UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DIFFERENTIAL PIXEL STATISTIC BASED PRIORS FOR IMAGE RESTORATION

A thesis submitted in partial satisfaction of the
requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING

by

David Odom

June 2006

The Thesis of David Odom
is approved:

Professor Peyman Milanfar, Chair

Professor Hany Farid

Professor Herbert Lee

Lisa C. Sloan
Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

Abstract

Differential Pixel Statistic Based Priors for Image Restoration

by

David Odom

The statistics of natural images play an important role in many image processing tasks. In particular, statistical assumptions about differences between neighboring pixel values are used extensively in the form of prior information for many diverse applications. The most common assumption is that these pixel difference values can be described be either a Laplace or Generalized Gaussian distribution. The statistical validity of these two assumptions is investigated formally in this paper by means of Chi-squared goodness of fit tests. The Laplace and Generalized Gaussian distributions are seen to deviate from real images, with the main source of error being the large number of zero and close to zero nearby pixel difference values. These values correspond to the relatively uniform areas of the image. A mixture distribution is proposed to retain the edge modeling ability of the Laplace or Generalized Gaussian distribution, and to improve the modeling of the effects introduced by smooth image regions. The Chi-squared tests of fit indicate that the mixture distribution offers a significant improvement in fit. A computationally efficient approximation to the mixture model is developed for use as a prior. A denoising application is presented that utilizes this proposed prior on pixel differences. Experiments are performed on real and simulated images, and a comparison is made with several other methods that are based on different statistical assumptions on pixel differences. The results show improved performance by the mixture model assumption.

Dedicated to my parents, Paul and Carolyn,

and my girlfriend, Liz.

Acknowledgments

I would like to thank my advisor, Peyman Milanfar, for providing support and guidance throughout the course of my research. I would also like to thank Dr. Morteza Shahram, Dr. Sina Farisu, and the members of the Multi-Dimensional Signal Processing research group (Amyn Poonawala, Hiro Takeda, and Mike Charest).

I would also like to express my gratitude to all of my friends in Santa Cruz, that made life here a great experience, and to my family who supported me from afar. Also, many thanks to Liz for dealing with me all of the times I got too stressed out about school.

The text of this thesis includes reprints of the following previously published material: Modeling Multiscale Differential Pixel Statistics [1], and Image Denoising Using Mixture Model Priors on Pixel Differences [2].

I want to thank the thesis reading committee (Professor Peyman Milanfar, Professor Hany Farid, and Professor Herbert Lee) for reviewing this thesis and for their valuable feedback.

Santa Cruz, California
June 15th, 2006
David Odom

Chapter 1

Introduction

In this chapter we briefly discuss the study of image statistics. First, a background of previous work on natural image statistics is presented. Next, we focus on the past work in the study of derivative statistics and discuss a few applications of these statistics to image processing problems. Finally, we present the organization of this thesis.

1.1 Natural Image Statistics

Natural images have been studied extensively from a variety of different viewpoints, with different applications in mind. The ability to accurately model natural images provides obvious benefits to image processing tasks such as denoising, compression and texture analysis, among others. The study of biological vision systems is also closely related to the study of natural images. As exact physical models for natural images may not be possible or convenient, statistical approaches may be taken to capture the variability present in these images. The study of natural image statistics arises from efforts to find and explain patterns that are present in natural images.

When studying the statistics of natural images, there are many choices of which particular aspect is under observation. Perhaps the first, most obvious choice would be to study the intensity values of the pixels themselves. However, the pixel intensity values do not display an obvious common structure, as shown in Figure **1.1**. It is not apparent that a simple explanation for these histograms exists. Instead of directly studying pixel values, the pixels are often transformed by some type of projection or filtering. After undergoing certain types of transformations, the resulting coefficients display some obvious patterns and characteristics. A few of the most commonly studied transformations include the wavelet transform [3], Discrete Cosine transform [4], and Fourier transform [5]. These statistics, among others, have been studied extensively, and a survey of this work can be found in [6]. The coefficients that are the particular focus of this thesis are those that arise from taking the difference between two neighboring pixels. The study of the marginal distributions of pixel differences is sometimes referred to as derivative statistics [7], as pixel differences are an approximation to the derivatives of an image.

1.1.1 Derivative Statistics

Differences between neighboring pixels in an image are of particular interest because they have been observed to have remarkably similar histograms across images, while their histograms differ drastically from the histograms of noise images [8]. The histograms of nearest horizontal neighbor pixel differences for the images of Figure **1.2** are shown in Figure **1.3**. Several of the commonly observed properties of pixel difference histograms are apparent in this example. The histograms of pixel differences have been observed to display non-gaussian behavior. The histograms have a high kurtosis, and are often said to have long, heavy, exponential tails. Several attempts have been made to model these derivative statistics with parametric models. Parametric models are of particular interest, because they allow fast and intuitive implementations as priors for image processing problems. Some explicit parametric models have been proposed, and some models have been implied by the use of certain priors.

The derivative statistics have been modeled by the Generalized Gaussian distribution in [3] and utilized in many other works, such as in [7] and in [9]. The Laplace assumption has been proposed explicitly [10], and implicitly by applications such as total variation denoising [11]. A Gaussian assumption also finds use in the form of Tikhonov regularization, despite the observed non-Gaussianity of the histograms. The statistical validity of these various assumptions is thoroughly investigated in this thesis. A new set of distributions, consisting of the mixture of a Gaussian and a Laplacian or Generalized Gaussian, is proposed as an alternate model for the pixel differences.

Another often cited property of natural images is the scale invariance of marginal distributions. This has been noted in [12], and in particular for image gradients in [8]. Scale invariance, in this context, is taken to mean that the marginal distributions of statistics remain unchanged if the image is scaled by a pixel averaging and decimation operation. A different approach is taken in this thesis, in which the pixel differences for pixels with varying spatial distances in between are studied. The histograms appear to show some patterns in the way they change as the distance between the different pixels is taken is increased. This is shown in the histograms of Figures **1.4** and **1.5**, where the position of the histogram relative to each image indicates the distance between pixels. This thesis attempts to describe this behavior in simple parametric terms to facilitate usage of the prior information in a multiscale context.

1.2 Thesis Contribution and Organization

Chapter 2 contains a statistical analysis of pixel differences in an image. Several previously proposed distributions for the pixel differences are presented. A mixture distribution is proposed as an alternate model. All distributions are analyzed by statistical goodness of fit tests, and the proposed mixture model is shown to provide a better fit.

Chapter 3 details the development of a prior from the proposed mixture model. Several approximations must be made to develop a computationally reasonable form that is implementable in image processing tasks. Both a single scale, and multiscale form of the prior are developed.

Chapter 4 describes the application of the mixture model prior in a denoising framework. Several other denoising methods that fit into the same framework are discussed, and related to different priors on pixel differences. A comparison is made between the denoising results of these different priors.

Chapter 5 contains conclusions and future work. The performance of the mixture model is discussed. Potential applications of the mixture model prior are included as future directions.

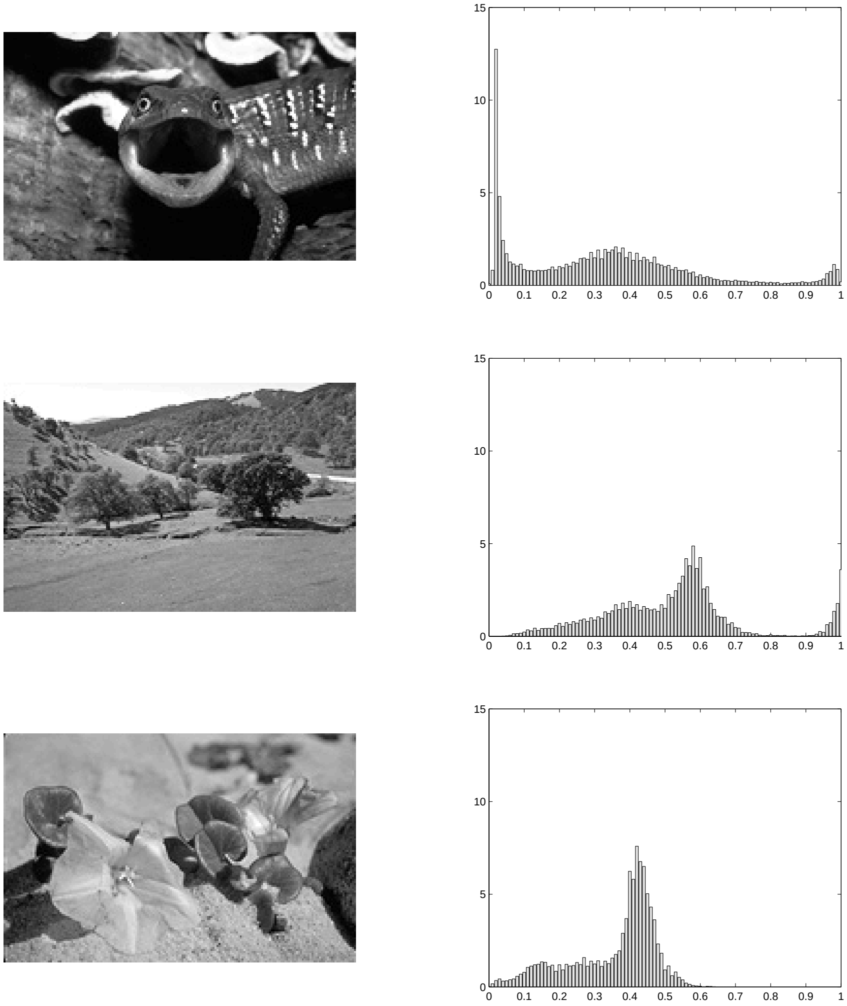


Figure 1.1: Histograms of pixel intensities for different images.

Figure 1.2: Samples of images.

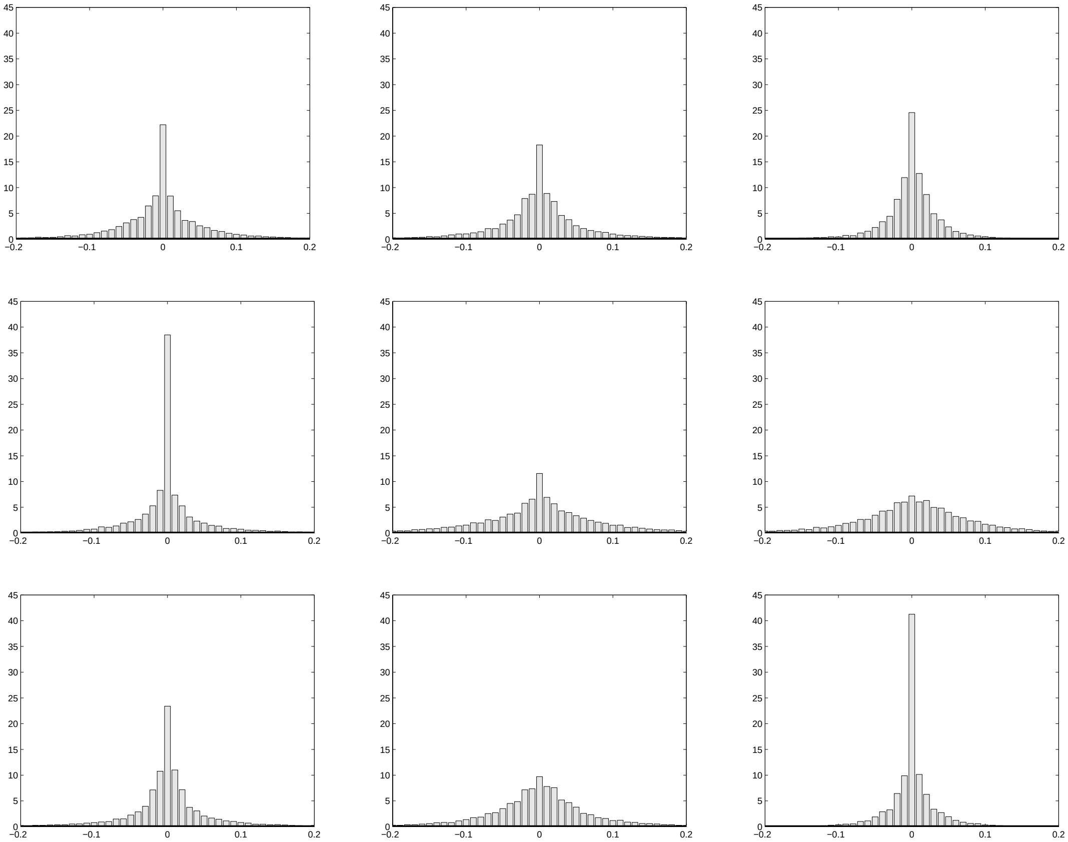


Figure 1.3: Histograms of nearest horizontal neighbor pixel differences for several images.

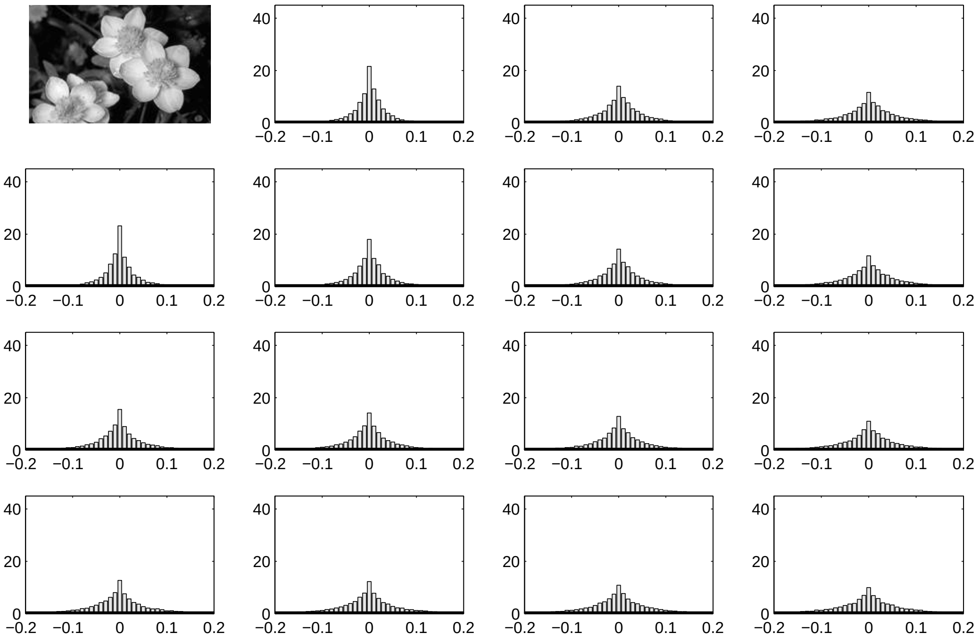


Figure 1.4: Histograms of pixel differences for different distances between pixels. Histogram positioning relative to the image indicates which pixel difference is displayed.

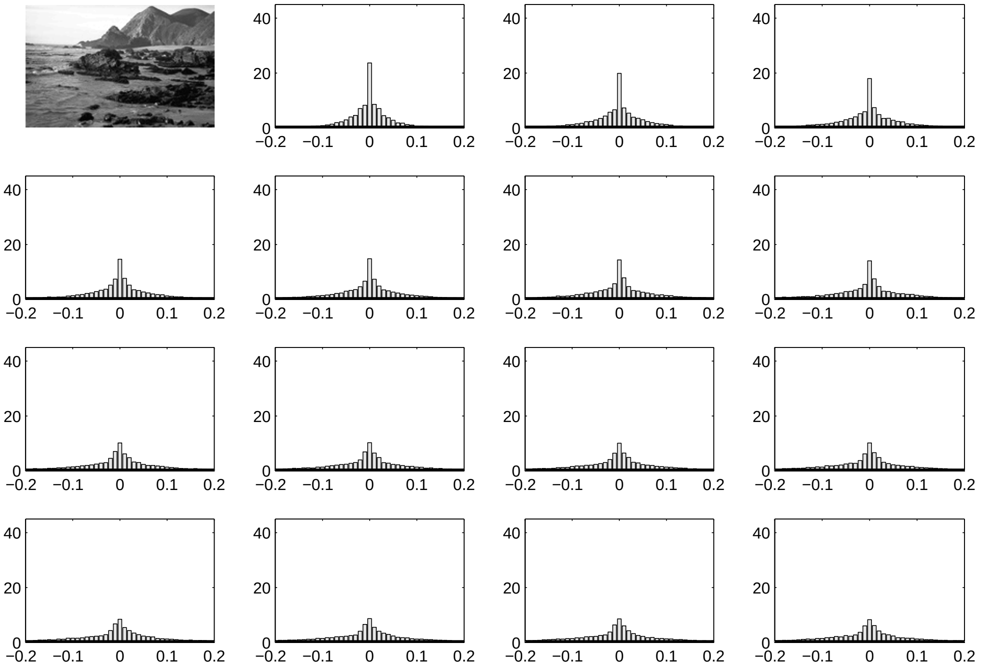


Figure 1.5: Histograms of pixel differences for different distances between pixels. Histogram positioning relative to the image indicates which pixel difference is displayed.

Chapter 2

Pixel Differences and Their Statistics

The statistics of neighboring pixel differences are often called derivative statistics, as pixel differences approximate directional derivatives in an image. These derivative statistics, in terms of differences between neighboring pixels in an image, have been described by several parametric distributions. In this chapter, we discuss these distributions, and propose a new distribution, consisting of the mixture of two component distributions. Goodness of fit tests are performed for each distribution to give a quantitative evaluation of their descriptive abilities.

2.1 Pixel Differences

To obtain the data for which the tests of goodness of fit are to be performed, a sample set of 300 grayscale images with pixel values in the range [0,1] was used. Each image was shifted vertically and horizontally by a maximum of three pixels, and then subtracted from the original. No reflection was used at border regions; the areas with no overlap were simply eliminated from the pixel difference data set. A sampling of 16 images from the test set is shown in Figure **2.1**.

Figure 2.1: Sample of images tested

Following the notation of [13], these pixel differences for a given image, $\mathbf{X}$, can be represented as follows

$$\underline{X}_{(l,m)} = \underline{X} - S_x^l S_y^m \underline{X}, \tag{2.1}$$

where $\underline{X}$ is a column-stacked vector of the original image. S_x^l and S_y^m are matrices that shift $\mathbf{X}$ by l and m pixels in horizontal and vertical directions respectively. $\underline{X}_{(l,m)}$ is the vector form of image $\mathbf{X}$ shifted horizontally by l pixels and vertically by m pixels and subtracted from the unshifted $\mathbf{X}$. For instance, an image from the test database, along with the difference image $\underline{X}_{(1,0)}$, are shown in Figure **2.2**. The histogram of this image is representative of the characteristics observed across a wide range of images. For increasing values of l and m, these pixel differences can also be viewed as neighboring pixel differences of successive pyramid layers, where each layer is formed by unfiltered decimation of a parent layer.

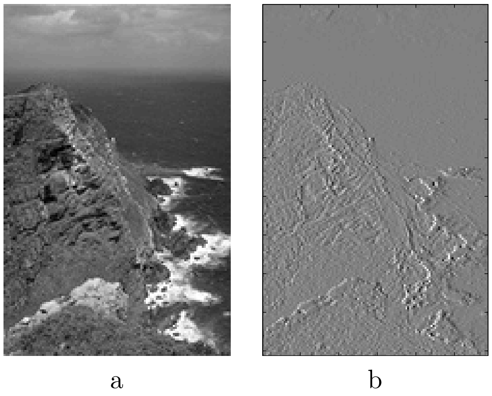

a
b

Figure 2.2: a: Original Image, b: $X_{(1,0)}$

2.2 Distributions

The pixel differences defined above have been modeled by several different distributions in the past. We present these distributions along with a proposed alternative below. All distributions are assumed to be zero-mean[1]. In what follows, we denote the pixels of the *difference* images under study by the random variable $\mathbf{x}$, with n samples denoted by $\{x_i\}$, $1 \leq i \leq n$.

