Multidimensional Median Filters for Finding Bumps

Jeffrey C. Miecznikowski
Department of Biostatistics
University of Buffalo
Buffalo, NY 14214
jcm38@buffalo.edu

Kimberly F. Sellers
Department of Mathematics
Georgetown University
Washington, DC 20057
kfs7@georgetown.edu

William F. Eddy
Departments of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
bill@stat.cmu.edu

August 2009

Abstract

One of the main topics within the field of mathematical morphology is the detection of objects, e.g. spots in an image. In general, spot detection or bump hunting in high-dimensional data is a well studied problem. Here, we propose a novel spot detection method based on the smoothing decomposition, *Data = Smooth + Rough* (Velleman, 1980). By studying the rough from a cross shaped smoother, we will show how to locate spots or "bumps" in multi-dimensional images. The essential feature of our novel method is the shape of the window. Usual practice is to choose a window which is (hyper-) cubical or (hyper-) spherical. We have found that by choosing the window to be (hyper-)crossical (i.e. shaped like a multi-dimensional cross), the resulting "rough" is also shaped like a cross centered on the local maxima. We choose to use the median as the summary statistic over the pixels in the smoothing window. Further, we show that the choice of a nonlinear summary statistic, such as the median, is critical in the ability to distinguish the spots in the image. We demonstrate a few properties of this procedure and apply it to a variety of images from various fields, including genetic and proteomic data analysis. The supplemental materials provided contain the supporting theory used for this manuscript.

Keywords: bump hunting, image analysis, spatial smoothing, feature detection, mathematical morphology

1 Introduction

The general subject of mathematical morphology (MM) began in the 1960s and encompasses methods from statistics, machine learning, topology, set theory, and computer science (Serra, 1986; Soille, 2003; Maragos, 1987). It is the science of analyzing and processing geometric structures in digital images. In MM, image processing techniques are applied to digital images to search for geometric structures, e.g. local maxima. Examples of common MM functions include opening, closing, thinning, binning, thresholding, and watershed techniques. A key component in MM is the structuring element, i.e. the shape used to interrogate the image. The two main characteristics that describe the structuring element are its shape and size. In digital images, the structuring element scans the image and alters the pixels in its window content using basic operators similar to Minkowski addition. We introduce and apply a "cross" structuring element in this paper.

The goal of processing images with MM methods can be to preserve the global features of the image, preserve large smooth objects in an image, denoise images, and detect objects within an image. Situations where MM methods are employed for detection include pedestrian detection (Gavrila et al., 2002), tumor mass detection (Tarassenko et al., 1995), and facial feature detection (Saber and Murat Tekalp, 1998; Wang et al., 2002). In this manuscript, we propose using a MM technique with a "cross" shaped structuring element in conjunction with residual analysis to aid in spot finding in biological images.

Spot detection in multi-dimensional images is presently being performed in many different applications ranging from astronomy to proteomics; see, for example, Agard et al. (1981), Bertin and Arnouts (1996), Lindeberg (1998), Cutler et al. (2003), and Lalush (2003). The previously mentioned papers employ MM techniques such as watershed methods, thresholding operators, and wavelet reconstruction methods to find the spots present in an image. In general, feature detection (which includes spot detection) has a growing body of research in high-dimensional datasets; see, for example, Jain and Zongker (1997) and Guo et al. (2008). Likewise, smoothing operators have a rich history throughout statistics and can be used in a variety of ways (Velleman, 1980), including:

- data exploration (Tukey, 1977; Tukey and Mosteller, 1977)
- robust regression (Beaton and Tukey, 1974)
- sequences of regressions (Wainer and Thissen, 1975)
- robust seasonal adjustment (Diggle, 1990)
- signal and image processing (Justusson, 1981; Polzehl and Spokoiny, 2000)
- spatial smoothing (Tibshirani and Wang, 2007).

