integrated with a MRF. Our main motivation has been to exploit the heavy-tail property of this density, which indeed seems to be a good way of preserving edges while imposing smoothness. The form of this prior has the additional advantage of allowing us to perform fully automated hyper-parameter estimation. Our recovery algorithm, although very simple to implement, achieves statistically significant improvements over Bayesian Compressive Sensing in under-determined problem settings, and is able to recover more textured images than Bayesian Compressive Sensing can.

- We devised and employed a *similarity-prior* based on Pearson type VII with a Markov Random Field to include the similarity information between two consecutive scenes that differ in colouring or lighting. This prior enables us to recover from fewer measurements than a general-purpose prior would, and can be applied, e.g. in medical imaging applications. We tested our methods both on 1D and 2D

signals in highly under-determined systems. Our results show that when the recovery algorithm received *useful* additional information, we were able to recover good quality signals from their linear transforms using either compressive matrices or classical super-resolution matrices $\boldsymbol{W}$ in highly under-determined conditions. We also demonstrated in low noise case that our automated parameter estimation is superior with respect to MSE in comparison with the well-known 5-folds cross validation method.

- We presented a new approach to multi-task signal recovery where the target signals need not have any overlap in their content but only share their higher level statistical characteristics. This can be used for simultaneous recovery of sets of natural images in a single run. We compared our approach with multi-task BCS, which is the state-of-the-art for multi-task signal recovery and we highlighted the settings in which our approach is advantageous.

7.2 Further Works

There are many avenues for future works for examples:

- Although the prior enables us to recover the high resolution image, the algorithm is still quite sensitive to the initialisation of the parameters. On the good side, we have a flexible model, but on the other side it has a wiggly objective function that is non-convex and hence it is hard to optimise. So far in our work, we found out empirically how to initialise the parameters but in order to remedy this issue, we would need a global optimisation technique. Evolutionary Algorithms (EA) are heuristic global optimisers that have the ability to find good quality solution (approximate solution) to difficult optimisation problem. However, the performance in EA is degrades in high dimensional problems. Indeed, scaling up evolutionary algorithms to high dimensions is recognise to be a major challenge and contests

are organised at the CEC conference[1]. The larger the dimensionality in the latest competition was 1000. This would corresponds to recovering an image no larger than 31×32. However in this thesis, we tackled a problem for 100×100 images where the number of pixels are 10 times bigger than the largest problem so far attempted by the best evolutionary algorithm in the competition.

- Image recovery from compressive or low information content measurements has a number of applications in areas as diverse as medical imaging and video surveillance. Deploying our results to such real-world applications would be an interesting direction for further work.

- A more distant future goal would be to incorporate super-resolution or compressive recovery into other probabilistic models that are used for clustering, visualisation, or data mining. This would allow taking low quality data and coming up with high quality cluster prototypes for example. We believe our probabilistic framework makes a step towards such applications.

[1] http://staff.ustc.edu.cn/~ketang/cec2012/cec2012lsgo.htm

LIST OF REFERENCES

[1] A. El-Baz and A. A. Farag, "Parameter estimation in gibbs-markov image models," in *Proceedings of the Sixth International Conference of Information Fusion*, vol. 2, pp. 934–942, 2003.

[2] S. Ji, Y.Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, pp. 2346–2356, June 2008.

[3] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, pp. 92–106, Jan. 2009.

[4] L. Pickup, *Machine Learning in Multi-frame Image Super-resolution*. PhD thesis, University of Oxford, Nov 2008.

[5] J.Yang and T.Huang, *Super-Resolution Imaging*. CRC Press, 2011.

[6] DL.Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, pp. 1289–1306, April 2006.

[7] E. Candés and M. Waki, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, March 2008.

[8] L. Brown, "A survey of image registration techniques," in *ACM Computer Surveys*, vol. 24, pp. 325–376, Dec. 1992.

[9] V. Dvorchenko, "Bounds on (deterministic)correlation functions with applications to registration," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 5, no. 2, pp. 206–213, 1983.

[10] Q. Tian and M. Huhns, "Algorithm for subpixel registration," in *Proc. in Computer Vision, Graphics, Image Proc.*, vol. 35, pp. 220–233, 1986.

[11] W. Xu, M. Wang, and A. Tang, "Sparse recovery from nonlinear measurements with applications in bad data detection for power networks," 2011.