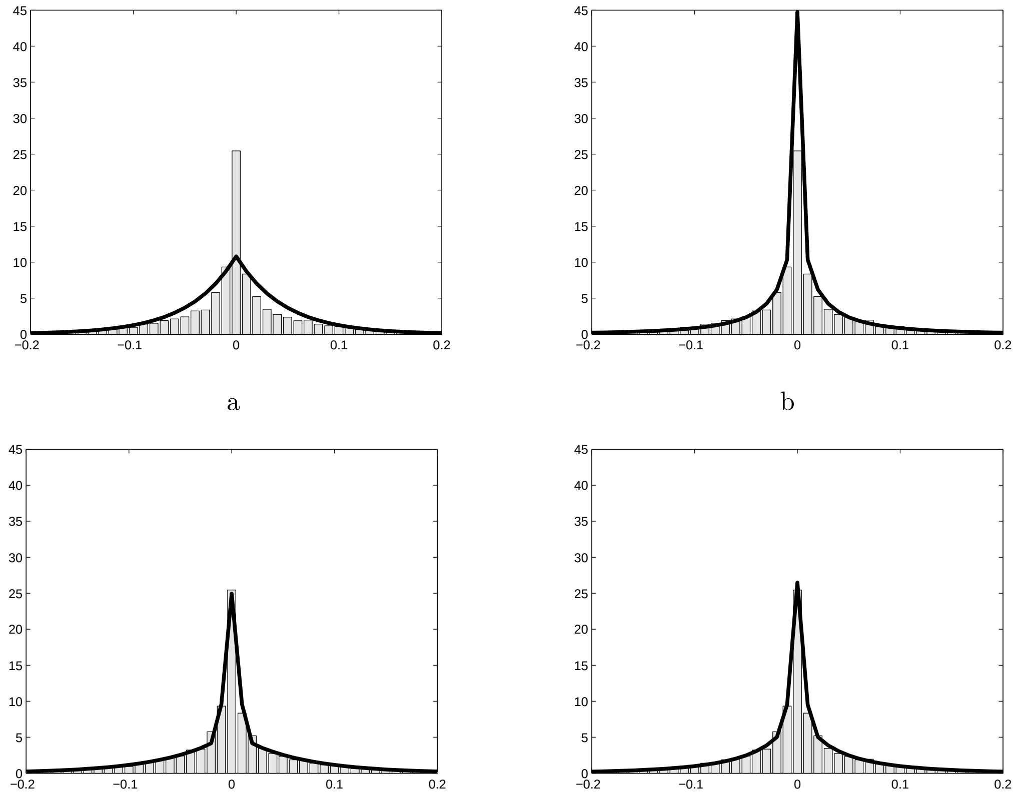
2.2.1 Laplace Distribution

The zero-mean classical Laplace distribution is given by the density function

$$p_L(x; s) = \frac{1}{2s} e^{-|x|/s}, \qquad -\infty < x < \infty \tag{2.2}$$

---

[1] This assumption was confirmed as valid by allowing the mean to vary for each distribution and comparing to the zero-mean case. The estimated means were indeed found to be essentially zero in all cases.



a b

c d

Figure 2.3: Histograms and fits for Fig. **2.2**b. a: Laplace ML estimate, b: Generalized Gaussian ML estimate, c: L-G Mixture model ML estimate, and d: GG-G Mixture model ML estimate

where $s > 0$ is the scale parameter.

The maximum likelihood[2] estimate for the scale parameter is given as follows

$$\hat{s} = \frac{1}{n}\sum_{i=1}^{n}|x_i| \tag{2.3}$$

In Figure **2.3**a, the histogram of pixel difference values for the pixel difference image of Figure **2.2**b, along with the superimposed Laplace ML fit, are shown. Table **2.1**

---

[2] Maximum likelihood (ML) estimators are used due to their desirable properties of consistency, asymptotic normality, and efficiency [14].

contains the average ML parameter estimates for the Laplace distribution across all images in the database.

2.2.2 Generalized Gaussian Distribution

The Generalized Gaussian (GG) Distribution is given by the following density function

$$p_{GG}(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})}\exp\left(-\left|\frac{x}{\alpha}\right|^{\beta}\right) \tag{2.4}$$

For the special case of $\beta = 1$ the GG distribution becomes a Laplace distribution, and for $\beta = 2$, it becomes a Gaussian distribution.

The ML estimate $\hat{\beta}$ is the solution of the following transcendental equation

$$1 + \frac{\psi(1/\hat{\beta})}{\hat{\beta}} - \frac{\sum_{i=1}^{n}|x_i|^{\hat{\beta}}\log|x_i|}{\sum|x_i|^{\hat{\beta}}} + \frac{\log\left(\frac{\hat{\beta}}{n}\sum_{i=1}^{n}|x_i|^{\hat{\beta}}\right)}{\hat{\beta}} = 0 \tag{2.5}$$

The Newton-Raphson iterative procedure as implemented by Do and Vetterli [15] can be used to solve this equation. With $\hat{\beta}$ determined, $\hat{\alpha}$ is given by

$$\hat{\alpha} = \left(\frac{\hat{\beta}}{n}\sum_{i=1}^{n}|x_i|^{\hat{\beta}}\right)^{1/\hat{\beta}} \tag{2.6}$$

In Figure **2.3**b, the histogram of pixel difference values for the pixel difference image of Figure **2.2**b along with the superimposed Generalized Gaussian ML fit are shown. Table **2.2** contains the average ML parameter estimates for the Generalized Gaussian distribution across all images in the database.

2.2.3 Mixture Distributions

Several observations can be made about the pixel difference data and the ML fits for the Laplace and Generalized Gaussian distributions shown above. It appears that the tails of the distributions describe the data well, but the center region is not described well by either distribution. In other words, the detail or edge regions of the image are well described, but the relatively uniform areas are not. On first inspection of the histograms, the pixel differences appear to be impulsive at the origin. However, this is not exactly the case. The center histogram bins contain both zero *and* close to zero pixel differences, and these small nonzero value pixel differences contribute significantly to the center bin. Motivated by these observations, we suggest a new set of distributions which consist of the mixture of a tail-modeling distribution, and a Gaussian distribution to model the more uniform regions.

The mixture distribution is defined as follows,

$$p(x;\underline{\theta}) = Ap_D(x;\Theta) + (1-A)p_G(x;\sigma^2), \tag{2.7}$$

where $\underline{\theta} = [\Theta, \sigma^2, A]$, p_D is some tail-modeling distribution, and p_G is a zero mean Gaussian distribution given by

$$p_G(x_i;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}. \tag{2.8}$$

Observations made above suggest that either a Laplace distribution or a Generalized Gaussian distribution are reasonable choices for p_D.

A further conceptual motivation for the mixture model is derived from the descriptive model for images that consists of a ”cartoon”, or piecewise-smooth component, and a texture component. A detailed analysis of various cartoon/texture decompositions can be found in [16, 17]. These mixture distributions can be related to this model by viewing the Gaussian component as describing the texture, and the Laplace or Generalized Gaussian component describing the cartoon.

The parameters for the mixture distribution given by (2.7) are estimated by using the EM algorithm [18]. Namely, the mixing parameter, A, is fixed, and the ML estimates for the two distributions are calculated. Next, the mixing parameter is estimated with the other distribution parameters fixed at the new estimate. This process is repeated until convergence [18]. A derivation of the ML estimates for the mixture distribution and a further explanation of the EM algorithm are contained in Appendix A.

In Figure **2.3**c, the histogram of pixel difference values for the pixel difference image of Figure **2.2**b, along with the superimposed Laplace/Gauss (L-G) mixture ML fit, are shown. Table **2.3** contains the average ML parameter estimates for the L-G mixture model. Figure **2.3**d contains the Generalized Gaussian/Gauss (GG-G) mixture fit. Table **2.4** contains the average ML parameter estimates for the GG-G mixture model.We observe that these fits are qualitatively superior to the fits of the Laplace or Generalized Gaussian distributions alone. In Section 2.4 we quantitatively verify these observations.

2.3 Parameter Behavior Across Scale

The histograms of pixel differences in natural images have been observed to remain constant across scale, when an averaging pyramid decomposition is used [8]. However, this is not the case for the pixel differences across scale we are describing, which can be viewed as being derived from an unfiltered, decimated pyramid as mentioned in Section 2.1. This choice of scale definition is motivated by a computationally efficient framework for multiscale priors, presented later in Chapter 4.

As the scale of the pixel difference varies from fine to coarse (increasing l and m from Equation 2.1) the parameters of each distribution vary as well. An example of the changes of the shape of the distributions is shown in Figures **2.4**, **2.5**, **2.6**, and **2.7**. In order to understand the overall behavior of these distributions, it is desirable to describe each parameter's behavior as some simple function of scale. Empirically, each parameter was observed to exhibit exponential, logarithmic, or linear behavior as shown in Figure **2.8** and described below.

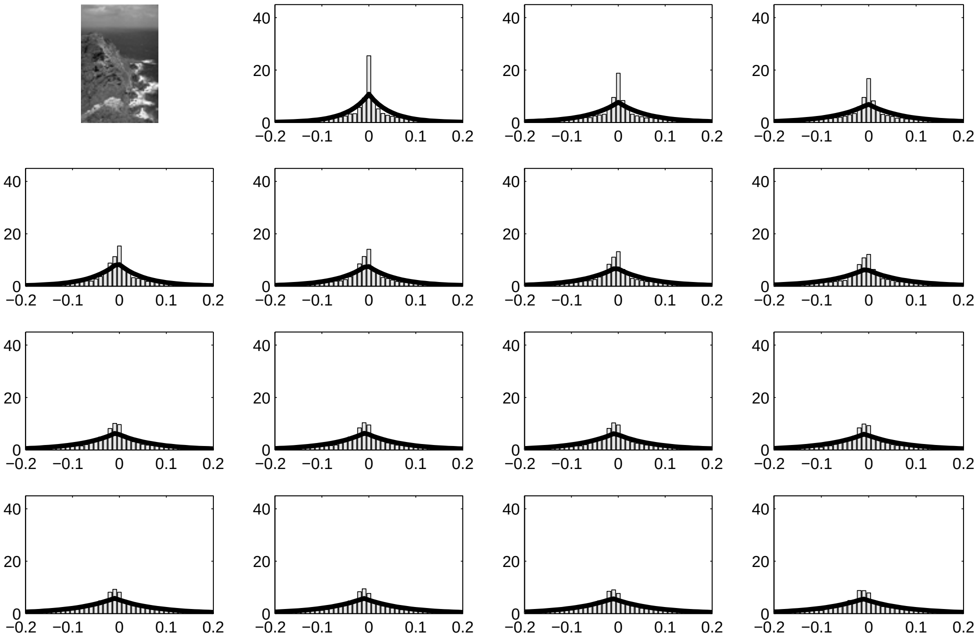


Figure 2.4: Histograms of pixel differences for different scales of pixel differences and Laplace ML estimates. With l of (2.1) increasing from left to right, with m of (2.1) increasing from top to bottom, and with the original image positioned at $(l, m) = (0, 0)$.

In particular, the s parameter of the Laplace distribution and the Laplace component of the mixture model was observed to grow according to the exponential relationship

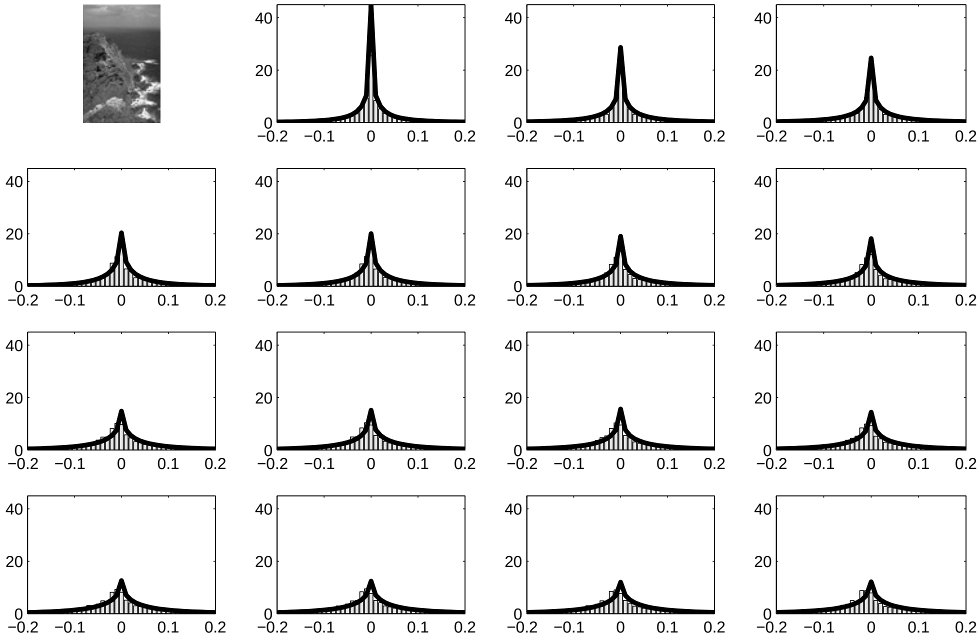


Figure 2.5: Histograms of pixel differences for different scales of pixel differences and Generalized Gaussian ML estimates. With l of (2.1) increasing from left to right, with m of (2.1) increasing from top to bottom, and with the original image positioned at $(l, m) = (0, 0)$.

$$s(l, m) = s_0 \exp\left(\frac{\gamma(|l| + |m| - 1)}{|l| + |m|}\right), \qquad |l| + |m| > 0 \tag{2.9}$$

where s_0 is the s value at the finest scale ($|l| + |m| = 1$).

The α parameter of the Generalized Gaussian distribution exhibited logarithmic growth of the form

$$\alpha(l, m) = \alpha_0 \frac{\ln(\zeta(|l| + |m|))}{\ln(\zeta)}, \qquad |l| + |m| > 0 \tag{2.10}$$

where α_0 is the α parameter value at the finest scale.

The β parameter of the Generalized Gaussian distribution grows linearly,

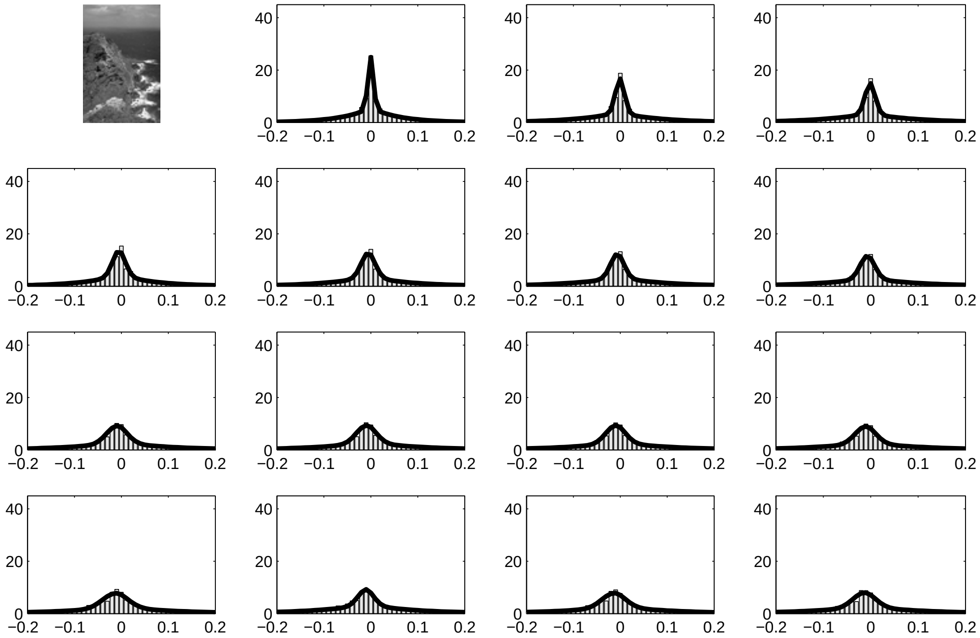


Figure 2.6: Histograms of pixel differences for different scales of pixel differences and Laplace/Gauss mixture ML estimates. With l of (2.1) increasing from left to right, with m of (2.1) increasing from top to bottom, and with the original image positioned at $(l, m) = (0, 0)$.

$$\beta(l, m) = \beta_0 + \eta(|l| + |m| - 1), \qquad |l| + |m| > 0 \tag{2.11}$$

where β_0 is the β parameter value at the finest scale.

The variance of the Gaussian component of the L-G mixture model behaves similarly to α of the Generalized Gaussian distribution.

$$\sigma^2(l, m) = \sigma_0^2 \frac{\ln(\zeta(|l| + |m|))}{\ln(\zeta)}, \qquad |l| + |m| > 0 \tag{2.12}$$

where σ_0^2 is the variance at the finest scale.

The mixing parameter, A, exhibits linear growth,

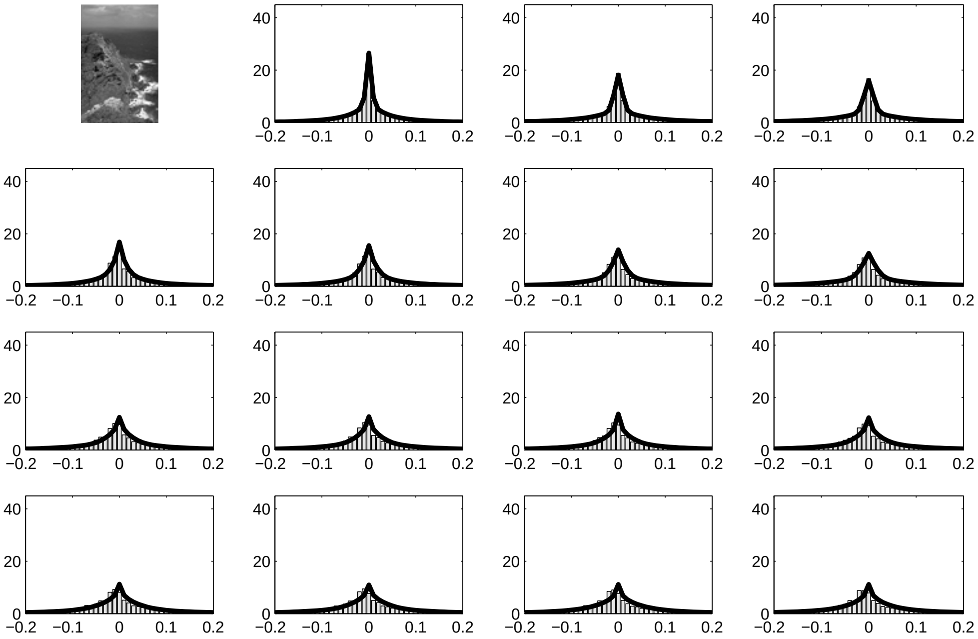


Figure 2.7: Histograms of pixel differences for different scales of pixel differences and Generalized Gauss/Gauss mixture ML estimates. With l of (2.1) increasing from left to right, with m of (2.1) increasing from top to bottom, and with the original image positioned at $(l, m) = (0, 0)$.

$$A(l, m) = A_0 + \eta(|l| + |m| - 1), \qquad |l| + |m| > 0 \tag{2.13}$$

where A_0 is the mixing parameter value at the finest scale.