This paper combines aspects of feature detection and smoothing to develop a new spot detection method for k-dimensional images. In our new method, a specialized median (referred to hereafter as an s-median) smoother is developed, where we focus on the residuals (the rough) as a method of feature detection. Figure 1 is an example of our spot detection method applied to an image consisting of four mountains shaped like Gaussian densities with different variances.

Residual plots, on the other hand, were developed as a diagnostic tool for checking common assumptions associated with model fit such as symmetry, normality, additivity, and variance stability; see, for example, Tukey (1977), Velleman and Hoaglin (1981), and Hoaglin et al. (1991). We propose instead to utilize the residuals obtained from the s-median as a means of feature detection in multi-dimensional datasets. Consider a k-dimensional (kD) image represented by $I_k(\mathbf{x})$, where $\mathbf{x}$ is a point location in the Cartesian coordinate system. Let $R_{k,c}$ denote the kD residual image obtained by using an s-median operator with "arms" of length c. Possible window sizes are illustrated in Figure 2. Accordingly, $R_{2,1}$ and $R_{2,2}$ refer to residual images obtained by using a 3×3 window (i.e. the 5-pixel cross shown in Figure 2(a)) or a 5×5 window (i.e. 9-pixel cross shown in Figure 2(b)), respectively. With this notation, we can explore the R image to locate peaks (in 1D) or spots/bumps (in 2D images), examine how the window size (structuring element) affects the $R_{k,c}$ image, and study how different-sized operators work in the presence of noise.

By introducing a window for smoothing (as illustrated in Figure 2) and utilizing the decomposition, *Rough = Data - Smooth* (Velleman, 1980), we illustrate spot detection in an image, along with the underlying theory that supports the success of this approach. The paper is outlined as follows. Section 2 demonstrates this procedure for the case where the image is noise-free. Section 3 extends these ideas to study the behavior of the s-median operator in the presence of noise. We illustrate the spot detection method via biological examples in Section 4, and give some final thoughts and discussion in Section 5.

2 The Noise-free Case

Consider a kD input image I_k, where $I_k(\mathbf{x})$ denotes the image intensity at the (Cartesian) coordinate $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. Similarly, let $S_{k,c}$ denote the s-median smoothed image, where $S_{k,c}(\mathbf{x})$ is the resulting s-median value at $\mathbf{x}$ (see Supplemental Materials for details regarding s-median computation). We compute the residual image $R_{k,c}$ via $R_{k,c}(\mathbf{x}) = I_{k,c}(\mathbf{x}) - S_{k,c}(\mathbf{x})$. The remainder of this paper characterizes R in terms of I and S.

2.1 One Mountain

Let I_1 be a one-dimensional (1D) mountain, that is, a 1D sequence consisting of a sequence of increasing values followed by decreasing values. Then, for any c, $R_{1,c} \geq 0$, and $R_{1,c} = 0$ when the sequence contained in the window span is monotone. In particular, $R_{1,c}(x) > 0$ if x is the location of the absolute maximum in the sequence. Extending this concept to a k-dimensional mountain, then $R_{k,c} \geq 0$ with a strict inequality occurring at the absolute maximum.

Figure 3(a) is an example of a 1D mountain whose maximum intensity value is at $x = 0$. Figure 3(b)-(c) show the associated residual images, $R_{1,1}$ and $R_{1,7}$, determined via a 3-pixel (i.e. arm size 1) smoothing window, and a 15-pixel (i.e. arm size 7) smoothing window, respectively. As a result, we see the impact that a varying window size can have on peak detection for a 1D mountain. The peak location in Figure 3(b) is clearly more precise than in Figure 3(c). In particular, because of the small window size associated with Figure 3(b), the peak location is the only location with a nonzero residual value. Meanwhile, in Figure

3(c), the larger window span allows for a greater number of cases where the window span does not contain a monotone sequence within the window. Thus, the interval containing nonzero residuals is [-3,3].