[12] T. Blumensath, "Compressed sensing with nonlinear observations and related nonlinear optimisation problems," *IEEE Transactions on Information Theory*, vol. PP, p. 1, Feb. 2013.

[13] S. I. Kabanikhin, "Definitions and examples of inverse and ill-posed problems," *J. Inv. Ill-Posed Problems*, vol. 16, p. 317357, 2008.

[14] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011.

[15] T. Soong, *Fundamentals of Probability and Statistics for Engineers*. Wiley, 2004.

[16] P. Cheeseman and J. Stutz, "On the relationship between bayesian and maximum entropy inference," in *In 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings*, vol. 735, 2004.

[17] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of Mathematical Psychology*, vol. 47, pp. 90–100, 2003.

[18] P. Cheeseman, B. Kanefsky, R. Kraft, and J. Stutz, "Super-resolved surface reconstruction from multiple images," tech. rep., NASA Ames Research Center, Dec. 1994.

[19] M.Elad and A.Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, 1997.

[20] H. Zhu, Y. Lu, and Q. Wu, "Super-resolution image restoration by maximum likelihood method and edge-orient diffusion," in *Proc. in International Symposium on Photoelectronic Detection and Imaging (SPIE)*, vol. 6625, 2008.

[21] D. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 627634, 2001.

[22] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur," *IEEE Transactions on Image Processing*, vol. 10, pp. 1187–1193, Aug. 2001.

[23] N. Nguyen, P. Milanfar, and G. Golub, "A computationally efficient superresolution image reconstruction algorithm," tech. rep., 2000.

[24] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1303–1310, MIT Press, 2003.

[25] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[26] D. Capel, *Image Mosaicing and Super-resolution*. PhD thesis, Department of Engineering Science, 2001.

[27] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1 ed., Jan. 1996.

[28] M. E. Tipping, *Bayesian Inference: An Introduction to Principles and Practice in Machine Learning*. 2004. Advanced Lectures on Machine Learning.

[29] B. J, "The case for objective bayesian analysis," *Bayesian Analysis*, vol. 1, no. 3, pp. 385–402, 2006.

[30] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[31] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, pp. 996–1011, June 1996. Member IEEE.

[32] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Processing*, vol. 6, pp. 621–633, Dec 1997.

[33] S. Borman and R. Stevenson, "Simultaneous multi-frame map super-resolution video enhancement using spatio-temporal priors," in *In Proceedings of the IEEE International Conference on Image Processing*, 1999.

[34] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to superresolution," in *Proceedings of the SPIE, Applications of Digital Image Processing XXVI*, vol. 5203, pp. 121–130, 2003.

[35] H. He and L. P. Kondi, "Map based resolution enhancement of video sequences using a huber-markov random field image prior model," in *IEEE Conference of Image Processing*, pp. 933–936, 2003.

[36] H. He and L. P. Kondi, "Choice of threshold of the huber-markov prior in map based video resolution enhancement," in *IEEE Electrical and Computer Engineering Canadian Conference*, vol. 2, pp. 801–804, Nov. 2004.

[37] F. Schubert and K. Mikolajczyk, "Combining high-resolution images with low-quality videos," in *BMVC*, 2008.

[38] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc. Ser. B*, vol. 36, p. 111147, 1974. With discussion and a reply by the authors.

[39] D. A. Dickey, "Introduction to predictive modeling with examples," in *SAS Global Forum*, Apr. 2012.

[40] A.Sylvain and C.Alain, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[41] K. Hashimoto, *Statistical Models of Machine Translation, Speech Recognition and Speech Synthesis for Speech-to-Speech Translation.* PhD thesis, Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Jan. 2011.

[42] R. Gutierrez-Osuna, "Validation." online.

[43] L. P. Devroye and T.J.Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory*, vol. 25, pp. 601–604, 1979.

[44] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[45] A. W. Moore, "Cross-validation for detecting and preventing overfitting." online.

[46] P. Refaeilzadeh, L.Tang, and H.Liu, "Cross-validation," in *In Proceedings of Encyclopedia of Database Systems*, pp. 532–538, 2009.

[47] S.Chevalier, E.Geoffrois, F.Preteux, and M.Lemaitre, "A generic 2d approach to handwriting recognition," in *Proc. 8th Inter. Conf. Document Analysis and Recognition*, pp. 489–493, 2005.