The parameters of the GG-G mixture model follow similar forms to the parameters of the previous distributions. More specifically, the mixing parameter, A, and the two parameters, α and β, of the Generalized Gaussian component all displayed linear behavior.

$$A(l, m) = A_0 + \eta(|l| + |m| - 1), \qquad |l| + |m| > 0 \tag{2.14}$$

$$\alpha(l, m) = \alpha_0 + \eta(|l| + |m| - 1), \qquad |l| + |m| > 0 \tag{2.15}$$

$$\beta(l,m) = \beta_0 + \eta(|l| + |m| - 1), \qquad |l| + |m| > 0 \tag{2.16}$$

The variance of the Gaussian component of the GG-G mixture model was well approximated by exponential growth of the form,

$$\sigma^2(l,m) = \sigma_0^2 \exp\left(\frac{\gamma(|l| + |m| - 1)}{|l| + |m|}\right), \qquad |l| + |m| > 0 \tag{2.17}$$

It is informative to observe the parameter behaviors for the average case. As such, the ML parameter estimates were found for each individual image, and then these estimates were averaged across images. In order to fit the models of Equations 2.9-2.17 to the average parameter behavior, (as shown in Figure **2.8**), a least-squares criterion was employed.

For the Laplace distribution, $s_0 = 0.0457$ and the exponential growth parameter γ was found to be 0.6731. The normalized[3] average s values for each scale along with the approximated exponential fit are shown in Figure **2.8**a.

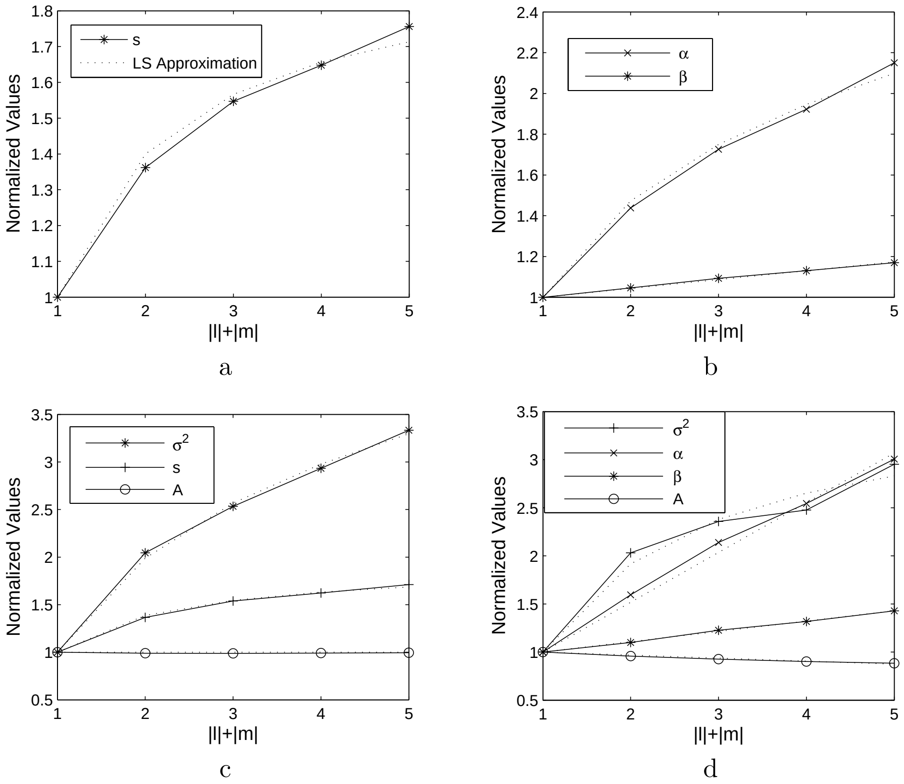
For the Generalized Gaussian distribution, $\alpha_0 = 0.0219$ and the parameter dictating its growth is given by $\zeta = 4.327$. $\beta_0 = 0.6022$ and the associated linear slope is $\eta = 0.0434$. The normalized average parameters for the Generalized Gaussian are shown with their approximations in Figure **2.8**b.

For the L-G mixture model, $s_0 = 0.0620$, and the corresponding γ parameter is 0.6534. It is interesting to note that the rate of growth of the s parameter of the Laplace component of the mixture model is very close to the rate of growth of the s component of the Laplace distribution when it is used alone. The variance of the Gaussian component is dictated by $\sigma_0^2 = 0.00075$ and $\zeta = 2.0157$. The mixing parameter, A, is seen to have an η value of 0, (i.e. constant),with $A_0 = 0.6795$. The normalized parameters for the L-G mixture model along with their least-squares fits are shown in Figure **2.8**c.

---

[3] Each parameter and its estimate are divided by the corresponding minimum parameter value for clarity in Figures **2.8**a,b, and c.

For the GG-G mixture model, $\alpha_0 = 0.0345$, and the corresponding η parameter is 0.5175. The value of β_0 was found to be 0.7187 with an η parameter value of 0.1073. The variance of the Gaussian component is dictated by $\sigma_0^2 = 0.0011$ and $\zeta = 1.3006$. The mixing parameter, A, is seen to have an η value of -0.0315,with $A_0 = 0.8177$. The normalized parameters for the GG-G mixture model along with their least-squares fits are shown in Figure **2.8**d.



a b c d

Figure 2.8: Parameter Behavior Across Scale. a: Laplace Distribution, b: Generalized Gaussian Distribution, c: L-G Mixture Distribution, and d: GG-G Mixture Distribution

The results presented above give insight into the behavior of the overall model. Several observations can be made about the average behavior of the two mixture models. The mixing parameter can be seen to describe what percentage of the image is described by the Laplacian or Generalized Gaussian component, and what percentage by the Gaussian component.

On average, for the L-G mixture model, $A \approx 0.64$ with a slope of 0 across scale, thus the Laplacian component describes approximately 64% of the image at *every* scale. The Laplace component has heavier tails at coarser scales, by an amount dictated by the relationship expressed in Equation 2.9. This reflects a loss of detail as scale becomes more coarse. Similarly, the Gaussian component has a larger variance at coarser scales, by an amount dictated by the relationship of Equation 2.12, reflecting an increase in local variance of the relatively uniform image regions.

It is interesting to note that the β parameter of the GG-G mixture model is close to 1 for $|l| + |m| = 4, 5$, and is increasing linearly. This indicates that the distribution that fits the tails best at finer scales has a higher kurtosis than a Laplace distribution, while at coarser scales, a Laplace distribution provides the best fit (bearing in mind the fact that a Generalized Gaussian distribution is identical to a Laplace distribution when $\beta = 1$). In other words, no improvement is made by the additional degree of freedom of the Generalized Gaussian distribution at a certain range of scales.

2.4 Goodness of Fit Tests

Visually observing the "quality" of fit of probabilistic distributions to pixel difference data is not sufficient to evaluate the use of each distribution as a prior. A statistical test is needed to quantitatively evaluate the goodness of fit, and to give a means of comparison between distributions. The Chi-squared test of fit [14, 19] was employed to test the goodness of fit for all distributions. Each specified distribution, $F_0(x)$, was divided into k classes, which were taken to be successive intervals in the range of x. The probability of an observation coming from each class can be easily calculated from the assumed distribution function, and is denoted by $p_{0i}, i = 1, 2, \ldots, k$. The observed frequency in each class is given by n_i, with n total observations. The Chi-squared test statistic follows, approximately, a Chi-squared distribution with $(k - c - 1)$ degrees of freedom where c is the number of estimated parameters. The Chi-squared statistic is then given by the following formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{0i})^2}{np_{0i}}. \tag{2.18}$$

The Chi-squared figures were calculated for every image in the test set and averaged values are computed for each scale of pixel difference. k was set to 75 for all calculations to facilitate comparison. The size of k is proportional to the size of the data sets and a minimum expected frequency per bin is assured. The results are shown in Figure **2.9**. A plot of the error contribution for each of the k bins of the Chi-squared test (argument of the summation of Equation 2.18) for the pixel difference data and fits shown in Figure **2.3** is shown in Figure **2.10**. The substantially smaller errors in the center region are apparent in the two mixture model plots.

Lower Chi-squared figures indicate a better fit. The Chi-squared values confirm what Figures **2.3**a, b, c, and d suggested earlier. The mixture models offer the best fit at all scales tested, even with the additional degrees of freedom over the Generalized Gaussian distribution taken into account. The test statistic follows a Chi-squared distribution with 73 degrees of freedom for the Laplace distribution, 72 for the GG, 71 for the L-G mixture, and 70 for the GG-G mixture for the specific test formulation presented above. The GG-G mixture model provides a slightly better fit than the L-G mixture model at finer scales. Tables **2.1**, **2.2**, **2.3**, and **2.4** contain the average Chi-squared figures for each distribution across the set of test images. All Chi-squared figures indicate that on average, for the

	l=0	l=1	l=2	l=3
m=0		s=0.0445 χ^2=13058	s=0.0630 χ^2=13230	s=0.0705 χ^2=12282
m=1	s=0.0469 χ^2=12019	s=0.0567 χ^2=11626	s=0.0671 χ^2=11691	s=0.0738 χ^2=11160
m=2	s=0.0670 χ^2=11991	s=0.0694 χ^2=11363	s=0.0742 χ^2=11083	s=0.0792 χ^2=10307
m=3	s=0.0758 χ^2=11178	s=0.0779 χ^2=10526	s=0.0813 χ^2=10122	s=0.0849 χ^2=9511

Table 2.1: Average Laplace Parameters

degrees of freedom used in the test, the hypothesis that the data is derived from any of the proposed distribution is rejected, even at a significance level of 0.01. This lack of statistical significance, however, does not imply that these distributions are not useful as priors, only that they are not perfect statistical descriptors of the data. However, the relative fit of each distribution is apparent in Figure **2.9**.

Several image features contribute to the large chi-squared figures. Long, high contrast edges lead to significant error in the extremities of the tail region. These occur frequently in a wide range of image types, resulting from the borders of dark shadows, skylines, or other similar region boundaries. In the example image of Figure **2.2**a, one such boundary occurs between the breaking surf and the surrounding darker water and land. This causes an unexpectedly high frequency in the bin of the Chi-squared test corresponding to pixel differences near 1, and can be seen most clearly in the errors of the less heavy-tailed Laplace and L-G mixture models shown in Figure **2.10**a and c. If we limited the test set to a specific class of images that does not contain these type of boundaries, the fits would be substantially better, but the tests would then not be representative of the quality of fit to a large portion of the images that are encountered in practice.

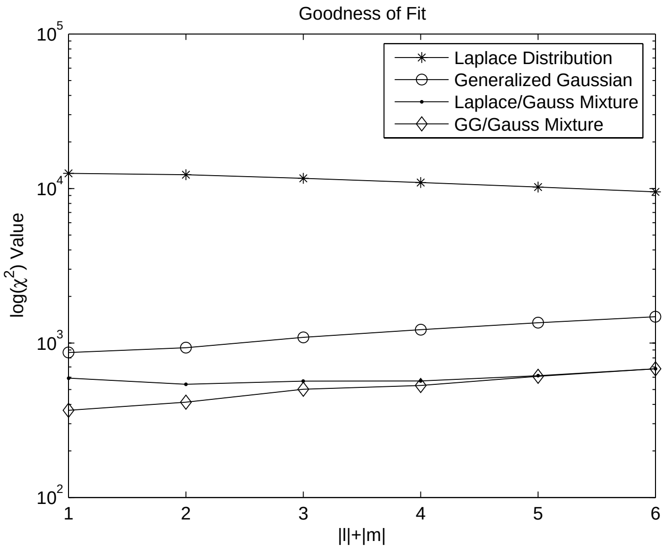


Figure 2.9: Average Chi-squared figures.

	l=0	l=1	l=2	l=3
m=0		α=0.0205 β=0.5790 χ^2=786.1	α=0.0303 β=0.6052 χ^2=888.8	α=0.0353 β=0.6315 χ^2=1005.6
m=1	α=0.0234 β=0.6254 χ^2=947.1	α=0.0286 β=0.6318 χ^2= 868.6	α=0.0341 β=0.6377 χ^2= 953.2	α=0.0390 β=0.6510 χ^2= 1053.0
m=2	α=0.0355 β=0.6529 χ^2=1037.1	α=0.0376 β=0.6660 χ^2=1066.2	α=0.0410 β=0.6758 χ^2=1125.2	α=0.0450 β=0.6897 χ^2=1232.1
m=3	α=0.0440 β=0.6973 χ^2=1324.0	α=0.0463 β=0.7099 χ^2=1478.4	α=0.0491 β=0.7186 χ^2=1473.9	α=0.0524 β=0.7296 χ^2=1478.9

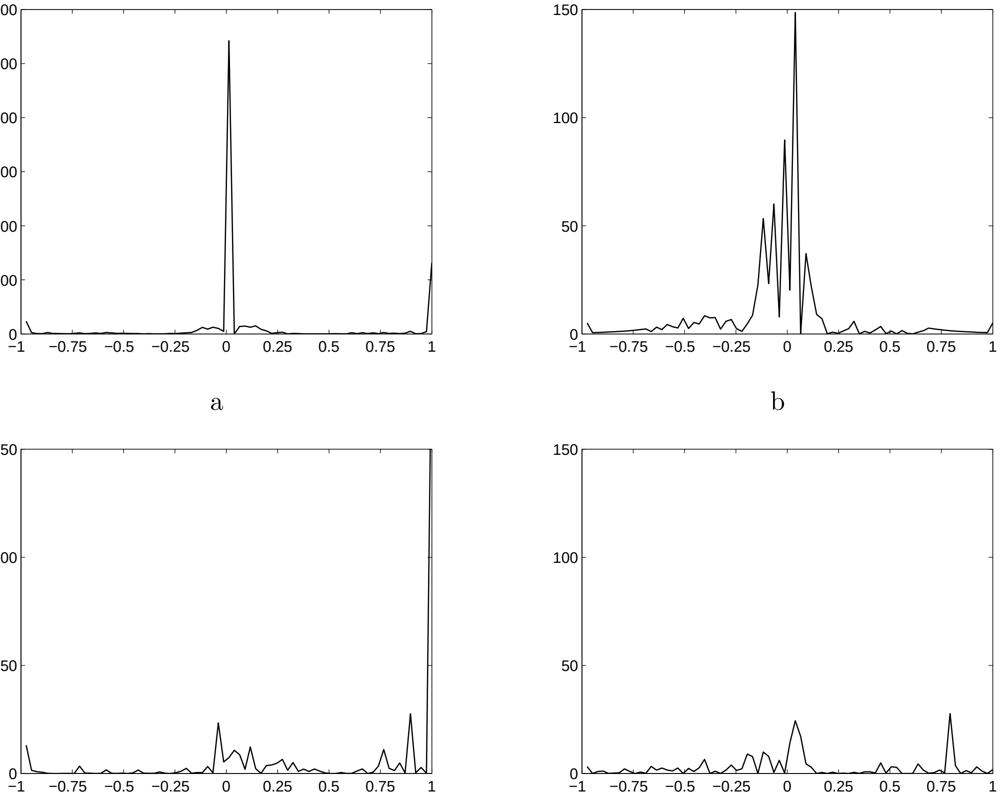
Table 2.2: Average Generalized Gaussian Parameters

	l=0	l=1	l=2	l=3
m=0		s=0.0599 σ^2=7e-4 A=0.6908 χ^2=609.74	s=0.0854 σ^2=1.6e-3 A=0.6774 χ^2=515.07	s=0.0949 σ^2=1.9e-3 A=0.6756 χ^2=496.46
m=1	s=0.0642 σ^2=8e-4 A=0.6683 χ^2=575.75	s=0.0770 σ^2=1.3e-3 A=0.6743 χ^2=508.56	s=0.0911 σ^2=1.7e-3 A=0.6718 χ^2=486.67	s=0.0987 σ^2=2.1e-3 A=0.6763 χ^2=482.75
m=2	s=0.0919 σ^2=1.7e-3 A=0.6641 χ^2=597.73	s=0.0944 σ^2=1.8e-3 A=0.6656 χ^2=548.23	s=0.0997 σ^2=2.0e-3 A=0.6734 χ^2=537.97	s=0.1050 σ^2=2.4e-3 A=0.6758 χ^2=541.86
m=3	s=0.1006 σ^2=2.2e-3 A=0.6710 χ^2=734.86	s=0.1037 σ^2=2.5e-3 A=0.6702 χ^2=683.79	s=0.1073 σ^2=2.6e-3 A=0.6765 χ^2=685.94	s=0.1110 σ^2=2.9e-3 A=0.6780 χ^2=680.6

Table 2.3: Average L-G Mixture Model Parameters

	l=0	l=1	l=2	l=3
m=0		α=0.0308 β=0.6817 σ^2=1.1e-3 A=0.8378 χ^2=339.21	α=0.0511 β=0.7439 σ^2=2.4e-3 A=0.8080 χ^2=349.82	α=0.0672 β=0.8578 σ^2=2.7e-3 A=0.7857 χ^2=417.11
m=1	α=0.0381 β=0.7557 σ^2=1e-3 A=0.7977 χ^2=394.58	α=0.0485 β=0.7822 σ^2=1.6e-3 A=0.7858 χ^2=350.44	α=0.0642 β=0.8278 σ^2=2.2e-3 A=0.7706 χ^2=358.08	α=0.0802 β=0.9072 σ^2=2.5e-3 A=0.7538 χ^2=393.99
m=2	α=0.0652 β=0.8445 σ^2=2.4e-3 A=0.7566 χ^2=540.60	α=0.0722 β=0.8752 σ^2=2.3e-3 A=0.7468 χ^2=498.84	α=0.0835 β=0.9308 σ^2=2.6e-3 A=0.7381 χ^2=518.29	α=0.0985 β=0.9974 σ^2=3.0e-3 A=0.7288 χ^2=535.63
m=3	α=0.0911 β=0.9667 σ^2=2.7e-3 A=0.7262 χ^2=731.41	α=0.0992 β=1.0031 σ^2=2.7e-3 A=0.7219 χ^2=680.13	α=0.1086 β=1.0554 σ^2=3.2e-3 A=0.7180 χ^2=682.42	α=0.1192 β=1.0869 σ^2=3.1e-3 A=0.7118 χ^2=680.11

Table 2.4: Average GG-G Mixture Model Parameters



a

b

c

d

Figure 2.10: Chi-squared error. a: Laplace Distribution, b: Generalized Gaussian Distribution, c: L-G Mixture Distribution, and d: GG-G Mixture Distribution

Chapter 3

Mixture Model Prior

The statistical study of the behavior of pixel differences discussed in Chapter 2 suggests that the mixture model of (2.7) could prove useful as a prior for image processing tasks. The density function can not be efficiently used as a prior in its current form, and must be adapted for efficient implementation. We choose to focus on the Laplace-Gauss (L-G) form of the mixture model given by

$$p_{MM}(x;\underline{\theta}) = Ap_L(x;s) + (1-A)p_G(x;\sigma^2), \tag{3.1}$$

where $\underline{\theta} = [A, s, \sigma^2]$, $A \in (0,1)$ is the mixing paramater, $p_L(x;s)$ is the zero-mean classical Laplace density function of Equation 2.2, and $p_G(x;\sigma^2)$ is the zero-mean Gaussian density function of Equation 2.8. Priors based on density functions often find use in the form of log-likelihood functions, in which the log of the joint density function is taken. These log-likelihood functions allow computationally efficient implementations in the case of densities from the exponential family. Simply taking the log of Equation 3.1 does not afford an easily implementable solution; several approximations must be made. These approximations, along with a discussion of the final version of the prior are discussed in this chapter.