Analogously, we consider the 2D example shown in Figure 3(d), where (50,50) is the center of the image and the location of the mountain's peak. Figure 3(e)-(f) show the corresponding $R_{2,c}$ images for s-medians of arm size 2 (i.e. $c = 2$, or a 5×5 cross smoother window consisting of 9 pixels; see Figure 2(b)), and 7 (i.e. $c = 7$, or a 15×15 cross smoother), respectively. The $R_{2,c}$ images each show a "cross" (or $''+''$ shape) formed at the local maximum, but with varying degrees of precision because of the chosen arm length associated with the s-median. Nonetheless, we see that the cross is characteristic of a mountain or spot present in I_2 and can be used to locate spot centers in an image. More interesting examples will involve multiple mountains within an image, as well as mountains in the presence of noise.

2.2 Multiple mountains

Considering multiple mountains requires the introduction of several more concepts and definitions to characterize their presence in an image. In one dimension, we will say that there are multiple mountains present in a 1D image if the 1D sequence of values $\{I_1(\mathbf{x})\}$, $\mathbf{x} = (1, ..., n)$ has more than one subsequence consisting of a strictly increasing sequence followed by a strictly decreasing sequence. Given the presence of multiple (say m) mountains, there exists an absolute maximum and an absolute minimum in the sequence. There also exist, however, m local maxima and $m - 1$ local minima between neighboring mountains. We restrict the local minimum value to be unique, and do not allow for constant sequences within the 1D sequence.

When we consider several mountains in two or more dimensions, it becomes difficult to characterize this scenario with general propositions regarding R. Several factors contribute to this problem, such as (1) the relative intensity of each mountain with respect to the others; (2) the location of the absolute maximum in the image; and (3) the mountain's support ("area"), volume, and shape/profile. Hence, any proposition regarding spot detection in multiple dimensions will require several restrictions and definitions regarding mountain shape and size, thus limiting the practicality of the proposition.

Since our applications will involve detection of multiple spots in biological images, we will present some specific examples showing the residual images obtained from images containing multiple spots. In the single mountain examples (Figure 3), it is already clear that different window sizes (i.e. different c values) impact the visibility and appearance of the cross in the $R_{k,c}$ image. One of the more interesting examples involves two nearby spots. Figure 4(a) shows a perspective plot of two mountains of different intensities and standard deviations. Here, the mountains are surrounded by largely flat regions, but the mountains are relatively close with a saddle point between them. As the value of c increases, the corresponding $R_{2,c}$ image changes from featuring two distinct crosses to showing only one cross. That is, as c increases, the crosses that first appear as two distinct objects become merged into one cross centered between the two spots. Figure 4(b)-(c) show the $R_{2,c}$ images obtained from two different values of c applied to Figure 4(a).

From the example in Figure 4, the change in window size significantly affects spot detection. When the window size (i.e. c) increases, the cross becomes wider. Ultimately, close

spots become blended together with large values of c. In contrast, a small smoothing window will produce a smaller width cross, thus allowing for the separation of two distinct peaks. Figure 4 shows that it can be difficult to characterize the $R_{k,c}$ images in the presence of multiple mountains. Large values of c tend to oversmooth an image such that, for large c, two mountains are likely to be "interpreted" as one large mass in the R image. Meanwhile, small values of c preserve small mountains near each other that may, in fact, be spurious depending on the nature of the noise (see Section 3). Although the $R_{k,c}$ operator for nearby mountains is difficult to characterize, it does allow for spot detection in images with multiple spots.

All of the examples thus far have focused on noise-free images. The following examples, however, explore an $R_{k,c}$ image when noise is introduced in the raw image I_k. As expected, it will make spot detection in $R_{k,c}$ more difficult.

3 The Noisy Case

Thus far, we have examined the R image for noise-free images, I. In the 1D case, we see that $R_{1,c}(x) > 0$ for any c when x is the location of the absolute maximum, and $R_{1,c} = 0$ when the sequence contained in each dimension of the smoothing window is monotone. Further, under certain circumstances associated with 1D images, $R_{1,c}(x) < 0$ when x is the location of a local minimum in our image. We can now extend these ideas to images in the presence of noise. As usual, when studying noisy data, we must bear in mind the signal-to-noise ratio.