[48] M.Berthod, Z.Kato, S.Yu, and J.Zerubia, "Bayesian image classification using markov random fields," *Image and Vision Computing*, vol. 14, p. 285295, May 1996.

[49] M.Gomez and R.A.Salinas, "A new technique for texture classification using markov random fields," *International Journal of Computers, Communications & Control*, vol. 1, pp. 41–51, 2006.

[50] S.Z.Li, *Markov Random Field Modeling in Computer Vision*. Springer, 1995.

[51] S.Geman and D.Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[52] F. Guichard and L. Rudin, "Image frame fusion by velocity estimation using region merging," 1997. US Patent 5,909,251.

[53] B. C. Tom and A. K. Katsaggelos, "Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images," in *In Proceedings of the IEEE International Conference on Image Processing (ICIP)*, p. 25392542, 1995.

[54] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, pp. 1–8, 2007.

[55] D. Allcroft and C. Glasbey, "A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 52, p. 487498, Oct 2003.

[56] A.Brezger, L.Fahrmeir, and A. Hennerfeind, "Adaptive gaussian markov random fields with applications in human brain mapping," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 56, p. 327345, 2007.

[57] J.Lindstrm and F.Lindgren, *A Gaussian Markov Random Field Model for Total Yearly Precipitation Over the African Sahel*. Lund University, 2008.

[58] Y.Yue and P.L.Speckman, "Nonstationary spatial gaussian markov random fields," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 96–116, 2010.

[59] K. Hanson and G. Wecksung, "Bayesian approach to limited-angle reconstruction in computed tomography," *Journal of the Optical Society of America*, vol. 73, p. 1501, 1983.

[60] E. Kaltenbacher and R. C. Hardie, "High resolution infrared image reconstruction using multiple, low resolution, aliased frames," in *In Proceedings of the IEEE National Aerospace Electronics Conference*, vol. 2, p. 702709, 1996.

[61] B.R.Hunt, "Bayesian methods in nonlinear digital image restoration," *IEEE Transactions on Computer*, vol. C-26, pp. 219–229, 1997.

[62] R. C. Hardie, K. J. Barnardb, J. G. Bognarc, E. E. Armstrongc, and E. A. Watsonb, "High resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system," *Optical Engineering 37*, vol. 37, pp. pp. 247–260, April 1998. Revised July 1997 - Winner of the 1998 Rudolf Kingslake Medal and Prize (selected by Awards Committee as the most noteworthy original paper to appear in Optical Engineering in 1998).

[63] R.C.Hardie and D.R.Droege, "A map estimator for simultaneous superresolution and detector nonunifomity correction," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, April 2007.

[64] S. Babacan, R.Molina, and A. Katsaggelos, "Generalized gaussian markov random field image restoration using variational distribution approximation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1265–1268, 2008.

[65] *Chap. 7 : The Central Gaussian, or normal distribution.* Dec. 1996.

[66] L.C.Pickup, S.J.Roberts, and A. Zisserman, "A sampled texture prior for image super-resolution," in *Advances in Neural Information Processing Systems*, pp. 1587–1594, Advances in Neural Information Processing Systems 16 (NIPS 2003), 2003.

[67] L. C. Pickup, D. P. Capel, S. J. Roberts, and A. Zisserman, "Bayesian methods for image super-resolution," *The Computer Journal*, vol. 52, pp. 101–113, Sept. 2009.

[68] L. Pickup, D. Capel, S. Roberts, and A. Zisserman, "Bayesian image super-resolution, continued," *NIPS*, pp. 1089–1096, 2006.

[69] D. Capel and A. Zisserman, "Automated mosaicing with super-resolution zoom," in *In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 885891, 1998.

[70] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," in *In Proceedings of the International Conference on Pattern Recognition*, vol. 1, p. 16001605, 2000.

[71] D. Capel and A. Zisserman, "Computer vision applied to superresolution," *IEEE Signal Processing Magazine*, vol. 20, no. 3, p. 7586, 2003.

[72] S. Pelletier and J. R. Cooperstock, "Fast image restoration with the huber-markov prior model," in *International Conference Image Processing*, (McGill University Department of Electrical and Computer Engineering Montreal, H3A 2A7, Canada), 2008.