3.1 Approximation to Log-Likelihood Function

The log-likelihood function for the L-G mixture model with pixel differences represented by the random variable $\mathbf{x}$, and n samples denoted by $\{x_i\}$, $1 \leq i \leq n$ is given by

$$L(\mathbf{x};\underline{\theta}) = \sum_{i=1}^{k} \log(Ap_L(x_i;s) + (1-A)p_G(x_i;\sigma^2)). \tag{3.2}$$

In order to make an approximation to this log-likelihood function, several observations can be made about the L-G mixture model, and its average parameter values. The function $p_{MM}(x;\underline{\theta})$ of Equation 3.1 can be seen to have two distinct parts: the tail region, and the center region. As a result, the approximation to the log-density function will consist of two distinct parts,

$$\log p_{MM}(x;\underline{\theta}) \approx \hat{l}(x;\underline{\theta}) = \begin{cases} \hat{l}_c(x;\underline{\theta}) & |x| < \rho(\underline{\theta}) \\ \hat{l}_t(x;\underline{\theta}) & |x| \geq \rho(\underline{\theta}) \end{cases}, \tag{3.3}$$

where $\hat{l}_t(x;\underline{\theta})$ is an approximation to the tail region, $\hat{l}_c(x;\underline{\theta})$ is an approximation to the center region, and $\rho(\underline{\theta})$ is some threshold determining when the center or tail approximation is applied.

The tail behavior of the distribution can be assumed to be approximated well by the Laplacian component of the mixture model alone. This can be seen by observing the typical values that the average variance of the Gaussian component takes on, presented in Table **2.3**. Four standard deviations of $p_G(x;\sigma^2)$, containing 99.99% of the area of the Gaussian component, includes the range of pixel difference values from ±0.11 to ±0.2 for different scales of pixel differences observed. This leaves a significant tail region which can therefore be seen to be drawn from the Laplacian component of the L-G mixture distribution. The tail approximation is then given by

$$\hat{l}_t(x;\underline{\theta}) = -\frac{|x|}{s} + \log\frac{A}{2s}. \tag{3.4}$$

However, the behavior of the mixture model near the origin is not described accurately by either component in isolation. Both the Laplacian component and the Gaussian component are necessary to accurately describe the behavior. An approximation to the region near the origin was made by formulating the log-density of the L-G mixture as follows, with $h = 1/(2\sigma^2)$ for the Gaussian component, and $f = \frac{1}{s}$ for the Laplacian component.

$$\begin{aligned}
\log p_{MM}(x;\underline{\theta}) &= \log\left[A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)\right] \\
&= \log\left[\frac{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}}\right] + \log\left[A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}\right] \\
&= \log\left[\frac{B\exp(-f|x|) + C\exp(-hx^2)}{B+C}\right] + \log\left[B+C\right], \qquad (3.5)
\end{aligned}$$

where $B = A\frac{f}{2}$ and $C = (1-A)\sqrt{h/\pi}$.

The first logarithmic term in the above equation was expanded about 1 into its Taylor series. Note that the argument of the logarithm is always greater than zero and less than or equal to one, with equality only when $x = 0$.

$$\begin{aligned}
\log p_{MM}(x;\underline{\theta}) \approx\ & \frac{B(\exp(-f|x|)-1) + C(\exp(-hx^2)-1)}{B+C} \\
& -\frac{[B(\exp(-f|x|)-1) + C(\exp(-hx^2)-1)]^2}{2(B+C)^2} \\
& +\frac{[B(\exp(-f|x|)-1) + C(\exp(-hx^2)-1)]^3}{3(B+C)^3} \\
& + \cdots + \log[B+C] \qquad (3.6)
\end{aligned}$$

Each exponential is then approximated by the first two terms of its Maclaurin series expansion.

$$\begin{aligned}
\exp(-f|x|) - 1 &\approx -f|x| \\
\exp(-hx^2) - 1 &\approx -hx^2 \\
(\exp(-f|x|) - 1)^2 &= \exp(-2f|x|) - 2\exp(-f|x|) + 1 \\
&\approx 0 \\
(\exp(-f|x|) - 1)^a(\exp(-hx^2) - 1)^b &\approx 0 \qquad \forall(a+b) > 1 \qquad (3.7)
\end{aligned}$$

Substituting the above expansion of the exponentials into the Taylor series of Equation 3.6 results in all terms of the expansion approximated by zero, except for the first. The resulting approximation is written as follows.

$$\hat{l}_c(x;\theta) = -K_1 f|x| - K_2 hx^2 + K_3, \qquad (3.8)$$

where

$$\begin{aligned}
K_1 &= \frac{B}{B+C} = \frac{A\frac{f}{2}}{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}} \\
K_2 &= \frac{C}{B+C} = \frac{(1-A)\sqrt{\frac{h}{\pi}}}{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}} \\
K_3 &= \log[B+C] = \log\left[A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}\right]. \qquad (3.9)
\end{aligned}$$

The complete approximation to the mixture model is then found by combining this center approximation with the tail approximation made by the Laplace component. The point of intersection between the center approximation and the tail approximation occurs at

$$
\begin{aligned}
\rho(\underline{\theta}) &= \frac{1}{2K_2h}\left(-f(K_1-1)+\sqrt{f^2(K_1-1)^2-4K_2h(\log(A\frac{f}{2})-K_3)}\right) \\
&= \frac{f}{2h}+\sqrt{\frac{f^2}{4h^2}-\frac{A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}}}{(1-A)h\sqrt{\frac{h}{\pi}}}\log\left(\frac{A\frac{f}{2}}{A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}}}\right)}. \qquad (3.10)
\end{aligned}
$$

Note that the approximation $\hat{l}(x;\underline{\theta})$ is not differentiable at the point $x = \rho(\underline{\theta})$. This does not affect the usage of the approximation in practical settings, as $\rho(\underline{\theta})$ does not generally occur at a discrete value, and the probability of a pixel difference being exactly equal to a continuous valued function is zero. In a practical implementation setting, the derivative of the approximation at the discontinuity can be arbitrarily defined as equal to the limit of the derivative as $x \to \rho(\underline{\theta})^+$.

A plot of the mixture model for the average parameter values at the finest scale, and its approximation are shown in Figure **3.1**.

3.1.1 Single Scale

The final approximation to the log-likelihood function at a single scale can then be expressed in terms of the log-density approximation. The approximation to the log of the L-G mixture density function can be written as follows

$$
\hat{l}(x;\underline{\theta}) = \begin{cases} -K_1f|x|-K_2hx^2+K_3 & \text{if } |x| \leq \rho(\underline{\theta}), \\ -f|x|+\log\left(A\frac{f}{2}\right) & \text{otherwise,} \end{cases} \qquad (3.11)
$$

where K_1, K_2, and K_3 are given by Equation 3.9.

The complete log-likelihood function is then found as the summation of the approximation to the log density function given in Equation 3.11 across pixel difference samples. The pixel difference samples $(x_1, \cdots, x_n)$ can be divided into two subsets $(\underline{x}_t, \underline{x}_c)$, where $\underline{x}_c$ contains all samples with a value less then $\rho(\underline{\theta})$, and $\underline{x}_t$ contains all samples with a value greater then or equal to $\rho(\underline{\theta})$. The vector $\underline{x}_t$ contains the n_t pixel difference samples that fall in the tail region of the density approximation, and the vector $\underline{x}_c$ contains the n_c pixel difference samples that fall in the center region, with $n_t + n_c = n$. The approximation to the log-likelihood function can then be written as follows

$$L(\mathbf{x}; \underline{\theta}) \approx -K_1 f \|\underline{x}_c\|_1 - K_2 h \|\underline{x}_c\|_2^2 + n_c K_3 - f \|x_t\|_1 + n_t \log\left(A\frac{f}{2}\right) \tag{3.12}$$

Note that $K_1 + K_2 = 1$, and these terms reflect the relative weights of the L_1 and L_2 terms in the center approximation region.

3.1.2 Multiscale

This single scale log-likelihood approximation is easily extended to a multiscale form by utilizing the parameter scale behavior observed in Section 2.3. Note that K_1, K_2, K_3, f, h, and ρ are functions of the parameters A, s, and σ^2 of the mixture model. The behavior of these mixture model parameters take the simple parametric forms discussed in Section 2.3. Following the notation of Chapter 2, $\underline{X}_{(l,m)}$ is the vector form of image $\mathbf{X}$ shifted horizontally by l pixels and vertically by m pixels and subtracted from the unshifted $\mathbf{X}$.The approximation to the L-G mixture log-density function can then be written as follows

$$\hat{l}(x_{(l,m)};\underline{\theta}) = \begin{cases} -K_1(l,m)f(l,m)|x_{(l,m)}| - K_2(l,m)h(l,m)x^2_{(l,m)} & \\ \quad +K_3(l,m) & \text{if } |x_{(l,m)}| <= \rho(\underline{\theta})_{(l,m)}, \\ -f(l,m)|x_{(l,m)}| + \log\left(A(l,m)\frac{f(l,m)}{2}\right) & \text{otherwise,} \end{cases} \tag{3.13}$$

with $x_{(l,m)}$ a pixel difference sample from $\underline{X}_{(l,m)}$.

Each vector $\underline{X}_{(l,m)}$ contains n samples, $(x_{(l,m)_1}, \cdots, x_{(l,m)_n})$. With each $\underline{X}_{(l,m)}$ divided into two subsets $(\underline{x}_{(l,m)_t}, \underline{x}_{(l,m)_c})$ as above, where $\underline{x}_{(l,m)_c}$ contains all samples with values less than $\rho(\underline{\theta})_{(l,m)}$, and $\underline{x}_{(l,m)_t}$ contains all samples with values greater than or equal to ρ. At the scale dictated by l and m, the vector $\underline{x}_{(l,m)_t}$ contains the $n_{(l,m)_t}$ pixel difference samples that fall in the tail region of the density approximation, and the vector $\underline{x}_{(l,m)_c}$ contains the $n_{(l,m)_c}$ pixel difference samples that fall in the center region, with $n_{(l,m)_t} + n_{(l,m)_c} = n$. The log-likelihood function for pixel differences up to P pixels apart in the horizontal and vertical direction can then be written as

$$\begin{aligned} L(\mathbf{x};\underline{\theta}) \approx & \sum_{\substack{l=-P \\ |l|+|m|>0}}^{P} \sum_{m=-P}^{P} -K_1(l,m)f(l,m)\|\underline{x}_{(l,m)_c}\|_1 - K_2(l,m)h(l,m)\|\underline{x}_{(l,m)_c}\|_2^2 \\ & +n_{(l,m)_c}K_3(l,m) - f(l,m)\|x_{(l,m)_t}\|_1 + n_{(l,m)_t}\log\left(A(l,m)\frac{f(l,m)}{2}\right) \end{aligned} \tag{3.14}$$

The total weights given to the L_1 and L_2 terms of the center approximation across scale ($-K_1(l,m)f(l,m)$ and $-K_2(l,m)h(l,m)$ respectively), for the observed average parameter values, are shown in Figure **3.2**. The L_2 coefficients have significantly larger magnitudes then the L_1 coefficients, reflecting the influence of the Gaussian component of the L-G mixture model near the origin.

3.2 Approximation Error

When approximating the log of the L-G mixture model density function, some error is introduced. In this section, we discuss the approximation error and its properties. We will divide the approximation into its center and tail components, as in Equation 3.3, and study the properties of each separately. Note that the log-density function, the tail approximation, and the center approximation are all even symmetric.

As discussed above, the tail region of the log density function is approximated by the Laplacian component of the mixture model, and is given by Equation 3.4. We will show that the tail approximation is bounded by the log-density function, and that the log-density function converges to the tail approximation for large $|x|$. We will then use these properties to derive an expression for the maximum error of the tail approximation.

First, we will show that the tail approximation, $\hat{l}_t(x;\underline{\theta})$, is bounded by the log-density function $\log p_{MM}(x;\underline{\theta})$ for all x,

$$\begin{aligned} \log p_{MM}(x;\underline{\theta}) &> \hat{l}_t(x;\underline{\theta}) \\ \log\left(A\frac{f}{2}\exp(-f|x|)+(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)\right) &> \log(A\frac{f}{2}\exp(-f|x|), \end{aligned} \tag{3.15}$$

where we have substituted Equations 3.1 and 3.4 for their respective terms on both sides of the inequality. Note that the arguments of the logarithms on both sides of the inequality of Equation 3.15 are strictly positive for all x. Using the fact that $\log y$ is a monotonically increasing function for $y > 0$, we can reduce the inequality of Equation 3.15 to an inequality of the arguments of the logarithms,

$$\begin{aligned} A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2) &> A\frac{f}{2}\exp(-f|x|) \\ \exp(-hx^2) &> 0, \end{aligned} \tag{3.16}$$

thus proving that $\hat{l}_t(x;\underline{\theta})$, is bounded by $\log p_{MM(x;\underline{\theta})}$ for all x.

Next, we show that the log-density function converges asymptotically to the tail approximation for large $|x|$. The convergence of the log-density function to the tail approximation is shown by taking the limit as $|x|$ approaches infinity of the difference of the two functions,

$$\begin{aligned} \lim_{|x|\to\infty} \log p_{MM}(x;\underline{\theta}) - \hat{l}_t(x;\underline{\theta}) &= \lim_{|x|\to\infty} \log\left(A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)\right) \\ &\quad - \log\left(A\frac{f}{2}\exp(-f|x|)\right) \\ &= \lim_{|x|\to\infty} \log\left(\frac{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2}\exp(-f|x|)}\right), \end{aligned} \tag{3.17}$$

where Equations 3.1 and 3.4 have again been substituted for their respective terms in the numerator and denominator.

The argument of the logarithm of Equation 3.17 can be shown to converge to 1 as $|x|$ approaches infinity by multiplying numerator and denominator by $\exp(f|x|)$,

$$\begin{aligned} \lim_{|x|\to\infty} \frac{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2}\exp(-f|x|)} &= \lim_{|x|\to\infty} \frac{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}\exp(f|x|-hx^2)}{A\frac{f}{2}} \\ &= \frac{A\frac{f}{2}}{A\frac{f}{2}} \\ &= 1. \end{aligned} \tag{3.18}$$

Thus, the limit of Equation 3.17 approaches 0, proving that the tail approximation converges to the log density function for large $|x|$.

Convergence in slope is shown in the following. The tail approximation is linear for $x \neq 0$, with its derivative taking the value

$$\frac{d}{dx}\hat{l}_t(x;\underline{\theta}) = -f sign(x). \tag{3.19}$$

The first derivative of the log of the L-G mixture density is given by

$$\begin{aligned} \frac{d}{dx}\log p_{MM}(x;\underline{\theta}) &= \frac{1}{p_{MM}(x;\underline{\theta})}\left(\frac{d}{dx}p_{MM}(x;\underline{\theta})\right) \\ &= -\frac{A\frac{f^2}{2} sign(x)\exp(-f|x|) + 2hx(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)} \end{aligned} \tag{3.20}$$

Observe that the derivative is strictly positive for all $x < 0$, and strictly negative for all $x > 0$. For sufficiently large values of x, the derivative of the log-density approaches the derivative of the tail approximation as seen by taking the limit of Equation 3.20 as $|x|$ goes to infinity.

$$\begin{aligned} \lim_{|x|\to\infty}\frac{d}{dx}\log p_{MM}(x;\underline{\theta}) &= \lim_{|x|\to\infty} -\frac{A\frac{f^2}{2} sign(x)\exp(-f|x|) + 2hx(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)} \\ &\to -f sign(x). \end{aligned} \tag{3.21}$$

Equation 3.21 is equal to Equation 3.19, showing that the log density function converges in slope to the tail approximation for large $|x|$.

With the fact that the tail approximation is bounded by the log-density function in mind, an expression for the absolute approximation error of $\hat{l}_t(x;\underline{\theta})$ is given as follows,

$$\begin{aligned} \epsilon_t(x;\underline{\theta}) &= \hat{l}_t(x;\underline{\theta}) - \log p_{MM}(x;\underline{\theta}) \\ &= \log\left(A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)\right) - \log A\frac{f}{2} + f|x|. \quad (3.22) \end{aligned}$$

We have already shown that the tail approximation converges to the log density function for large $|x|$, indicating that the absolute error $\epsilon_t(x;\underline{\theta})$ approaches 0 for large $|x|$. The error begins to increase as $|x|$ decreases from infinity (observe that $\epsilon_t(x;\underline{\theta})$ takes on positive values for $x < \infty$). The point at which the maximum error of the tail approximation occurs can be found by equating the first derivative of $\epsilon_t(x;\underline{\theta})$ to zero. The first derivative of the absolute error of the tail approximation of Equation 3.22 is given by

$$\begin{aligned} \frac{d}{dx}\epsilon_t(x;\underline{\theta}) &= \frac{d}{dx}\log\left(A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)\right) - \log A\frac{f}{2} + f|x| \\ &= f sign(x) - \frac{A\frac{f^2}{2} sign(x)\exp(-f|x|) + (1-A)2hx\sqrt{\frac{h}{\pi}}\exp(-hx^2)}{A\frac{f}{2}\exp(-f|x|) + (1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)}. \quad (3.23) \end{aligned}$$

The root of Equation 3.23 is $|x| = \frac{f}{2h}$. This is the value of $|x|$ which maximizes the expression given by Equation 3.22. This maximum absolute error for the tail approximation can be shown to always occur at $|x| \leq \rho(\underline{\theta})$ (shown only for positive x below)

$$\begin{aligned} \frac{f}{2h} &\leq \frac{f}{2h} + \sqrt{\frac{f^2}{4h^2} - \frac{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}}{(1-A)h\sqrt{\frac{h}{\pi}}}\log\left(\frac{A\frac{f}{2}}{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}}\right)} \\ 0 &\leq \sqrt{\frac{f^2}{4h^2} - \frac{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}}{(1-A)h\sqrt{\frac{h}{\pi}}}\log\left(\frac{A\frac{f}{2}}{A\frac{f}{2} + (1-A)\sqrt{\frac{h}{\pi}}}\right)}, \quad (3.24) \end{aligned}$$

where $\rho(\underline{\theta})$ is given by Equation 3.10.

Recall the fact that the complete approximation to the log density function consists of the tail approximation, $\hat{l}_t(x;\underline{\theta})$, only for $|x| > \rho(\underline{\theta})$ (see Equation 3.3). Equation

3.24 shows that the maximum error of the complete approximation, $\hat{l}(x;\underline{\theta})$, due to the tail approximation component ,$\hat{l}_t(x;\underline{\theta})$, will occur at $|x| = \rho(\underline{\theta})$.

The error can be seen to be derived from the influence of the Gaussian component of the mixture model, as its amplitude begins to increase near the origin. A plot of $\epsilon_t(x;\underline{\theta})$ for the average parameter values at the finest scale (taken from Table **2.3**) is shown in Figure **3.3**.

The second part of the approximation of Equation 3.3 is the center approximation, $p_c(x;\underline{\theta})$. In a similar manner, we will show that the center approximation is bounded by the log-density function, and then show that the approximation error increases as $|x|$ increases.