We begin with the 1D case, which presents some interesting results. Consider adding independent and identically distributed (i.i.d.) Gaussian noise to the 1D monotonic sequence $\{I(\mathbf{x})\}$, where $\mathbf{x} = (1, ..., n)$. Let $I(i) = g_i + N_i$, where g_i denotes the true signal at location i, s_i equals the step size at i in the monotonic sequence such that $g_i + s_i = g_{i+1}$, and $N_i \sim N(0, \sigma)\ \forall\, i$ denotes normally distributed noise of mean zero and standard deviation σ. We fix $s_i = s$ for our examples such that the signal-to-noise ratio (s/σ) remains constant within each simulation.

We will examine the case when $R_{1,c}(x) = 0$ at an arbitrary location x. We know that $R_{1,c} = 0$ on a strictly increasing or decreasing sequence. In certain cases, a closed-form solution exists to compute the probability that the R image is zero ($Pr(R = 0)$) over a monotone sequence with i.i.d. noise. However, in most cases, it is not possible to compute a closed-form solution for $Pr(R = 0)$. Figure 5 shows the estimated $Pr(R_{1,c} = 0)$ over a monotone sequence for different values of c in the 1D situation. Note that the x-axis is the ratio of stepsize to standard deviation of the noise, and that each curve begins at 1/(2c+1) and asymptotes at 1. The signal-to-noise ratio, s/σ, is the critical value in evaluating how $R_{1,c}$ operates in a noisy environment. If the step size is 0, then each point in the window is equally likely to be the median, hence $Pr(R_{1,c} = 0) = \frac{1}{2c+1}$ for any c when $s = 0$. As the step size increases relative to the standard deviation of the noise, naturally for any c, we expect the probability to converge to one ($Pr(R_{1,c} = 0) \to 1$). Hence, for noise-free monotone images, we have that $R_{1,c} = 0$ for all c.

In the presence of noise, however, the monotone signal becomes contaminated such that $Pr(R_{1,c} = 0)$ decreases as c increases. Intuitively, as c increases, the number of points in the smoothing window increases hence there are more "opportunities" for other points to be the

median, thus making the residual nonzero at that location.

Given the local maximum at $x = p$ in a noise-free 1D mountain, $R_{1,c}(p) > 0$. In the presence of noise, we can estimate (via simulation) the $Pr(R_{1,c}(p) > 0)$ when $I(i) = g_i + N_i$ and

$$g_i = \begin{cases} g_{i+1} - s, & i < p \\ g_{i+1} + s, & i \geq p, \end{cases}$$

i.e., we can estimate the probability that the 1D residual image intensity value at the local maximum location is positive. Figure 6(a) shows the estimated $Pr(R_{1,c}(p) > 0)$ at the absolute maximum location p as a function of s/σ for different values of c in a 1D image. Analogously, Figure 6(b) shows $Pr(R_{2,c}(\boldsymbol{p}) > 0)$ for a maximum location $\boldsymbol{p} = (p_1, p_2)$ when I is a 2D image. For all c, the simulation results show that $Pr(R_{k,c}(\boldsymbol{p}) > 0) \rightarrow 1$ monotonically as the signal-to-noise ratio increases, $k = 1, 2$. Further, as shown in Figures 6(a)-(b), as c increases, the rate in which $Pr(R_{k,c}(\boldsymbol{p}) > 0)$ converges to 1 also increases, $k = 1, 2$.