[73] E. Candés and J. Caltech, *L1-magic : Recovery of Sparse Signals via Convex Programming*, October 2005.

[74] C. M.Bishop, *Pattern Recognition and Machine Learning*. Cambridge,UK: Springer, 6th ed., February 2007.

[75] D. Peel and G. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, p. 339348, 2000.

[76] L. C. Pickup, S. J. Roberts, and A. Zisserman, "Optimizing and learning for super-resolution," in *Proceedings of the British Machine Vision Conference*, 2006.

[77] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multi-frame super-resolution," *IEEE Transactions on Image Processing*, 2003.

[78] X. Zhang and E. Y. Lam, "Superresolution reconstruction using nonlinear gradient-based regularization," *Multidimensional Systems and Signal Processing*, vol. 20, pp. 375–384, December 2008.

[79] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research 13*, vol. 13, pp. 281–305, Feb. 2012.

[80] A.Kában, "Estimation of the regularization parameter in huber-mrf for image resolution enhancement.," tech. rep., University of Birmingham, 2011. http://www.cs.bham.ac.uk/∼axk/huber.pdf.

[81] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.

[82] S. LLC, "Bayesian inference." online, 2012.

[83] Y.Nagahara, "Non-gaussian distribution for stock returns and related stochastic differential equation.," *Asia-Pacific Financial Markets*, vol. 3, no. 2, pp. 121–149, 1996.

[84] P. S. Prevéy, "The use of pearson vii distribution functions in x-ray diffraction residual stress measurement.," *Advances in X-Ray Analysis*, vol. 29, pp. 103–111, 1986.

[85] J.Sun, A. Kabán, and J.Garibaldi, "Robust mixture modeling using the pearson type vii distribution.," *Pattern Recognition Letters.*, vol. 31, pp. 2447–2454, Dec 2010.

[86] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G.Baraniuk., "Signal processing with compressive measurements.," *IEEE Journal of Selected Topics in Signal Processing*, 2009.

[87] K. Pearson, "Mathematical contributions to the theory of evolution, xix: Second supplement to a memoir on skew variation.," *Transactions of the Royal Society of London*, vol. 216, pp. 429–457, 1916. Series A, Containing Papers of a Mathematical or Physical Character.

[88] Y. Qi, T. Minka, R. Picard, and Z. Ghahramani., "Predictive automatic relevance determination by expectation propagation.," in *International Conference Machine Learning*, 2004.

[89] M. Nikolova, "Local strong homogeneity of a regularised estimator," *Siam Journal of Applied Mathematics*, vol. 61, no. 2, pp. 633–658, 2000.

[90] A.Kabán and S.AliPitchay, "Single-frame image recovery using a pearson type vii mrf," *Neurocomputing*, vol. 80, pp. 111–119, March 2012. MLSP2010 special issue.

[91] L. Anamy, *Introduction to the Design and Analysis of Algorithms*. Addison Wesley, 3rd ed., 2011.

[92] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," *IEEE Trans. on Signal Processing*, vol. 58, Sept. 2010.

[93] J. Giraldo, J. Trzasko, S. Leng, C. McCollough, and A. Manduca, "A non-convex prior image constrained compressed sensing (nc-piccs).," in *Proc. of SPIE : Physics of Medical Imaging*, vol. 7622, 2010.

[94] R.G.Baraniuk, V.Cevher, M.F.Duarte, and C.Hegde, "Model-based compressive sensing," *IEEE Trans. Information Theory*, vol. 56, pp. 1982–2001, Dec. 2010.

[95] E.Candes, J.Romberg, and T.Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, p. 489 509, Feb. 2006.

[96] W. Lu and Vaswani, "Regularized modified bpdn for noisy sparse reconstruction with partial erroneous support and signal value knowledge," *IEEE Trans. on Signal Processing*, vol. 60, pp. 182–196, Oct 2011.

[97] S. Pitchay and A. Kabán, "Single-frame signal recovery using a similarity-prior based on pearson type vii mrf," in *1st International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pp. 123–133, SciTePress Digital Library, Feb. 2012.

[98] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*. W.H. Freeman & Co, 5th ed., 2006. Chap 15: Nonparametric Tests.