The approximation to the center of the log-density function, and its first two derivatives are given by

$$\hat{l}_c(x;\underline{\theta}) = -K_1 f|x| - K_2 h x^2 + K_3 \tag{3.25}$$

$$\frac{d}{dx}\hat{l}_c(x;\underline{\theta}) = -K_1 f sign(x) - 2K_2 h x \tag{3.26}$$

$$\frac{d^2}{dx^2}\hat{l}_c(x;\underline{\theta}) = -2K_2 h. \tag{3.27}$$

It is apparent that $\hat{p}_c(x;\underline{\theta})$ is a convex function, with the second derivative being negative and constant for all x.

Next, we show that the center approximation is bounded by the log-density function, with increasing error as x increases. First, note that the approximation, $\hat{l}_c(x;\underline{\theta})$ is equal to the log-density function at $x = 0$. Also, the limit of the first derivative of $\hat{l}_c(x;\underline{\theta})$ as x approaches 0^+ is equal to the same limit taken for the log-density function, shown as follows

$$\begin{aligned} \lim_{x\to 0^+} \frac{d}{dx}\hat{l}_c(x;\underline{\theta}) &= -K_1 f && (3.28) \\ \lim_{x\to 0^+} \frac{d}{dx}\log p_{MM}(x;\underline{\theta}) &= \frac{-A\frac{f^2}{2}}{A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}}} = -K_1 f && (3.29) \end{aligned}$$

The second derivatives of $\hat{l}_c(x;\underline{\theta})$ and $\log p_{MM}(x;\underline{\theta})$ as x approaches 0^+ are given by

$$\begin{aligned} \lim_{x\to 0^+} \frac{d^2}{dx^2}\hat{l}_c(x;\underline{\theta}) &= -2K_2 h && (3.30) \\ \lim_{x\to 0^+} \frac{d^2}{dx^2}\log p_{MM}(x;\underline{\theta}) &= -2K_2 h + f^2 K_1 + \frac{2}{A}K_1^2 && (3.31) \end{aligned}$$

The second derivative of the log-density function can be seen to be greater than the second derivative of the approximation in the limit as x approaches 0^+. This fact, taken in conjunction with the equality in the functions at $x = 0$ and their first derivatives in the limit, show that the approximation is bounded by the log-density function near the origin. As x increases, we can show that the first derivative of $\hat{l}_c(x;\underline{\theta})$ is less than the first derivative of $\log p_{MM}(x;\underline{\theta})$ for all $x > 0$,

$$\begin{aligned} \frac{d}{dx}\hat{l}_c(x;\underline{\theta}) &< \frac{d}{dx}\log p_{MM}(x;\underline{\theta}) \\ -\frac{A\frac{f^2}{2}sign(x)+2x(1-A)\sqrt{\frac{h}{\pi}}}{A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}}} &< -\frac{A\frac{f^2}{2}sign(x)+2x(1-A)\sqrt{\frac{h}{\pi}}}{A\frac{f}{2}\exp(-f|x|)+(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)} \\ \frac{1}{A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}}} &< \frac{1}{A\frac{f}{2}\exp(-f|x|)+(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2)} \\ A\frac{f}{2}\exp(-f|x|)+(1-A)\sqrt{\frac{h}{\pi}}\exp(-hx^2) &< A\frac{f}{2}+(1-A)\sqrt{\frac{h}{\pi}} \qquad \forall x > 0, && (3.32) \end{aligned}$$

where we have substituted the expressions from Equation 3.20 and 3.26 for their respective terms.

It follows directly that the center approximation error is strictly increasing for $|x|$ increasing, and in terms of the complete approximation, $\hat{l}(x;\underline{\theta})$, the maximum error will occur at $x = \rho(\underline{\theta})$. The error is on the order of x^2 due to the neglected terms of the Maclaurin series expansion of Equation 3.7. An expression for the absolute error of the center approximation is given as follows

$$\epsilon_c(x;\underline{\theta}) = \log p_{MM}(x;\underline{\theta}) + K1f|x| + K2hx^2 - K3. \tag{3.33}$$

A plot of $\epsilon_c(x;\underline{\theta})$ for the average parameter values at the finest scale (taken from Table **2.3**) is shown in Figure **3.3**.

It was shown above that the error is increasing for the tail approximation as $|x|$ decreases with a maximum value at $|x| < \rho(\underline{\theta})$, and is increasing for the center approximation as $|x|$ increases. Therefore, the maximum error of the complete log-density approximation, $\hat{l}(x;\underline{\theta})$, occurs at the point $|x| = \rho(\underline{\theta})$, and can be computed from Equation 3.22 or 3.33. The maximum approximation error, expressed as a percent error is given by

$$\epsilon_{max} = 100 \cdot \frac{\epsilon_t(\rho(\underline{\theta});\underline{\theta})}{\log p_{MM}(\rho(\underline{\theta});\underline{\theta})} \tag{3.34}$$

The integral absolute error of the approximation may also be expressed as a percent error, and is given by

$$\epsilon_{tot} = 100 \cdot \frac{\int_0^{\rho(\underline{\theta})} \epsilon_c(x;\underline{\theta})dx + \int_{\rho(\underline{\theta})}^1 \epsilon_t(x;\underline{\theta})dx}{\int_0^1 |\log p_{MM}(x;\underline{\theta})|dx}, \tag{3.35}$$

where we have used the fact that both the approximation error and the log density function are symmetric about the origin.

The maximum approximation error (occurring at $x = \rho(\underline{\theta})$) and the integral absolute error for the average parameter values of Table **2.3** are shown in Table **3.1**. We see

	l=0	l=1	l=2	l=3
m=0		$\epsilon_{max} = 19.34\%$ $\epsilon_{tot} = 0.062\%$	$\epsilon_{max} = 34.85\%$ $\epsilon_{tot} = 0.137\%$	$\epsilon_{max} = 41.56\%$ $\epsilon_{tot} = 0.170\%$
m=1	$\epsilon_{max} = 22.84\%$ $\epsilon_{tot} = 0.077\%$	$\epsilon_{max} = 29.85\%$ $\epsilon_{tot} = 0.044\%$	$\epsilon_{max} = 38.03\%$ $\epsilon_{tot} = 0.157\%$	$\epsilon_{max} = 46.28\%$ $\epsilon_{tot} = 0.184\%$
m=2	$\epsilon_{max} = 39.64\%$ $\epsilon_{tot} = 0.163\%$	$\epsilon_{max} = 41.57\%$ $\epsilon_{tot} = 0.171\%$	$\epsilon_{max} = 44.47\%$ $\epsilon_{tot} = 0.187\%$	$\epsilon_{max} = 55.15\%$ $\epsilon_{tot} = 0.209\%$
m=3	$\epsilon_{max} = 50.60\%$ $\epsilon_{tot} = 0.194\%$	$\epsilon_{max} = 61.73\%$ $\epsilon_{tot} = 0.209\%$	$\epsilon_{max} = 62.37\%$ $\epsilon_{tot} = 0.219\%$	$\epsilon_{max} = 76.16\%$ $\epsilon_{tot} = 0.235\%$

Table 3.1: Percent error measures for the L-G mixture log-density approximation

increasing error as the scale increases (increasing l,m). As the scale increases, the value of $\rho(\underline{\theta})$ increases, indicating that the center approximation is used in the complete approximation for larger values of $|x|$. We have observed that the center approximation error is on the order of x^2, explaining the increasing percent error measures.

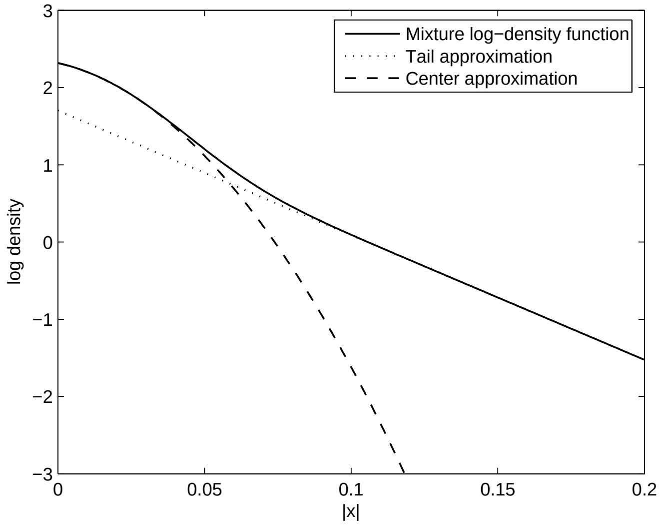


Figure 3.1: Log of L-G mixture density function, $\log p_{MM}(x;\underline{\theta})$, and approximation.

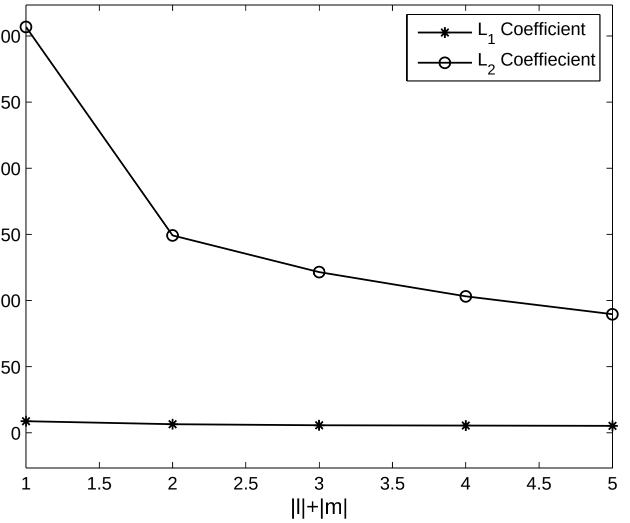


Figure 3.2: L_1 and L_2 weights for the center approximation of the log density function for the L-G mixture model across scale.

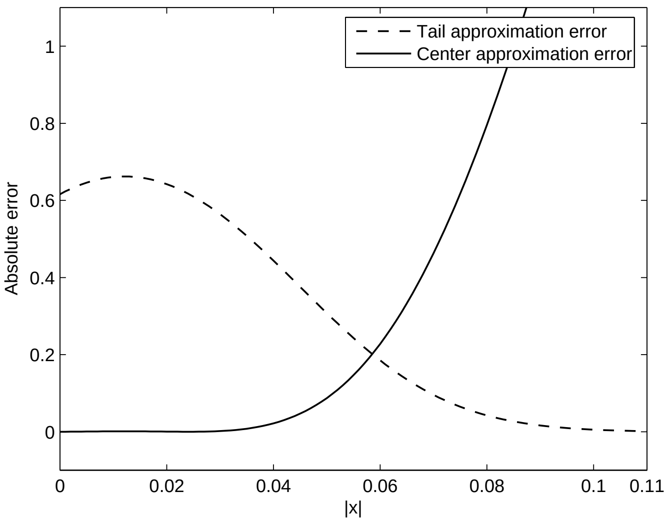


Figure 3.3: Absolute errors of the center and tail approximations to the log-density function.

Chapter 4

Image Denoising

Many image processing tasks utilize statistical assumptions about pixel differences between neighboring pixels in an image. This chapter focuses on image denoising, in which several standard methods can be derived from assumptions that pixel differences can be described by the distributions presented in Chapter 2.

Many different approaches to image denoising have been taken, with the major division being the domain in which the denoising is implemented. Wavelet domain denoising methods have been studied extensively, and a sampling of this work can be found in [9, 20, 21, 22, 23, 24]. More relevant to the work in this thesis, direct pixel domain methods have also been studied extensively in works such as [11, 25, 26, 27, 28].

The prior developed in Chapter 3 is adapted to fit into a pixel domain denoising framework. Several denoising experiments are performed, and the results of denoising with the proposed prior are compared to the methods based on different distributional assumptions.

4.1 Denoising Framework

Recalling the notation of Chapter 2, pixel differences for a given image, $\mathbf{X}$, can be represented by the vector, $\underline{X}_{(l,m)}$. The image $\mathbf{X}$ is shifted by l and m pixels in horizontal and vertical directions respectively, and subtracted from the unshifted image, with the result column-stacked into a vector as in Equation 2.1.

Image denoising is often posed in the framework of cost function minimization with regularization. If $\underline{Y}$ is the noise-corrupted version of the original image, $\underline{X}$, the denoised estimate is given as follows,

$$\underline{\hat{X}} = \arg\min_{\underline{X}} \left[\|\underline{Y} - \underline{X}\|_2^2 + \lambda \Upsilon(\underline{X}) \right], \tag{4.1}$$

where $\Upsilon(\underline{X})$ is the regularization term, and $\lambda > 0$ is the regularization weight. The regularization term is a means of incorporating prior information about the original image into the estimate. Many commonly used regularization terms can be derived from different statistical assumptions about the pixel differences in an image.

The regularization terms implemented in this chapter utilize a pairwise interaction Markov random field assumption [26, 27]. Under this assumption, the regularization term penalizes the cost function at each pixel location according to pairwise differences between pixels deemed to be neighbors. Taking the neighborhood to include only nearest neighbor pixel differences results in single scale methods, while larger neighborhood sizes result in multiscale methods.

In the following, we present several well-known regularization methods that rely on statistical assumptions about pixel differences. Two which rely only on nearest neighbor pixel differences are Tikhonov regularization, and Total Variation (TV) regularization. A method which attempts to incorporate pixel differences from more distant neighbors is Bilateral-Total Variation (B-TV). Finally we present two regularization terms based on the prior developed in Chapter 3, one that operates on nearest neighbor pixel differences, and one that operates on more distant neighbors as well.

4.1.1 Tikhonov Regularization

One frequently referenced regularization term is the Tikhonov cost function. The Tikhonov cost function can be defined as follows, where Γ is some operator, usually highpass.

$$\Upsilon_T(\underline{X}) = \|\Gamma \underline{X}\|_2^2 \tag{4.2}$$

If Γ is chosen to be the gradient operator, the Tikhonov cost function becomes

$$\Upsilon_T(\underline{X}) = \|\nabla \underline{X}\|_2^2. \tag{4.3}$$

The gradient can be approximated by taking neighboring pixel differences, giving the following

$$\Upsilon_T(\underline{X}) = \|\underline{X}_{(0,1)}\|_2^2 + \|\underline{X}_{(1,0)}\|_2^2. \tag{4.4}$$

The regularization term of Equation 4.4 can be related to a Gaussian assumption on pixel differences. If a constant term is added to the regularization term of Equation 4.1, the argument of the maximization is not affected. The regularization term of Equation 4.4, with regularization weight λ, can be rewritten as

$$\lambda \Upsilon_T(\underline{X}) = \lambda(\|\underline{X}_{(0,1)}\|_2^2 + \|\underline{X}_{(1,0)}\|_2^2) + \log C,$$

where C is a constant.

If Equation 4.5 is exponentiated, it can then be equated to a joint density of an independent, identically distributed (IID) random variable $\mathbf{x}$, with each sample $(x_1, \cdots, x_n)$ drawn from a zero-mean Gaussian distribution.

$$\begin{aligned}
\exp(\lambda \Upsilon_T) &= C \exp\left(\lambda(\|\underline{X}_{(0,1)}\|_2^2 + \|\underline{X}_{(1,0)}\|_2^2)\right) \\
&= C \exp(\lambda \|\mathbf{x}\|_2^2) \\
&= \prod_{i=1}^{n} p_G(x_i; \sigma^2), \qquad (4.5)
\end{aligned}$$

where the above equation implies $C = (2\pi\sigma^2)^{-n/2}$ and $\sigma^2 = -1/2\lambda$. The density function $p_G(x_i; \sigma^2)$ is a zero-mean Gaussian density, with variance σ^2, given by Equation 2.8. Note that all elements of $\underline{X}_{(0,1)}$ and $\underline{X}_{(1,0)}$ are assumed to be samples of the random variable $\mathbf{x}$. The Tikhonov regularization term of Equation 4.4 can then be seen to be the log of the joint density of the samples of $\mathbf{x}$.

4.1.2 Total Variation Regularization

Another well recognized method utilizes the Total Variation (TV) regularization term [11]. TV can be approximated by the computationally efficient regularization term [29]

$$\Upsilon_{TV}(\underline{X}) = \|\nabla \underline{X}\|_1, \qquad (4.6)$$

where ∇ is the gradient operator. If the gradient is again defined in terms of pixel differences,

$$\Upsilon_{TV}(\underline{X}) = \|\underline{X}_{(0,1)}\|_1 + \|\underline{X}_{(1,0)}\|_1. \qquad (4.7)$$

In a similar fashion this regularization term can be related to an assumption of a Laplace distribution on the pixel differences. Equation 4.7 multiplied by the regularization weight, λ, and exponentiated, can then be equated to a joint density of an IID random variable $\mathbf{x}$, with each sample drawn from a zero-mean Laplacian distribution.

$$\begin{aligned}\exp(\lambda \Upsilon_{TV}) &= C \exp\left(\lambda(\|\underline{X}_{(0,1)}\|_1 + \|\underline{X}_{(1,0)}\|_1)\right) \\ &= C \exp(\lambda \|\mathbf{x}\|_1) \\ &= \prod_{i=1}^{n} p_L(x_i; s), \qquad (4.8)\end{aligned}$$

where C is again a constant term that does not affect the argument of the maximization of the cost function of Equation 4.1. The above Equation implies $C = 2s^{-n}$ and $s = -1/\lambda$. The density function $p_L(\mathbf{x}; s)$ is a zero-mean Laplacian density, with scale parameter s, given by Equation 2.2. All elements of $\underline{X}_{(0,1)}$ and $\underline{X}_{(1,0)}$ are again assumed to be samples of the random variable $\mathbf{x}$. The TV regularization term of Equation 4.7 can then be seen to be the log of the joint density of the samples of $\mathbf{x}$.

4.1.3 Bilateral Total Variation Regularization

The total variation criterion has been extended to a regularizer called Bilateral-TV [13]. This regularizer combines ideas from total variation and bilateral filtering [28, 30]. This regularization term takes the form

$$\Upsilon_{BTV}(\underline{X}) = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|l|+|m|} \|\underline{X}_{(l,m)}\|_1 \qquad (4.9)$$

with α being a scalar weight, $0 < \alpha < 1$.

This regularization term can be derived from the same Laplace assumption made for the TV regularizer with one change. Namely, the $\alpha^{|l|+|m|}$ term is introduced to model parameter changes of the Laplace distribution across scale.

$$\begin{aligned}
\exp(\lambda\Upsilon_{BTV}) &= C\exp\left(\sum_{l=-P}^{P}\sum_{m=-P}^{P}\lambda\alpha^{|l|+|m|}\|\underline{X}_{(l,m)}\|_1\right) \\
&= C\prod_{l=-P}^{P}\prod_{m=-P}^{P}\exp(\lambda\alpha(l,m)\|\mathbf{x}_{(l,m)}\|_1) \\
&= \prod_{l=-P}^{P}\prod_{m=-P}^{P}\prod_{i=1}^{n}p_L(x_{(l,m)_i};s(l,m)), \qquad (4.10)
\end{aligned}$$

where the n samples of $\underline{X}_{(l,m)}$, $(x_{(l,m)_1},\cdots,x_{(l,m)_n})$, are drawn from the random variable $\mathbf{x}_{(l,m)}$. The density function $p_L(x_i;s(l,m))$ is a zero-mean Laplacian density, with parameter s a function of l and m. The above Equation implies $C=\prod_{l=-P}^{P}\prod_{m=-P}^{P}(2s(l,m))^{-n}$ and $s(l,m) = -1/\lambda\alpha^{(|l|+|m|)}$. The introduction of the $\alpha^{(|l|+|m|)}$ term has the effect of assuming the shape of the decay of the s parameter of the Laplace distribution as scale increases ($|l|+|m|$ increasing). The B-TV regularization term of Equation 4.9 can then be seen to be the log of the joint density of the samples of $\mathbf{x}_{(l,m)}$ for $|l|+|m|\leq 2P$.