To further illustrate the importance of the size of the smoothing window in detecting spots, Figure 7 shows the R image for Figure 1 when noise (of the type described above) is added to the image. With a fixed smoothing window, as the standard deviation of the noise increases, the ability to discern the cross decreases. Specifically at a standard deviation of 50, the crosses are nearly indistinguishable from noise for the top two mountains. Recall that, with the noise-free single mountain example, we clearly detected a cross in the rough operator image at the mountain's maximum. Further, the size of the observed cross was directly related to the operating window. As the window size increased, the size of the cross increased as well. Consider adding noise to the mountain in Figure 3. Figure 8(a) shows a single mountain with added i.i.d. N(0,$\sigma = 200$) noise at each pixel. Figures 8(b)-(c) show the associated $R_{2,c}$ images for a $c = 9$ cross and a $c = 27$ cross, respectively. There are several interesting features to note in this example. We see the respective crosses associated with the smoothing window; however, when using the $c = 9$ arm, it is much harder to distinguish the cross from the remaining picture. With the $c = 27$ arm, the cross is more apparent, mainly due to the cross being wider than in Figure 8(b).

To confirm that large values of c more effectively find spots, Figures 9(a)-(c) show a sequence of three spots in order of increasing size with normally distributed noise of mean 0 and standard deviation 48. Figures 9 (d)-(f) are the $R_{2,2}$ images corresponding to Figures 9 (a)-(c), respectively. Figures 9 (g)-(i) are the $R_{2,3}$ images corresponding to Figures 9 (a)-(c). From the image, it is easier to detect larger spots in the presence of noise and in the presence of noise, larger values of c are more effective for detecting spots.

Collectively, Figures 5-9 illustrate the tradeoff that must be considered when determining the arm size for the s-median smoother. We see that large values of c are more likely to yield positive residuals at the maximum in the I image; however, the residuals associated with large values of c are also more likely to be nonzero in the presence of noise over monotonic regions. In other words, for spot finding, large values of c will improve spot detection in noisy images, however, it may cause two distinct spots to become one large spot in the presence of noise. A balance between these two issues will be critical in choosing the optimal c value(s) for peak or spot finding (see Section 5).

4 Examples

The biological imaging domain is attractive for this spot detection method. Numerous biological applications involve spot or feature detection in images including mass spectrometry, gel electrophoresis, and genetic microarrays. In mass spectrometry, the relevant data are represented as spectra where the associated peaks in the intensity plots represent proteins (or peptides) present in a sample. Obtaining the location and intensity of these peaks aides in identifying sample proteins for further study consideration. Gel electrophoresis data are represented in the form of 2D images comprised of protein spots. Again, investigators are interested in detecting these features in order to isolate their location in the image and potentially extract the associated protein sample for further analysis. Finally, microarray data are represented as two-dimensional images of spots in a 2D matrix structure. Feature detection is key in order for the genetic data to be properly summarized for future analysis. Thus, feature detection in biological images is important since image and subsequent analyses are critical for these technologies to have utility in diagnosing disease or assessing putative biomarkers.

4.1 Mass Spectrometry

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS) is a technology that can be used to profile protein markers from tissue or bodily fluids, such as serum or plasma in order to compare biological samples from different patients or different conditions. The output from a MALDI experiment consists of a measured intensity for each mass-to-charge ratio (m/z) value; see Figure 10(a). The sets of expressed proteins are identified within each spectrum in order to ultimately determine differentially expressed proteins between conditions or samples. See Karas et al. (1990) for further details describing the MALDI technology.

Our s-median derived R image can be used to detect peaks in MALDI based images and thus locate peptides present in the sample. The MALDI spectrum for each sample consists of a single vector, I, thus applying the s-median is equivalent to applying a running median to the I image. Using a MALDI dataset obtained from the Proteomics Core Laboratory at Roswell Park Cancer Institute, we can examine the results of applying the s-median to a MALDI spectrum. For this data, we set the bandwidth (i.e. the value of c in $R_{1,c}$) to 500 data points, which corresponds to approximately a 95 m/z bandwidth. Figure 10(b) shows the resulting s-median image using the chosen bandwidth; Figure 10(c) shows the associated $R_{1,500}$ image. From examining the R image, we note that the spikes in the original spectrum are preserved, thus aiding in the identification of the peak location. Further, we notice that we also have "negative peaks" in the residual image. Further work will explore the utility of these "negative peaks" in quantifying the peaks in a MALDI image.