Appendices

APPENDIX A

Neighbourhood feature

Using the example size of image $\boldsymbol{z}$ (3×4), the neighbourhood feature $\boldsymbol{Dz}$ is computed as follows:

$$\boldsymbol{Dz} = D_{ij} \times [z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_9, z_{10}, z_{11}, z_{12}]^T$$

$$=$$

$$\left[\begin{pmatrix} D_{11}, & D_{12}, & . & . & D_{1,12} \\ D_{21}, & D_{22}, & . & . & D_{2,12} \\ ... & & & & \\ ... & & & & \\ D_{12,1}, & D_{12,2}, & . & . & D_{12,12} \end{pmatrix} \times \begin{pmatrix} z_1 \\ z_2 \\ . \\ . \\ z_{12} \end{pmatrix} \right]$$

=

$$\begin{bmatrix}
(1.z_1) + (-\frac{1}{4}.z_2) + (0.z_3) + (-\frac{1}{4}.z_4) + (0.z_5) + (0.z_6) + (0.z_7) + (0.z_8) + (0.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(-\frac{1}{4}.z_1) + (1.z_2) + (-\frac{1}{4}.z_3) + (0.z_4) + (-\frac{1}{4}.z_5) + (0.z_6) + (0.z_7) + (0.z_8) + (0.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(0.z_1) + (-\frac{1}{4}.z_2) + (1.z_3) + (0.z_4) + (0.z_5) + (-\frac{1}{4}.z_6) + (0.z_7) + (0.z_8) + (0.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(-\frac{1}{4}.z_1) + (0.z_2) + (0.z_3) + (1.z_4) + (-\frac{1}{4}.z_5) + (0.z_6) + (-\frac{1}{4}.z_7) + (0.z_8) + (0.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(0.z_1) + (-\frac{1}{4}.z_2) + (0.z_3) + (-\frac{1}{4}.z_4) + (1.z_5) + (-\frac{1}{4}.z_6) + (0.z_7) + (-\frac{1}{4}.z_8) + (0.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(0.z_1) + (0.z_2) + (-\frac{1}{4}.z_3) + (0.z_4) + (-\frac{1}{4}.z_5) + (1.z_6) + (0.z_7) + (0.z_8) + (-\frac{1}{4}.z_9) + (0.z_{10}) + (0.z_{11}) + (0.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (-\frac{1}{4}.z_4) + (0.z_5) + (0.z_6) + (1.z_7) + (-\frac{1}{4}.z_8) + (0.z_9) + (-\frac{1}{4}.z_10) + (0.z_{11}) + (0.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (0.z_4) + (-\frac{1}{4}.z_5) + (0.z_6) + (-\frac{1}{4}.z_7) + (1.z_8) + (-\frac{1}{4}.z_9) + (0.z_{10}) + (-\frac{1}{4}.z_{11}) + (0.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (0.z_4) + (0.z_5) + (-\frac{1}{4}.z_6) + (0.z_7) + (-\frac{1}{4}.z_8 + (1.z_9) + (0.z_10) + (0.z_{11}) + (-\frac{1}{4}.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (0.z_4) + (0.z_5) + (0.z_6) + (-\frac{1}{4}.z_7) + (0.z_8) + (0.z_9) + (1.z_{10}) + (-\frac{1}{4}.z_{11}) + (0.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (0.z_4) + (0.z_5) + (0.z_6) + (0.z_7) + (-\frac{1}{4}.z_8) + (0.z_9) + (-\frac{1}{4}.z_{10}) + (1.z_{11}) + (-\frac{1}{4}.z_{12}) \\
(0.z_1) + (0.z_2) + (0.z_3) + (0.z_4) + (0.z_5) + (0.z_6) + (0.z_7) + (0.z_8) + (-\frac{1}{4}.z_9) + (0.z_{10}) + -\frac{1}{4}.z_{11}) + (1.z_{12})
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix} z_1 - \frac{1}{4}.z_2 - \frac{1}{4}.z_4 \\ -\frac{1}{4}.z_1 + z_2 - \frac{1}{4}.z_3 - \frac{1}{4}.z_5 \\ -\frac{1}{4}.z_2 + z_3 - \frac{1}{4}.z_6 \\ -\frac{1}{4}.z_1 + z_4 - \frac{1}{4}.z_5 - \frac{1}{4}.z_7 \\ -\frac{1}{4}.z_2 - \frac{1}{4}.z_4 + z_5 - \frac{1}{4}.z_6 - \frac{1}{4}.z_8 \\ -\frac{1}{4}.z_3 - \frac{1}{4}.z_5 + z_6 - \frac{1}{4}.z_9 \\ -\frac{1}{4}.z_4 + z_7 - \frac{1}{4}.z_8 - \frac{1}{4}.z_10 \\ -\frac{1}{4}.z_5 - \frac{1}{4}.z_7 + z_8 - \frac{1}{4}.z_9 - \frac{1}{4}.z_{11} \\ -\frac{1}{4}.z_6 - \frac{1}{4}.z_8 + z_9 - \frac{1}{4}.z_{12} \\ -\frac{1}{4}.z_7 + z_{10} - \frac{1}{4}.z_{11} \\ -\frac{1}{4}.z_8 - \frac{1}{4}.z_{10} + z_{11} - \frac{1}{4}.z_{12} \\ -\frac{1}{4}.z_9 - \frac{1}{4}.z_{11} + z_{12} \end{bmatrix}$$