4.1.4 Mixture Model Regularization

The statistical study of the behavior of pixel differences discussed in Chapter 2 suggests a different regularization term. The proposed mixture model offers a better fit to real image data, and should therefore offer better denoising performance. The L-G mixture model of Equation 3.1 offers a means to derive a computationally efficient regularization term. The regularizer we are looking for should maximize the log-likelihood function for the Laplace-Gauss (L-G) mixture model of Equation 3.1. An approximation to the log-likelihood function was developed in Chapter 3, and presents a means to implement a computationally efficient form of the L-G mixture model assumption.

The approximation to the log-likelihood function for a single derivative scale then takes the form of Equation 3.12. The constant terms that do not affect the cost function maximization can be removed. The vector $\underline{X}_{(l,m)}$ can be divided into two subsets $(\underline{x}_{(l,m)_t}, \underline{x}_{(l,m)_c})$, where $\underline{x}_{(l,m)_c}$ contains all samples with a value less than $\rho(\underline{\theta})$ (given by Equation 3.10), and $\underline{x}_{(l,m)_t}$ contains all samples with a value greater than or equal to $\rho(\underline{\theta})$. The L-G mixture model regularizer for a single scale is then given by

$$\begin{aligned} \Upsilon_{MM_{single}}(\underline{X}) &= K_1 f(\|\underline{x}_{(1,0)_c}\|_1 + \|\underline{x}_{(0,1)_c}\|_1) + K_2 h(\|\underline{x}_{(1,0)_c}\|_2^2 + \|\underline{x}_{(0,1)_c}\|_2^2) \\ &\quad + f(\|\underline{x}_{(1,0)_t}\|_1 + \|\underline{x}_{(0,1)_t}\|_1), \end{aligned} \tag{4.11}$$

where K_1 and K_2 are given by Equation 3.9, $f = 1/s$, and $h = 1/(2\sigma^2)$. Note that $K_1 + K_2 = 1$, and these terms reflect the relative weights of the L_1 and L_2 terms.

This single scale mixture model regularizer is easily extended to a multiscale form by utilizing the parameter scale behavior discussed in Chapter 2. Note that K_1 and K_2 are functions of A, s, and σ^2 of the mixture model, and the behavior of these mixture model parameters take simple parametric forms. The multiscale mixture model regularizer can then be written in the following form

$$\begin{aligned} \Upsilon_{MM_{multi}}(\underline{X}) &= \sum_{l=-P}^{P} \sum_{m=-P}^{P} K_1(l,m) f(l,m) \|\underline{x}_{(l,m)_c}\|_1 + K_2(l,m) h(l,m) \|\underline{x}_{(l,m)_c}\|_2^2 \\ &\quad + f(l,m) \|\underline{x}_{(l,m)_t}\|_1, \end{aligned} \tag{4.12}$$

recalling that ρ is also a function of (l, m).

The regularization terms of equations 4.11 and 4.12 represent global cost functions, where the specific weights are chosen spatially by the values of the pixel differences at each location. If the pixel difference value is above the threshold (i.e. edge region) its regularization is weighted by the L_1 norm, and if below (i.e. smooth region) its regularization is weighted by the mixture of L_1 and L_2 terms.

The mixture of L_1 and L_2 terms makes intuitive sense. L_1 regularization alone tends to lead to piecewise constant images, and L_2 regularization alone leads to excessively blurry images. The mixture of L_1 and L_2 can reasonably be expected to perform better, with a correct choice of the mixture. The statistical analysis of Chapter 2 is the mechanism for making a reasonable choice of how to form the mixture.

We also note that the other regularization methods presented closely correspond to specific parameter choices of the multiscale L-G mixture regularization term. If the neighborhood size, P, is set to one, the mixing parameter A set to zero, and appropriate shape parameters chosen to ensure only nearest neighbor pixel differences (not diagonal neighbors), the Tikhonov regularization of Equation 4.4 is achieved. This is seen most simply from the distribution from which the L-G mixture regularization is derived. Equation 3.1 simply becomes a Gaussian distribution when $A = 0$, which corresponds to the underlying assumption for Tikhonov regularization. In a similar fashion, with $A = 1$, and a neighborhood size of $P = 1$, a correspondence can be drawn to TV regularization. Likewise, with $A = 1$, a larger neighborhood size, and appropriate shape parameters, a close match to B-TV can be found. The match to B-TV is inexact due to the different shape of parameter decay with scale that is assumed for the two methods, but can be very close for small neighborhood sizes.

4.2 Denoising Experiments

Five methods were implemented using the regularization terms discussed above: Tikhonov, TV, Bilateral-TV, single scale LG-Mix, and multiscale LG-Mix. Each regularization term was used in the denoising framework presented in equation 4.1. This minimization was carried out by means of iterative steepest descent.

$$\underline{\hat{X}}^{(n+1)} = \underline{\hat{X}}^{(n)} - \beta \left\{ 2(\underline{\hat{X}}^{(n)} - \underline{Y}) + \lambda \left. \frac{\partial}{\partial \underline{X}} \right|_{\underline{\hat{X}}^{(n)}} \Upsilon(\underline{X}) \right\} \tag{4.13}$$

Recalling the notation of Equation 2.1, where S_x^l and S_y^m are matrices that shift image $\mathbf{X}$ by l and m pixels in horizontal and vertical directions respectively, the derivatives of the first three regularization terms can be written as follows

$$\left. \frac{\partial}{\partial \underline{X}} \right|_{\underline{\hat{X}}^{(n)}} \Upsilon_T = 2(I - S_x^{-1})(\underline{\hat{X}}^{(n)} - S_x^1 \underline{\hat{X}}^{(n)}) + 2(I - S_y^{-1})(\underline{\hat{X}}^{(n)} - S_y^1 \underline{\hat{X}}^{(n)}) \tag{4.14}$$

$$\left. \frac{\partial}{\partial \underline{X}} \right|_{\underline{\hat{X}}^{(n)}} \Upsilon_{TV} = (I - S_x^{-1}) sign(\underline{\hat{X}}^{(n)} - S_x^1 \underline{\hat{X}}^{(n)}) + (I - S_y^{-1}) sign(\underline{\hat{X}}^{(n)} - S_y^1 \underline{\hat{X}}^{(n)}) \tag{4.15}$$

$$\left. \frac{\partial}{\partial \underline{X}} \right|_{\underline{\hat{X}}^{(n)}} \Upsilon_{BTV} = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|l|+|m|} (I - S_y^{-m} S_x^{-l}) sign(\underline{\hat{X}}^{(n)} - S_x^l S_y^m \underline{\hat{X}}^{(n)}). \tag{4.16}$$

The first derivative of the multiscale L-G mixture regularization term takes the following form

$$\left. \frac{\partial}{\partial \underline{X}} \right|_{\underline{\hat{X}}^{(n)}} \Upsilon_{MM_{multi}} = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \begin{cases} (I - S_y^{-m} S_x^{-l})(K_1(l, m,) f(l, m) sign(\underline{\hat{X}}^{(n)} - S_x^l S_y^m \underline{\hat{X}}^{(n)}) \\ +2K_2(l, m) h(l, m)(\underline{\hat{X}}^{(n)} - S_x^l S_y^m \underline{\hat{X}}^{(n)})) \\ \qquad\qquad\qquad\qquad \text{if } |\underline{\hat{X}}^{(n)}(i)| \leq \rho, \\ f(l, m)(I - S_y^{-m} S_x^{-l}) sign(\underline{\hat{X}}^{(n)} - S_x^l S_y^m \underline{\hat{X}}^{(n)}) \\ \qquad\qquad\qquad\qquad \text{otherwise,} \end{cases} \tag{4.17}$$

where $\underline{\hat{X}}^{(n)}(i)$ denotes the ith element of the vector $\underline{\hat{X}}^{(n)}$.

The first derivative of the single scale L-G mixture regularization term takes the same form as Equation 4.17, with (l, m) of the summation only equal to $(1, 0)$ and $(0, 1)$.

We first perform denoising experiments on a simple simulation image. The image shown in Figure **4.1**a was corrupted by additive white Gaussian noise (AWGN) with standard deviation $\sigma = 10$, with the noisy image shown in Figure **4.1**b. The noisy version was then denoised using the different regularization terms to illustrate their effects [1]. Note the gradation of the background of the test image. The denoising results are shown in Figure **4.2**. Tikhonov regularization can be seen to blur the image excessively, thus removing edges. TV and B-TV regularization retain edges, but remove the gradation of the background. Both single scale and multiscale L-G mixture regularization retain the sharp edges in the image as well as recover the gradation of the background. The advantage of the multiscale methods can be seen by observing the pixel in the center of each letter. The light background pixels are surrounded by dark foreground pixels, and are incorporated into the foreground by the single scale methods, with the effect most noticeable in the images denoised by TV and Tikhnonov regularization. However, both multiscale [2] methods are able to retain edges and correctly interpret the two light pixels surrounded by dark. This example serves to illustrate the conceptual advantages of the L-G mixture regularization terms. The mixture of L_1 and L_2 terms allows the L-G mixture regularization to both recover edges and remove noise from the smooth regions simultaneously.

Further experiments are performed next, in which we investigate the various regularization terms' relative abilities to reconstruct an image from a noise corrupted version of the image. Two major criteria can be employed to evaluate these abilities: quantitative and qualitative. Quantitative measures such as PSNR give an indication of signal recovery ability, but are not necessarily good indicators of visual quality.

We first perform a quantitative experiment, in which we denoise an image when the ground truth, or original image, is known. The *Barbara* image, shown in Figure **4.3**

---

[1] Parameters were chosen to visually give the best denoising performance

[2] Neighborhood size, P, equals 2 for all experiments

is corrupted by additive white Gaussian noise (AWGN) with the noise standard deviation, σ ranging from 5 to 15. The parameters of each regularization method are optimized to give the maximum PSNR, taking a fixed number of iterations and step size. In the case of Tikhonov and TV regularization, the only parameter to optimize is the regularization weight , λ, of Equation 4.1. B-TV regularization has the additional scale term, α. The single scale L-G mixture regularization has the terms, s, σ^2, and A as well as the regularization weight, λ. The multiscale L-G mixture regularization has the same terms of the single scale version, along with the shape parameters, γ (2.9), ζ (2.12), and η (2.13). Monte Carlo simulations were performed, with 30 realizations at each noise level. The results of these experiments are shown in Table **4.1**. The L-G mixture model regularization method resulted in the highest PSNR for every noise level tested. The multiscale methods (B-TV and multiscale L-G mixture) offer only a slight improvement in PSNR over their single scale counterparts (TV and single scale L-G mixture). It is not surprising that the L-G mixture regularization gives the best PSNR due to the fact that the other denoising methods correspond to particular parameter choices of the mixture model, as discussed in Section 4.1.4. The denoised images from one realization of this experiment with WGN of standard deviation $\sigma = 10$ are shown in Figure **4.5**, and an enlarged region in Figure **4.6**. It is apparent from these figures that even though the PSNR is only marginally higher for the multiscale L-G mixture regularizer, the visual quality has changed. The forehead region of the image has been returned more closely to the original texture by the multiscale L-G mixture regularization than by any of the other methods.

As discussed above, B-TV and the L-G mixture regularization terms have several parameters that must be adjusted to give the results shown in Table **4.1**. It is not practical in most frameworks to allow this many degrees of freedom, as the complexity of any opti-

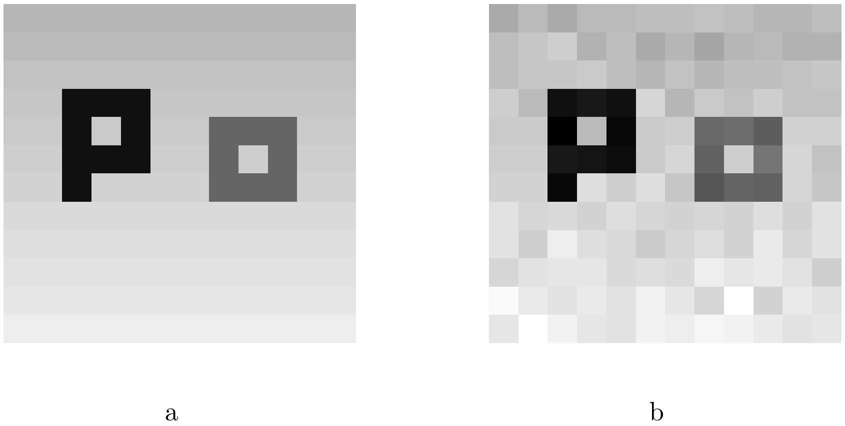
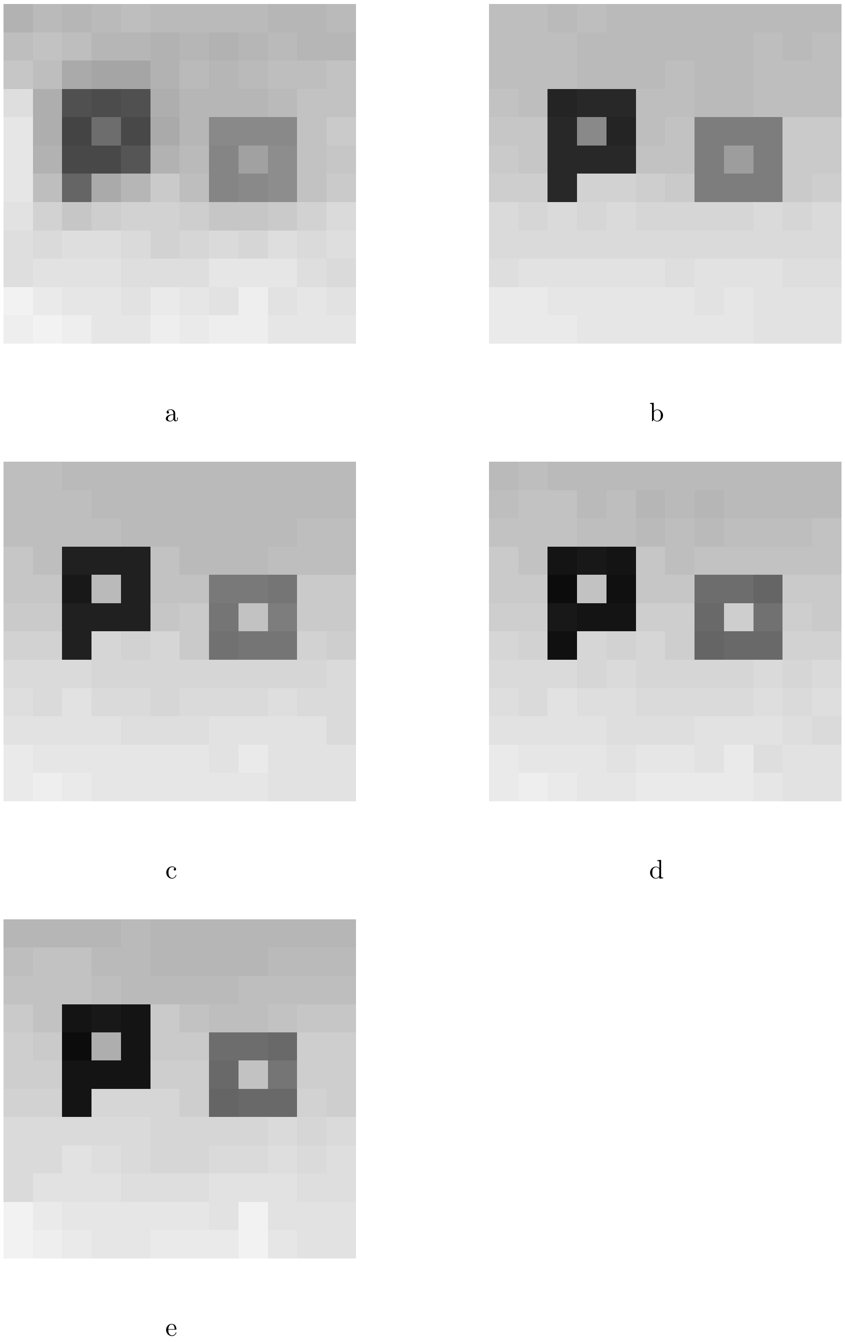

a
b

Figure 4.1: Simulated image used for denoising a: Original, b:Noisy($\sigma = 10$)

	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
Tikhonov	34.6079	29.7668	27.4194
TV	35.8449	31.0770	28.5476
B-TV	35.8519	31.0839	28.5532
L-G Mixture (multiscale)	36.0275	31.2714	28.6684
L-G Mixture (single scale)	36.0248	31.2206	28.6522

Table 4.1: Average PSNR from denoising experiment on *Barbara* image using 30 Monte Carlo runs.



a

b

c

d

e

Figure 4.2: Simulation results using different regularization terms. a:Tikhonov(MSE=566.1023), b:TV(MSE=89.1713), c:Bilateral-TV(MSE=63.6256), d:Multiscale L-G mixture(MSE=15.9847), e:Single scale L-G mixture(MSE=25.5127)

	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$
B-TV ($\alpha = 0.7$)	35.6678	30.8989	28.3260
Single Scale L-G Mixture (avg. dist. parameters)	36.0043	31.2022	28.6036
Multiscale L-G Mixture (avg. shape and dist. params)	35.7990	31.1230	28.5222
Multiscale L-G Mixture (avg. shape parameters)	36.0168	31.1842	28.5319

Table 4.2: Average PSNR from denoising experiment on *Barbara* image using 30 Monte Carlo runs, with degrees of freedom removed.

mization process increases significantly. To take this fact into account we also performed the same denoising experiment as above with several degrees of freedom removed. For B-TV regularization we fix the scale term α to a standard value of 0.7, and the neighborhood size to $P = 2$, and allow only the regularization weight λ to vary. For the single scale L-G mixture regularization, we fix the distribution parameters s, σ^2, and A to the average values presented in Table **2.3**, and allow only the regularization weight to vary. In a similar fashion we perform the denoising experiments with the shape parameters of the multiscale L-G mixture regularization fixed, allowing the distribution parameters and regularization weight to vary, and then finally for the case where all parameters except for the regularization weight are fixed to average values. The results of these experiments are shown in Table **4.2**. The multiscale methods can be seen to be quite sensitive to the shape parameters. In fact, TV regularization, is capable of giving a higher PSNR than all of the multiscale methods when fine tuning of the spatial parameter decay is not allowed. However, the single scale L-G mixture method performs quite well even with an average parameter set, outperforming all methods with full flexibility, except for the multiscale L-G mixture method.

As apparent from Figure **4.5**, maximizing PSNR does not necessarily correspond to the best visual quality. The added Gaussian noise is still visible in the denoised images. A second experiment was performed, in which no knowledge of the original image was required. A corrupted image was denoised using each of the regularization terms presented above. Only one degree of freedom was allowed, the regularization weight λ, with all other parameter values fixed to the average values. For each image and method, the regularization weight was gradually increased, until noise was no longer visible. In order to give a clear view of the effects of each regularization method, the residuals are examined. The residuals are obtained by subtracting the denoised image from the noisy version and taking the absolute value. The residuals from a perfect denoising method would consist solely of the noise that was added to the original image. The appearance of image structure and content in the residuals imply that some of the image was removed in the denoising process. If the residual values in smooth regions are larger than the added noise in those smooth regions, image texture has been removed.