$$=$$

$$\begin{bmatrix} z_1 - \frac{1}{4}(z_2 + z_4) \\ z_2 - \frac{1}{4}(z_1 + z_3 + z_5) \\ z_3 - \frac{1}{4}(z_2 + z_6) \\ z_4 - \frac{1}{4}(z_1 + z_5 + z_7) \\ z_5 - \frac{1}{4}(z_2 + z_4 + z_6 + z_8) \\ z_6 - \frac{1}{4}(z_3 + z_5 + z_9) \\ z_7 - \frac{1}{4}(z_4 + z_8 + z_{10}) \\ z_8 - \frac{1}{4}(z_5 + z_7 + z_9 + z_{11}) \\ z_9 - \frac{1}{4}(z_6 + z_8 + z_{12}) \\ z_{10} - \frac{1}{4}(z_7 + z_{11}) \\ z_{11} - \frac{1}{4}(z_8 + z_{10} + z_{12}) \\ z_{12} - \frac{1}{4}(z_9 + z_{11}) \end{bmatrix}$$

From the details above, we can see only z_5 and z_8 contain four neighbours which can be simplify into this notation $z_i - \frac{1}{4}\sum_{j \in 4neigb(i)} z_j$. However the real data is an image with higher dimensionality where there are more intensity value in $\mathbf{z}$ with four neighbours compared the pixels with two and three neighbours. Therefore, we can neglect the minority pixels since it does not affect the majority difference.

APPENDIX B

Derivation

B.0.1 Derivative of σ

The objective function for σ is written in (6.17) and we solve it as follow. Notice that $\frac{M}{2}\log(2\pi)$ does not depend on the σ^2, therefore we can ignore this as a constant and remains $\frac{1}{2}\log[\Sigma]$. Denoting the likelihood term that depends on σ^2 by $Obj(\sigma^2)$,

$$
\begin{aligned}
Obj(\sigma^2) &= \sum_{k=1}^{K}\left\{\frac{1}{2\sigma^2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2 + M\log\pi + \frac{1}{2}\log[\Sigma]\right\} \\
\frac{\partial Obj(\sigma^2)}{\partial\left(\frac{1}{\sigma^2}\right)} &= \sum_{k=1}^{K}\left\{\frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2 + 0 + \frac{1}{2}\frac{\partial[\log\Sigma]}{\partial\left(\frac{1}{\sigma^2}\right)}\right\} \text{ where } \Sigma = \sigma^2 I \\
&= \sum_{k=1}^{K}\left\{\frac{1}{2}\log\left(\det(\sigma^2 I)\right) + \frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2\right\} \\
&= \sum_{k=1}^{K}\left\{\frac{1}{2}\log\left\{(\sigma^2)^M(\det(I))\right\} + \frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2\right\} \\
&= \sum_{k=1}^{K}\left\{\left(\frac{M}{2}\log\sigma^2\right)(1) + \frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2\right\} \\
&= \sum_{k=1}^{K}\left\{\left(-\frac{M}{2}\log\frac{1}{\sigma^2}\right) + \frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2\right\} \\
&= \sum_{k=1}^{K}\left\{-\frac{M}{2}\sigma^2 + \frac{1}{2}\left(y^{(k)} - W^{(k)}\theta^{(k)}\right)^2\right\}
\end{aligned}
$$

Equate the above to zero, yields a closed form estimate for σ^2 in (6.18) and the term that depends on σ^2 for multi-task recovery is written by:

$$\begin{aligned}
\sum_{k=1}^{K}\left\{-\frac{1}{2}M\sigma^2+\frac{1}{2}\left(y^{(k)}-W^{(k)}\theta^{(k)}\right)^2\right\} &= 0\\
\sum_{k=1}^{K}\left\{\frac{1}{2}\left(-M\sigma^2+\left(y^{(k)}-W^{(k)}\theta^{(k)}\right)^2\right)\right\} &= 0\\
\sum_{k=1}^{K}\left\{-M\sigma^2+\left(y^{(k)}-W^{(k)}\theta^{(k)}\right)^2\right\} &= 0\\
\sigma^2=\sum_{k=1}^{K}\left\{\frac{1}{M}\left(y^{(k)}-W^{(k)}\theta^{(k)}\right)^2\right\}&
\end{aligned}$$

B.0.2 Derivative of λ

There are two terms that involve λ from this equation (6.19). Recall the objective function of λ as follow:

$$Obj_{(\lambda)}=\sum_{k=1}^{K}\left\{\underbrace{\frac{N\nu}{2}\log\lambda}_{term1}\underbrace{-\frac{1+\nu}{2}\sum_{i=1}^{N}\log[(\theta_i^{(k)})^2+\lambda]}_{term2}\right\}$$

Then we solve the derivative for the first term of λ. Denote $\sqrt{\lambda}$=a, therefore $\lambda=a^2$.

$$\begin{aligned}
\frac{\partial Obj_{term1}(a^2)}{\partial a} &= \sum_{k=1}^{K}\left\{\frac{N\nu}{2}\frac{\partial\log a^2}{\partial a}\right\}\\
&= \sum_{k=1}^{K}\left\{\frac{N\nu}{2}\frac{1}{a^2}\frac{\partial a^2}{\partial a}\right\}\\
&= \sum_{k=1}^{K}\left\{\frac{N\nu}{2}\frac{1}{a^2}2a\right\}\\
&= \sum_{k=1}^{K}\left\{\frac{N\nu}{a}\right\}\\
&= \sum_{k=1}^{K}\left\{\frac{N\nu}{\sqrt{\lambda}}\right\}
\end{aligned}$$

Now, solve the second term that consists of λ:

$$\begin{aligned}
\frac{\partial Obj_{term2}(a^2)}{\partial a} &= \sum_{k=1}^{K}\left\{-\frac{1+\nu}{2}\sum_{i=1}^{N}\frac{\partial \log\{\theta_i^2+a^2\}}{\partial a}\right\} \\
&= \sum_{k=1}^{K}\left\{-\frac{1+\nu}{2}\sum_{i=1}^{N}\frac{1}{\theta_i^2+a^2}2a\right\} \\
&= \sum_{k=1}^{K}\left\{-\frac{1+\nu}{2}\sum_{i=1}^{N}\frac{2\sqrt{\lambda}}{\theta_i^2+\lambda}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{i=1}^{N}-\frac{(1+\nu)\sqrt{\lambda}}{\theta_i^2+\lambda}\right\} \\
&= \sum_{k=1}^{K}\left\{\frac{-\sqrt{\lambda}-\nu\sqrt{\lambda}}{\theta_i^2+\lambda}\right\}
\end{aligned}$$

Finally sum up both terms, the derivative of $\sqrt{\lambda}$ can be obtained as in (6.20):

$$\begin{aligned}
\frac{\partial Obj(\lambda)}{\partial\sqrt{\lambda}} &= \sum_{k=1}^{K}\left\{\frac{N\nu}{\sqrt{\lambda}}\sum_{i=1}^{N}\frac{-\sqrt{\lambda}-\nu\sqrt{\lambda}}{\theta_i^2+\lambda}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{i=1}^{N}\frac{\nu(\theta_i^2+\lambda)+(\sqrt{\lambda}\times(-\sqrt{\lambda}))+(\sqrt{\lambda}\times-(\nu\sqrt{\lambda}))}{(\theta_i^2+\lambda)\sqrt{\lambda}}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{i=1}^{N}\frac{\nu\theta_i^2+\nu\lambda-\lambda-\nu\lambda}{(\theta_i^2+\lambda)\sqrt{\lambda}}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{i=1}^{N}\frac{\nu\theta_i^2-\lambda}{(\theta_i^2+\lambda)\sqrt{\lambda}}\right\}
\end{aligned}$$