The experiment described above was performed on the *Barbara* image (shown in Figure **4.3**) corrupted by AWGN noise with $\sigma = 10$. We are now denoising the same image as in the first experiment, with a qualitative criterion determining the regularization levels used. The denoising results are displayed in Figure **4.7**. Optimizing by visual criteria leads to a similar relationship between quantitative reconstruction error as was shown in the first experiment. The two L-G mixture regularization terms give the maximum PSNR, with the single scale version performing better than the average parameter multiscale form. A better understanding of the effects of the regularization terms is seen if the residuals are examined. The residuals for the TV, B-TV, and both L-G mixture methods are displayed in Figure **4.8**. The residuals are scaled to use the full dynamic range, with each residual using the same scaling. The noise that was added to the original image is also displayed with the same scaling to give a reference. The L-G mixture methods residuals resemble the noise more closely than the two L_1 methods, with the single scale L-G mixture residuals resembling the noise much more closely than the multiscale L-G mixture. Both TV and B-TV have significantly darker patterns present in their residuals, indicating that more image content was removed.

The experiment described above was repeated on the *Lena* image (shown in Figure **4.4**) corrupted by white Gaussian noise with $\sigma = 5$. The denoising results are displayed in Figure **4.9**. The two L-G mixture regularization terms again give the maximum PSNR, with the single scale version performing better than the average parameter multiscale form. The residuals for the TV, B-TV, and both L-G mixture methods are displayed in Figure **4.10**. The residuals are scaled as before to use the full dynamic range, with each residual using the same scaling. The noise that was added to the original image is also displayed with the same scaling to give a reference. The L-G mixture methods residuals resemble the noise more closely than the two L_1 methods. More image content is visible in the multiscale mixture residual, indicating that less noise was removed from the edge regions. Both TV and B-TV residuals have significantly darker patterns when compared to the reference noise image, indicating that more image texture was removed along with the noise.

A third experiment was performed in which a color image corrupted by unknown noise was restored. The *JFK* image of Figure **4.11** is corrupted by film grain noise and scanning process noise. The image was decomposed into YCrCb channels, with each denoised separately. In a similar manner to the experiment above, the regularization level was increased until noise was no longer apparent. The results of this denoising can be seen in Figure **4.11**. A detail region is shown in Figure **4.12** to further illustrate the effects of the different regularization terms. The "piecewise constant" effect of L_1 regularization can be seen in the overly smooth cheek region, while, the L_1/L_2 combination of the $L - G$ mixture regularization retains a more realistic texture. The inverted residuals for several of the denoising methods can be seen in Figure **4.13**. The larger residuals (darker grayscale values) of the two L_1 methods are apparent. Recalling the methodology of this experiment (increasing regularization until noise was not visible), the brighter values indicate that more content of the image was removed along with the noise.

The L-G mixture regularization terms consistently give the best denoising performance of the methods tested. This holds for both types of experiments: maximizing PSNR, and qualitative noise removal. The full flexibility of adjusting all underlying parameters of the L-G mixture distribution is not required for the regularizer to provide an improvement.

Figure 4.3: *Barbara* Image

Figure 4.4: *Lena* image

Figure 4.5: Denoising results using different regularization terms. From left to right and top to bottom: Noisy image, Tikhonov(MSE=68.6237), TV(MSE=50.8512), B-TV(MSE=50.7703), Multiscale L-G mixture(MSE=49.1996), Single scale L-G Mixture(MSE=49.2032)

Figure 4.6: Enlarged region of denoised image. From left to right and top to bottom: Original image, Tikhonov, TV, B-TV Multiscale L-G mixture, single scale L-G mixture

Figure 4.7: Denoised *Barbara* image, AWGN $\sigma = 10$. From left to right and top to bottom: Noisy Image, Tikhonov(PSNR-25.0190), TV(PSNR-29.8692), B-TV(PSNR-30.0962) , Multiscale L-G mixture(PSNR-30.6981), Single scale L-G mixture(PSNR-30.8475).

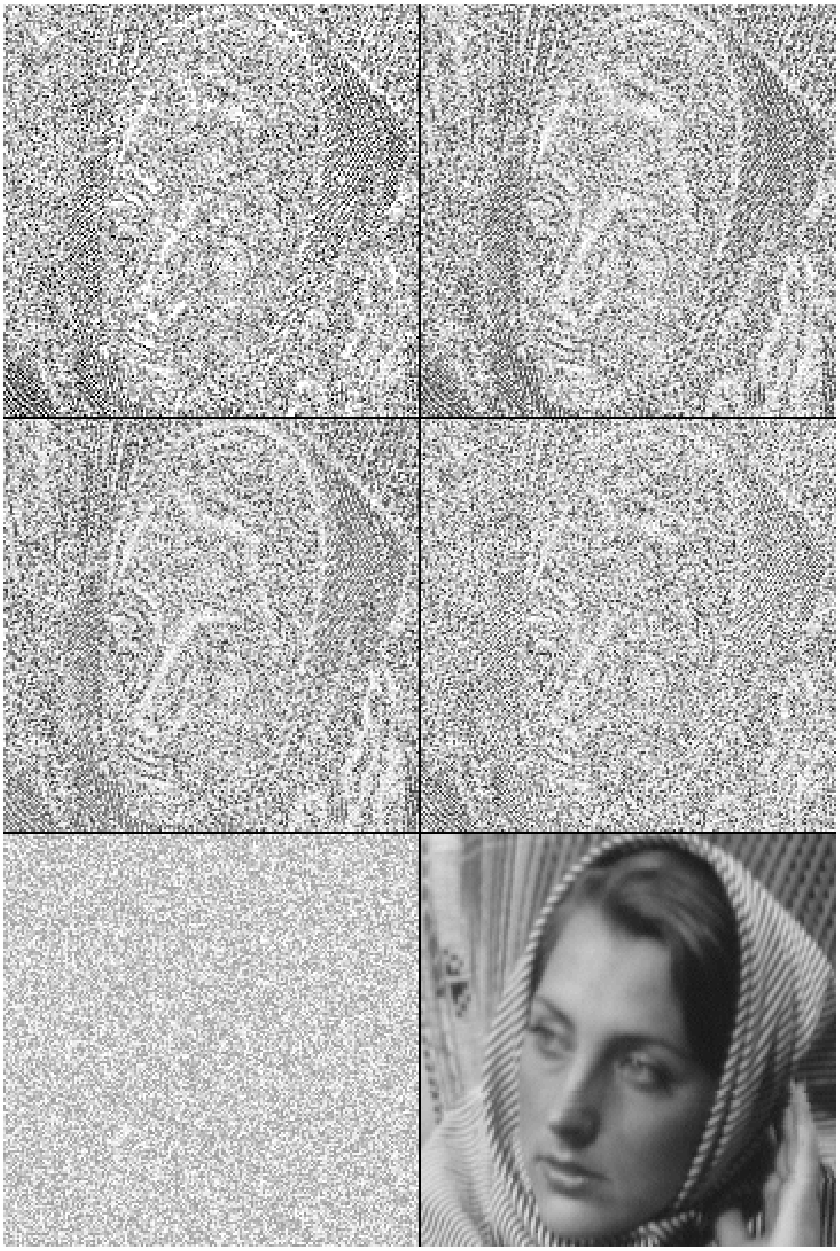
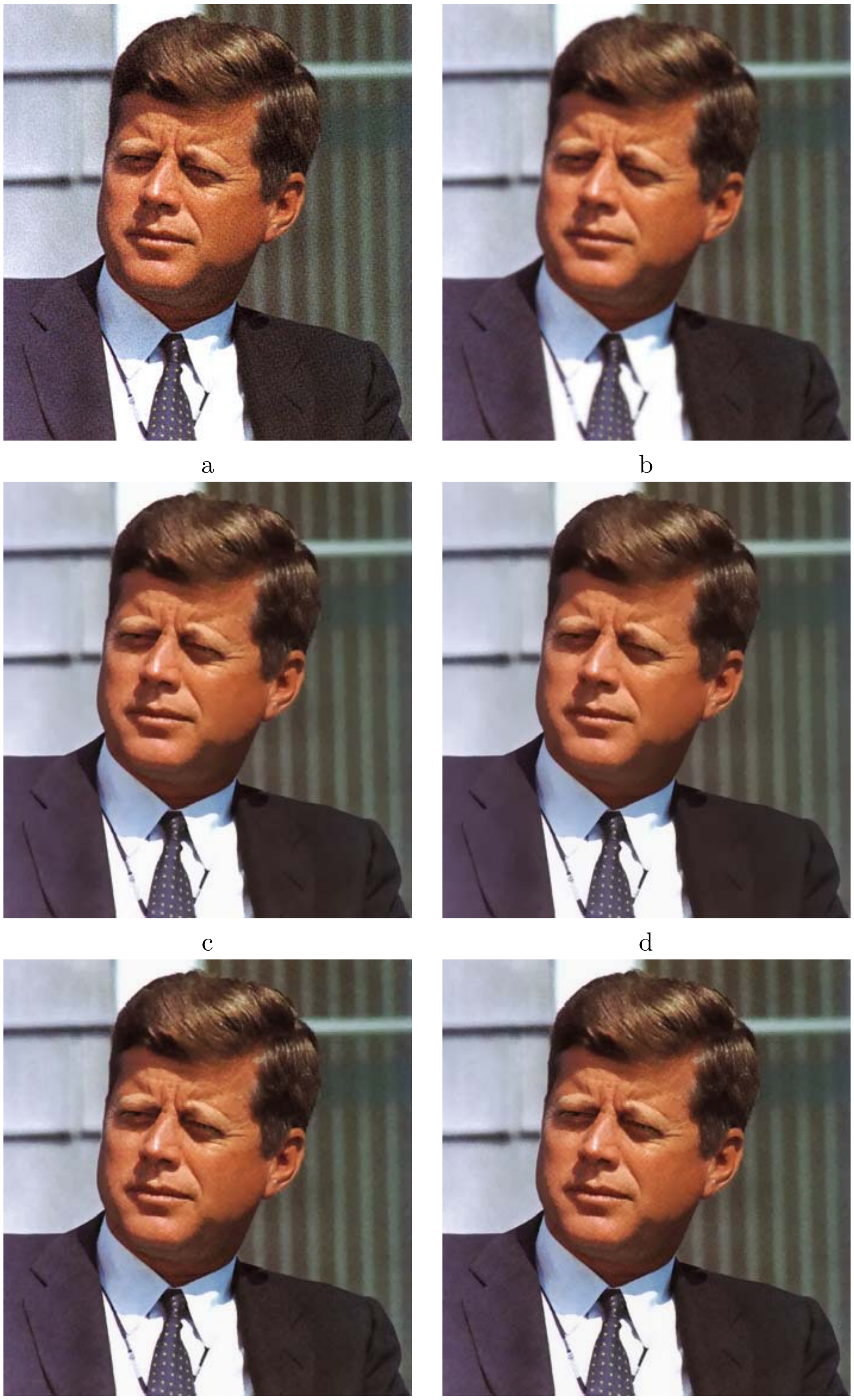


Figure 4.8: Denoising experiment results from *Barbara* image. From left to right and top to bottom: TV residual, B-TV residual, Multiscale L-G mixture residual, Single scale L-G mixture residual, added noise, original image.

Figure 4.9: Denoised *Lena* image, AWGN $\sigma = 5$. From left to right and top to bottom: Noisy Image, Tikhonov(PSNR-33.2206), TV(PSNR-36.6980), B-TV(PSNR-36.7120) , Multiscale L-G mixture(PSNR-37.1114), Single scale L-G mixture(PSNR-37.3630).

Figure 4.10: Denoising experiment results from *Lena* image. From left to right and top to bottom: TV residual, B-TV residual, Multiscale L-G mixture residual, Single scale L-G mixture residual, added noise, original image.



a b c d e f

Figure 4.11: JFK color denoising experiment. a: Original image, b: Tikhonov regularization, c: TV regularization, d: B-TV regularization, e: single scale L-G regularization, f: multiscale L-G regularization

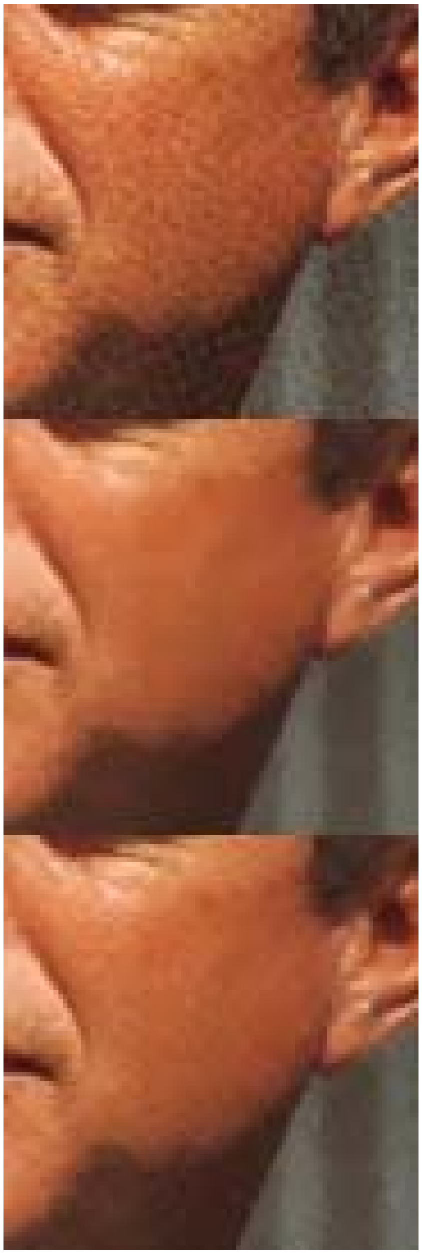


Figure 4.12: Detailed region of JFK color denoising experiment. From top to bottom: Original image, image denoised with TV regularization, image denoised with multiscale L-G mixture regularization.

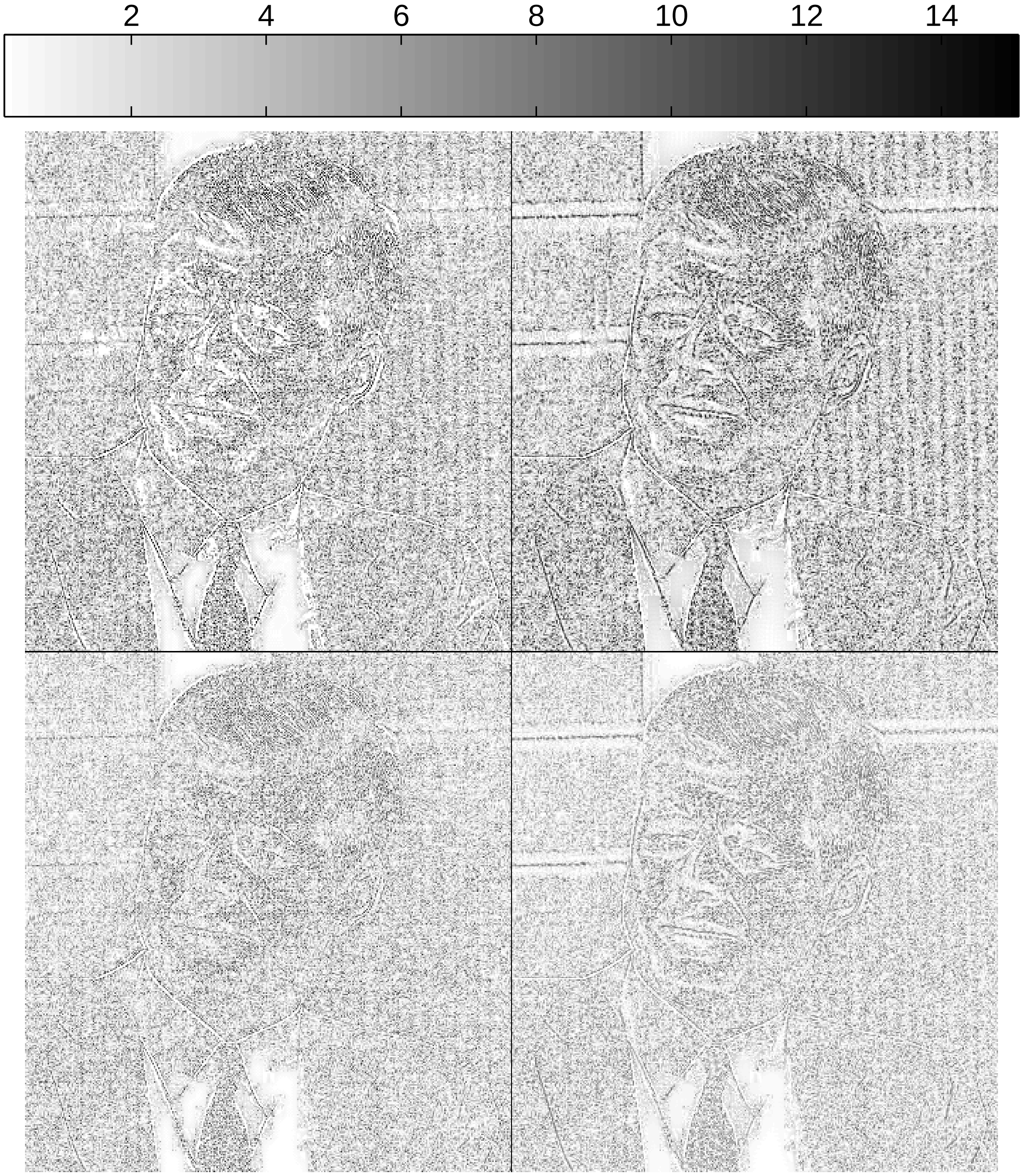


Figure 4.13: Luminance residuals from JFK color denoising experiment. From left to right and top to bottom: TV regularization, B-TV regularization, single scale L-G mixture regularization, multiscale L-G mixture regularization.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The statistics of pixel differences were studied in Chapter 2. Several common assumptions were studied, and a two component mixture model was proposed, and shown to provide a better fit on average by means of Chi-squared tests of fit. A prior was developed from the Laplace-Gauss (L-G) form of the mixture model in Chapter 3. The prior takes the simple form of a mixture of L_1 and L_2 terms. This prior was then applied to a denoising framework in Chapter 4 in a single scale and multiscale form.

In Chapter 4, experiments were performed under two major criteria. First denoising parameters were adjusted to give maximum PSNR when an original image was known. Second, parameters were adjusted to give the best visual results, with the original image not necessarily known.

In both cases, where an original image was available, the L-G mixture model prior was capable of providing the highest PSNR of the methods tested. The visual results of this experiment agreed with the quantitative results. Tikhonov regularization caused excessive blurring, while TV and B-TV regularization preserved edges but removed texture from smooth regions.

One drawback of parametric regularization methods is the sometimes difficult process of parameter tuning. These difficulties motivated the experiments for which an average set of parameter values was used, with the L-G mixture prior average parameters determined from the statistical analysis of Chapter 2. Fixing average parameter values and adjusting regularization levels by visual inspection more closely mimics the way in which parametric regularization terms are frequently used in practice. Under these tests the L-G mixture regularizer performed better then the other regularization methods, apparent in the residuals of Figures **4.10** and **4.13**.

The L-G mixture regularization provides an attractive option for pixel domain denoising. As shown above, the mixture prior provides improved performance in terms of PSNR as well as visual quality. This increase in performance is at the cost of only a small amount of computation added on to efficient methods. Tikhonov regularization as implemented in Equation 4.4 consists of a few simple matrix shift and addition operations for each iteration. TV regularization, as implemented in Equation 4.7 adds only the *sign* operation for each shift. The single scale L-G mixture regularization only adds an inequality and scalar multiplication for each pixel difference value. The multiscale methods require a more substantial increase in computational cost, with the computations increasing as a function of neighborhood size P. B-TV takes on the order of $P(P+1)$ times more computations than TV, while the multiscale L-G mixture regularizer is a simliar factor more than the single scale L-G mixture regularizer.