B.0.3 Derivative of ν

There are four terms that involve ν from this equation (6.21). Recall the objective function of ν as follow.

$$Obj_{(\nu)} = \sum_{k=1}^{K}\left\{\underbrace{N\log\Gamma\left(\frac{1+\nu}{2}\right)}_{term1}\underbrace{-N\log\Gamma\left(\frac{\nu}{2}\right)}_{term2}+\underbrace{\frac{N\nu}{2}\log\lambda}_{term3}\underbrace{-\frac{1+\nu}{2}\sum_{i=1}^{N}\log[(\theta_i^{(k)})^2+\lambda]}_{term4}\right\}$$

Again, solve the first term that consists of ν. Denotes the $\sqrt{\nu}$=b, therefore $\nu = b^2$.

$$\begin{aligned}
\frac{\partial Obj_{term1}(b^2)}{\partial b} &= \sum_{k=1}^{K}\left\{N\frac{\partial \log\Gamma\left(\frac{1+b^2}{2}\right)}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{N\psi\left(\frac{1+b^2}{2}\right)\times\partial\frac{\left(\frac{1+b^2}{2}\right)}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{N\psi\left(\frac{1+b^2}{2}\right)\times\frac{2b}{2}\right\} \\
&= \sum_{k=1}^{K}\left\{N\psi\left(\frac{1+\nu}{2}\right)\sqrt{\nu}\right\}
\end{aligned}$$

Again, solve the derivative of the second term that consists of ν:

$$\begin{aligned}
\frac{\partial Obj_{term2}(b^2)}{\partial b} &= \sum_{k=1}^{K}\left\{-N\frac{\partial \log\Gamma\left(\frac{b^2}{2}\right)}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{-N\psi\frac{b^2}{2}\frac{\partial\frac{b^2}{2}}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{-N\psi\frac{b^2}{2}\frac{2b}{2}\right\} \\
&= \sum_{k=1}^{K}\left\{-N\psi\frac{\nu}{2}\sqrt{\nu}\right\}
\end{aligned}$$

The derivative of the third term that consists of ν, yields:

$$\begin{aligned}
\frac{\partial Obj_{term3}(b^2)}{\partial b} &= \sum_{k=1}^{K}\left\{\frac{\partial\frac{Nb^2}{2}\log\lambda}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{\frac{N2b}{2}\log\lambda\right\} \\
&= \sum_{k=1}^{K}\left\{N\sqrt{\nu}\log\lambda\right\}
\end{aligned}$$

Next, the derivative of the final term that consists of ν:

$$\begin{aligned}
\frac{\partial Obj_{term4}(b^2)}{\partial b} &= \sum_{k=1}^{K}\left\{\frac{\partial\left(-\frac{1+b^2}{2}\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right)}{\partial b}\right\} \\
&= \sum_{k=1}^{K}\left\{-\frac{2b}{2}\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right\} \\
&= \sum_{k=1}^{K}\left\{-b\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right\} \\
&= \sum_{k=1}^{K}\left\{-\sqrt{\nu}\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right\}
\end{aligned}$$

Finally, summing up all the terms, the derivative of $\sqrt{\nu}$ can be obtained as in (6.22):

$$\begin{aligned}
\frac{\partial Obj(\nu)}{\partial\sqrt{\nu}} &= \sum_{k=1}^{K}\left\{N\psi\left(\frac{1+\nu}{2}\right)\sqrt{\nu}-N\psi\frac{\nu}{2}\sqrt{\nu}+N\sqrt{\nu}\log\lambda-\sqrt{\nu}\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right\} \\
&= \sum_{k=1}^{K}\left\{\left(N\psi\left(\frac{1+\nu}{2}\right)-N\psi\frac{\nu}{2}+N\log\lambda-\sum_{i=1}^{N}\log(\theta_i^2+\lambda)\right)\sqrt{\nu}\right\}
\end{aligned}$$