The two different forms of the L-G mixture regularization term would be best applied in different settings. In settings where computational cost is a main concern, the single scale version provides an improvement over TV at minimal cost, even with the parameters fixed to the average values. If better denoising performance is necessary, the additional flexibility of adjusting parameters or increasing neighborhood size can easily be applied.

5.2 Future Work

Several interesting extensions of the ideas presented in this thesis can be considered. The mixture model developed in Chapter 2 and the prior of Chapter 3 are not application specific, and have been developed to describe pixel differences across a wide variety of images.

The optimal reconstruction parameters for the denoising experiments are not the same as the parameters that best describe the original image. The optimal parameters also change as a function of noise variance. The relationship between the model parameters describing the original image, and those that give optimal denoising performance could be investigated. The correlation between these parameters and noise levels could also be studied.

The denoising application presented in Chapter 4 is only one specific use of the prior. Many implementations of image reconstruction tasks rely on assumptions on pixel differences for regularization. Tasks such as image deblurring or segmentation could be undertaken with minimal modification to the mixture model prior. Variations on the denoising theme such as interpolation could also be investigated.

The statistical analysis of Chapter 2 was limited to grayscale photographic images. We applied the prior to the separate YCrCb channels of a color image in the last experiment of Chapter 4. However, the analysis was only performed on the luminance channel. A study of the other color channels could provide some insight into the statistics of color images, and provide some benefits to color image processing.

The statistical analysis could also be extended to images from different types of sensors. For example, medical imaging, infrared sensors, or satellite imaging could be studied. Application specific priors could be developed based on the statistics of pixel differences for the different systems.

Another area of interesting theoretical work lies in the cartoon/texture representations of images [16, 17]. There is clearly a relationship between the mixture model and these representations. The exact manner in which they are related is not obvious however. An in depth study of this relationship could prove to be a fruitful area of research.

Appendix A

Expectation Maximization Algorithm and Mixture Model Parameter Estimation

The problem of estimating unknown parameters of mixture density functions has been studied extensively. One particular iterative method for numerically approximating the maximum-likelihood (ML) estimates for mixture densities is the Expectation-Maximization (EM) algorithm. A survey of work relating to the EM algorithm, and mixture density estimation can be found in [31]. This usage of the EM algorithm is the specialization of the more general problem of ML estimation for incomplete data sets [32, 18].

The general EM algorithm is formulated by assuming that an observed set of data, $\mathbf{x}$, is incomplete, and that some missing (or hidden) data, $\mathbf{y}$, is present. The EM algorithm maximizes the log-likelihood function $L(\mathbf{\Theta}) = \log p(\mathbf{x}|\mathbf{\Theta})$, where $\mathbf{\Theta}$ is the vector that parameterizes the density function p, by utilizing the relationship between $p(\mathbf{y}|\mathbf{\Theta})$ and

$p(\mathbf{x}|\mathbf{\Theta})$. The EM algorithm is an iterative process, where $\mathbf{\Theta^{(t)}}$ is a vector of parameter estimates at the t-th iteration of the algorithm. For $\mathbf{\Theta}$ and $\mathbf{\Theta^{(t)}}$ the log-likelihood function can be written as follows

$$L(\mathbf{\Theta}) = Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) - H(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}), \tag{A.1}$$

where $Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) = E(\log p(\mathbf{y}|\mathbf{\Theta})|\mathbf{x}, \mathbf{\Theta^{(t)}})$ and $H(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) = E(\log p(\mathbf{y}|x, \mathbf{\Theta})|\mathbf{x}, \mathbf{\Theta^{(t)}})$. With a current guess, $\mathbf{\Theta^{(t)}}$, of the ML parameter estimate, the next approximation, $\mathbf{\Theta^{(t+1)}}$, is found by an iteration of the EM algorithm [18]. First, the E-step is performed, in which $Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}})$ is determined. Next, the M-step finds $\mathbf{\Theta^{(t+1)}} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}})$. It has been shown that $L(\mathbf{\Theta})$ is monotonically increasing on any iterative sequence of the EM algorithm [18]. The EM algorithm is guaranteed to converge to a local maximum of the likelihood function, and in the case of mixture density estimation problems, the solution of the likelihood function is the strongly consistent ML parameter estimate [31]. (*Strongly consistent* means converging to the true parameter values as the sample size approaches infinity with probability *equal* to 1.)

The relevant problem to this thesis is ML estimation of finite mixture densities. A mixture density of m components can be defined as follows

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{i=1}^{m} \alpha_i p_i(\mathbf{x}|\theta_i), \tag{A.2}$$

where $\mathbf{x}$ is a vector of N samples $\{x_1, \cdots, x_N\}$, and where the mixing parameters α_i are nonnegative and $\sum_{i=1}^{m} \alpha_i = 1$. Each p_i is a density function parameterized by θ_i. The parameters for the entire mixture, $\mathbf{\Theta}$ are therefore $(\alpha_i, \cdots, \alpha_m, \theta_1, \cdots, \theta_m)$.

It is apparent that a direct solution for the ML parameter estimates for the mixture densities of Equation A.2 is not viable. If $\mathbf{x}$ is considered to be incomplete and a hidden variable $\mathbf{y} = \{y_1, \cdots, y_N\}$ is defined, the mixture density estimation problem can be posed in the framework of the EM algorithm. Each y_i indicates which mixture component generated the corresponding x_i. The variable $\mathbf{y}$ therefore takes on the values $(1, \cdots, m)$. The conditional density $p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta^{(t)}})$ is given by

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta^{(t)}}) &= \prod_{k=1}^{N} p(y_k|\mathbf{x}, \mathbf{\Theta^{(t)}}) \\ &= \prod_{k=1}^{N} \frac{\alpha_{y_k}^{(t)} p_{y_k}(x_k|\theta_{y_k}^{(t)})}{p(x_k|\mathbf{\Theta^{(t)}})} \\ &= \prod_{k=1}^{N} \frac{\alpha_{y_k}^{(t)} p_{y_k}(x_k|\theta_{y_k}^{(t)})}{\sum_{l=1}^{m} \alpha_l^{(t)} p_l(x_k|\theta_l^{(t)})} \end{aligned} \tag{A.3}$$

The function $Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}})$ for a finite mixture density is then given by

$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) &= \sum_{i=1}^{m} \left[\sum_{k=1}^{N} \frac{\alpha_i^{(t)} p_i(x_k|\theta_i^{(t)})}{p(x_k|\mathbf{\Theta^{(t)}})} \right] \log \alpha_j + \sum_{i=1}^{m} \sum_{k=1}^{N} \log p_i(x_k|\theta_i) \frac{\alpha_i^{(t)} p_i(x_k|\theta_i^{(t)})}{p(x_k|\mathbf{\Theta^{(t)}})} \\ &= \sum_{i=1}^{m} \sum_{k=1}^{N} p(y_k = i|x_k, \mathbf{\Theta^{(t)}}) \log \alpha_j + \sum_{i=1}^{m} \sum_{k=1}^{N} \log p_i(x_k|\theta_i) p(y_k = i|x_k, \mathbf{\Theta^{(t)}}) \end{aligned} \tag{A.4}$$

All terms of this equation can be solved for, and in general, if closed form ML parameter estimates exist for each individual component of the mixture distribution, then a closed form solution will exist for each iteration of the EM algorithm. The two sets of summations of Equation A.4 can be maximized independently. The estimate for α_i is given by

$$\begin{aligned} \hat{\alpha}_i &= \frac{1}{N} \sum_{k=1}^{N} p(y_k = i|x_k, \mathbf{\Theta^{(t)}}) \\ &= \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i^{(t)} p_i(x_k|\theta_i^{(t)})}{\sum_{l=1}^{m} \alpha_l^{(t)} p_l(x_k|\theta_l^{(t)})}. \end{aligned} \tag{A.5}$$

The estimates for the other distribution parameters are found by differentiating $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)})$ with respect to each parameter and equating to zero. The derivations for the two mixture model forms used in this thesis follow directly from Equation A.4. The Laplace-Gauss mixture model is defined as follows

$$\begin{aligned} p(x_k|\mathbf{\Theta}) &= \alpha_1 p_L(x_k|\mu_L, s) + \alpha_2 p_G(x_k|\mu_g, \sigma^2) && \text{(A.6)} \\ &= \alpha \frac{1}{2s} \exp\left(\frac{|x_k - \mu_L|}{s}\right) + (1-\alpha)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_k - \mu_G)^2}{2\sigma^2}\right) && \text{(A.7)} \end{aligned}$$

where $\mathbf{\Theta} = (\alpha, \mu_L, s, \mu_G, \sigma^2)$ and where $\alpha = \alpha_1$ and where $\alpha_2 = 1 - \alpha$. The Q function of Equation A.4 can be written in terms of this specific mixture distribution.

$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) &= \sum_{i=1}^{2}\sum_{k=1}^{N} p(y_k = i|x_k, \mathbf{\Theta^{(t)}}) \log \alpha_j \\ &\quad + \sum_{k=1}^{N}\left(-\log(2s) - \frac{|x_k - \mu_L|}{s}\right) p(y_k = i|x_k, \mathbf{\Theta^{(t)}}) \\ &\quad + \sum_{k=1}^{N}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i - \mu_G)^2}{2\sigma^2}\right) p(y_k = i|x_k, \mathbf{\Theta^{(t)}}), \end{aligned}$$

where $y_k = 1$ indicates that the x_k was derived from the Laplace distribution, and $y_k = 2$ indicates that it is drawn from the Gaussian component.

Differentiating with respect to each parameter, equating to zero, and solving each equation gives the following estimators

$$
\begin{aligned}
\hat{\mu}_L &= \arg_{\mu} \sum_{k=1}^{N} sign(x_k - \mu) p(y_k = 1 | x_i, \mathbf{\Theta}^{(t)}) && \text{(A.8)}\\
\hat{s} &= \frac{\sum_{k=1}^{N} |x_i - \mu_L| p(y_k = 1 | x_i, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 1 | x_i, \mathbf{\Theta}^{(t)})} && \text{(A.9)}\\
\hat{\mu}_G &= \frac{\sum_{k=1}^{N} x_i p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})} && \text{(A.10)}\\
\hat{\sigma}^2 &= \frac{\sum_{k=1}^{N} (x_i - \hat{\mu}_G)^2 p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})} && \text{(A.11)}
\end{aligned}
$$

The corresponding estimators for the zero-mean Laplace-Gauss mixture distribution are easily shown to be

$$
\begin{aligned}
\hat{s} &= \frac{\sum_{k=1}^{N} |x_i| p(y_k = 1 | x_i, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 1 | x_i, \mathbf{\Theta}^{(t)})} && \text{(A.12)}\\
\hat{\sigma}^2 &= \frac{\sum_{k=1}^{N} (x_i)^2 p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 2 | x_i, \mathbf{\Theta}^{(t)})} && \text{(A.13)}
\end{aligned}
$$

Another distribution examined is the Generalized Gauss-Gauss (GG-G) mixture defined by the following equation

$$
\begin{aligned}
p(x_k|\mathbf{\Theta}) &= \alpha_1 p_{GG}(x_k|\mu_L, s) + \alpha_2 p_G(x_k|\mu_g, \sigma^2)\\
&= \alpha \frac{b}{2a\Gamma(1/b)} \exp - \left(\frac{|x_k|}{a}\right)^b + (1-\alpha)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_k - \mu_G)^2}{2\sigma^2}\right),
\end{aligned}
\tag{A.14}
$$

where $\Gamma(\cdot)$ denotes the gamma function. The corresponding Q function for the GG-G mixture is given as follows

$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta^{(t)}}) &= \sum_{i=1}^{2}\sum_{k=1}^{N} p(y_k = i|x_k|\mathbf{\Theta^{(t)}}) \log \alpha_j \\ &+ \sum_{k=1}^{N}\left(\log b - \log(2a\Gamma(1/b)) - \left(\frac{|x_k|}{a}\right)^b\right) p(y_k = 1|x_k, \mathbf{\Theta^{(t)}}) \\ &+ \sum_{k=1}^{N}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i - \mu_G)^2}{2\sigma^2}\right) p(y_k = 2|x_k, \mathbf{\Theta^{(t)}}), \end{aligned} \tag{A.15}$$

The estimator for the a parameter has a closed form expression, and can be shown to be

$$\hat{a} = \left(\frac{b\sum_{k=1}^{N} |x_k|^b p(y_k = 1|x_k, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 1|x_k, \mathbf{\Theta}^{(t)})}\right)^{1/b}. \tag{A.16}$$

The estimator for the b parameter does not have a closed form solution, and can be shown to be the solution to the following equation

$$\begin{aligned} 1 \quad + & \quad \frac{\Psi(1/b)}{b} - \frac{\sum_{k=1}^{N} |x_k|^b \log|x_k| p(y_k = 1|x_k, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} |x_k|^b p(y_k = 1|x_k, \mathbf{\Theta}^{(t)})} \\ & + \frac{1}{b}\log\left(\frac{b\sum_{k=1}^{N} |x_k|^b p(y_k = 1|x_k, \mathbf{\Theta}^{(t)})}{\sum_{k=1}^{N} p(y_k = 2|x_k, \mathbf{\Theta}^{(t)})}\right) = 0, \end{aligned} \tag{A.17}$$

where $\Psi(\cdot)$ denotes the digamma function, which is the first order derivative of $\log\Gamma(\cdot)$, and is given by

$$\Psi(\tau) = -\zeta + \int_0^1 (1 - t^{\tau-1})(1-t)^{-1} dt, \tag{A.18}$$

with ζ denoting the Euler constant.

As in the case of the ML estimates for the Generalized Gaussian function, the Newton-Raphson iterative procedure provides a solution. An implementation for the Generalized Gaussian ML estimates, such as the one written by Do and Vetterli [15] can be modified to estimate the parameters of the GG component of the mixture distribution. The second partial derivative with respect to b is used by this iterative procedure, and is given by

$$
\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})}{\partial b^2} \quad = \quad & \frac{1-\Psi(1/b)}{b^2} - \frac{\Psi\prime(1/b)}{b^3} - \frac{\sum_{k=1}^{N}(|x_k|^b \log|x_k|)^2 p(y_k=1|x_k,\boldsymbol{\Theta}^{(t)})}{\sum_{k=1}^{N}|x_k|^b p(y_k=1|x_k,\boldsymbol{\Theta}^{(t)})} \\
& + \left(\frac{\sum_{k=1}^{N}|x_k|^b \log|x_k| p(y_k=2|x_k,\boldsymbol{\Theta}^{(t)})}{\sum_{k=1}^{N}|x_k|^b p(y_k=1|x_k,\boldsymbol{\Theta}^{(t)})}\right)^2 + \frac{\sum_{k=1}^{N}|x_k|\log|x_k| p(y_k=2|x_k,\boldsymbol{\Theta}^{(t)})}{b\sum_{k=1}^{N}|x_k|^b p(y_k=1|x_k,\boldsymbol{\Theta}^{(t)})} \\
& - \frac{1}{b^2}\log\left(\frac{b\sum_{k=1}^{N}|x_k|^b \sum_{k=1}^{N}}{\sum_{i=1}^{N} p(y_k=1|x_k,\boldsymbol{\Theta}^{(t)})}\right). \qquad \text{(A.19)}
\end{aligned}
$$

Bibliography

[1] D. Odom and P. Milanfar, "Modeling multiscale differential pixel statistics," in *Proceedings of the SPIE Conf. on Computational Imaging*, vol. 6065, Feb. 2006.

[2] ——, "Image denoising using mixture model priors on pixel differences," *submitted to IEEE Signal Processing Letters.*

[3] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[4] R. Reininger and J. Gibson, "Distributions of the two dimensional DCT coefficients for images," *IEEE Transactions on Communication*, vol. COM-31, no. 6, pp. 835–839, June 1983.

[5] D. Field, "Relations between the statistics of natural images and the response properties of corical cells," *Journal of the Optical Society of America*, vol. 4, no. 12, pp. 2379–2394, 1987.

[6] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[7] J. Huang and D. Mumford, “Statistics of natural images and models,” in *CVPR*, vol. 1, June 1999, pp. 541–547.

[8] S. Zhu and D. Mumford, “Prior learning and gibbs reaction-diffusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1236–1250, Nov. 1997.

[9] E. Simoncelli and E. Adelson, “Noise removal via Bayesian wavelet coring,” in *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 1, 1996, pp. 379–382.

[10] M. Green, “Statistics of images, the TV algorithm of Rudin-Osher-Fatemi for image denoising and an improved denoising algorithm,” UCLA CAM, Report 02-55, 2002.

[11] L. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D*, vol. 60, pp. 259–268, Feb. 1992.

[12] D. Ruderman and W. Bialek, “Statistics of natural images: Scaling in the woods,” *Physical Review Letters*, vol. 73, no. 6, pp. 814–818, Aug. 1994.

[13] S. Farsiu, D. Robison, M. Elad, and P. Milanfar, “Fast and robust multiframe super resolution,” *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[14] A. Stuart, J. Ord, and S. Arnold, *Kendall’s Advanced Theory of Statistics*, 6th ed. Arnold, 1999, vol. 2A.

[15] M. Do and M. Vetterli, “Wavelet-based texture retrieval using generalized gaussian density and Kullback-Leibler distance,” *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[16] L. Vese and S. Osher, "Modeling textures with total variation minimization and oscillating patterns in image processing," *Journal of Scientific Computing*, vol. 19, no. 1-3, pp. 553–572, 2003.

[17] T. Le and L. Vese, "Image decomposition using total variation and div(BMO)," *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal*, vol. 4, no. 2, pp. 390–423, 2005.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[19] J. Chang, J. Shin, N. Kim, and S. Mitra, "Image probability distribution based on generalized gamma function," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 325–328, Apr. 2005.

[20] D. Donaho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[21] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive bayesian wavelet shrinkage," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1413–1421, 1997.

[22] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a bayesian approach," *Journal of the Royal Statistical Society B*, vol. 60, pp. 725–749, 1998.

[23] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized gaussian and complexity priors," *IEEE Transactions on Information Theory*, vol. 45, pp. 909–919, 1999.

[24] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, “Image denoising using scale mixtures of gaussians in the wavelet domain,” *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[25] J. Lee, “Digital image enhancement and noise filtering by use of local statistics,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 165–168, Mar. 1980.

[26] J. Besag, P. Green, D. Higdon, and K. Mengersen, “Bayesian computation and stochastic systems,” *Statistical Science*, vol. 10, no. 1, pp. 3–66, 1995.

[27] J. Besag, J. York, and A. Molliè, “Bayesian image restoration, with two applications in spatial statistics (with discussion),” *Annals of the Institute of Statistical Mathematics*, vol. 43, pp. 1–59, 1991.

[28] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Proceedings of IEEE International Conference on Computer Vision*, Jan. 1998, pp. 836–846.

[29] Y. Li and F. Santosa, “A computatinal algorithm for minimizing total variation in image restoration,” *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 987–995, June 1996.

[30] M. Elad, “On the bilateral filter and ways to improve it,” *IEEE Transactions on Image Processing*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002.

[31] R. Redner and H. Walker, “Mixture densities, maximum likelihood and the EM algorithm,” *SIAM Review*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[32] H. Hartley, “Maximum likelihood estimation from incomplete data,” *Biometrics*, vol. 14, pp. 174–194, 1958.