4.2 Gel Electrophoresis

Another application of this spot detection technique is on images obtained from two dimensional difference gel electrophoresis (2D-DIGE) experiments such as those described in Sellers et al. (2007). For our 2D-DIGE examples, we will focus on images representing portions of

the 2D gels examining morphogenesis in *Drosophila* obtained from the Minden laboratory at Carnegie Mellon University (Gong et al., 2004; Mergliano and Minden, 2003). Studying the protein spots in the images allows the researchers to obtain a protein expression signature of the sample under a given condition or given time point. The images under study have been normalized according to the model described in Sellers et al. (2007). The full images are 1024 pixels $\times$ 1280 pixels and densely populated with protein spots, making it difficult to observe individual protein spots in detail. We therefore focus on a 50 pixel $\times$ 50 pixel sub-image to better understand the result of applying the s-median.

Figure 11(a) shows the protein gel sub-image selected for illustration, I, with the associated perspective plot shown in Figure 11(b). Figure 11(c) displays the associated residual image, $R_{2,6}$. From Figure 11(c), we can see the crosses associated with the protein spots shown in Figure 11(a). As well, we also see that each protein spot is outlined in black since, in the noise-free case, the R image is negative at local minima in an image. We can use the black outline as a boundary identification tool to determine spot size in order to more accurately determine summary information and excise the protein sample(s) of interest from the gel. In 2D-DIGE experiment, after quantification of the protein spots under different channels or conditions, similar to gene microarrays, the spot ratios are computed and compared to assess the degree of differential expression.

4.3 Pin Microarrays

Genetic microarrays are a popular analysis tool to study genetic changes associated with disease. The image obtained from a microarray experiment consists of a series of spots indicating the measured fluorescence of a probe (or "gene") deposited at that location. See Schena et al. (1995) for a detailed description of the microarray technology, and Gentleman (2005) for an overview of the methods used for microarray analysis. In this technology, image analysis software is required to summarize the signal for a given spot on a chip. In this situation, we can examine the R image obtained from a pin microarray image for proper identification of spot locations and spots sizes to aid in spot quantitation and data summarization.

Figure 12(a) shows an example of a microarray image obtained from a cell cycle yeast experiment (Fink et al., 1998). Similar to the gel electrophoresis example, Figure 12(b) examines a subsection of the microarray chip shown in Figure 12(a). Figure 12(c) shows the associated residual image ($R_{2,6}$) obtained from applying an s-median to Figure 12(b). A closer inspection of Figure 12(c) reveals a black spot within the center of each microarray probe. This is an interesting phenomenon attributed to the manufacturing of the microarray. Occasionally, the impact of the pin onto the microarray chip displaces the probe material and causes a "donut" shape probe hybridization profile. The hybridization spot has a "hole" in the middle since there was little or no probe material deposited to hybridize. This effect is not obvious in Figure 12(b) but is clearly distinguished in Figure 12(c)-(d). This kind of information can be used to improve the estimation of spot intensity in the microarray image. The spot intensity estimates are used as input for downstream processing, ultimately, yielding the expression value for each probe representing the amount of hybridized genetic material.

5 Discussion

The classic equation, $rough = data - smooth$, has been the standard paradigm within statistics. In this paper, we demonstrate a new MM operator where the residual image derived from a novel smoother can be used to locate spots or mountains in an image. Thus this method combines the residual operator from statistics with the structuring element (cross shape window) in the field of mathematical morphology. The method is fast to implement and is not biased by the scale of the mountain, e.g. regardless of the mountain's size and height, its location will be detected via a cross in the associated R image. This aspect alleviates the need to alter or change the grey scales in an image when searching for spots of varying intensities.

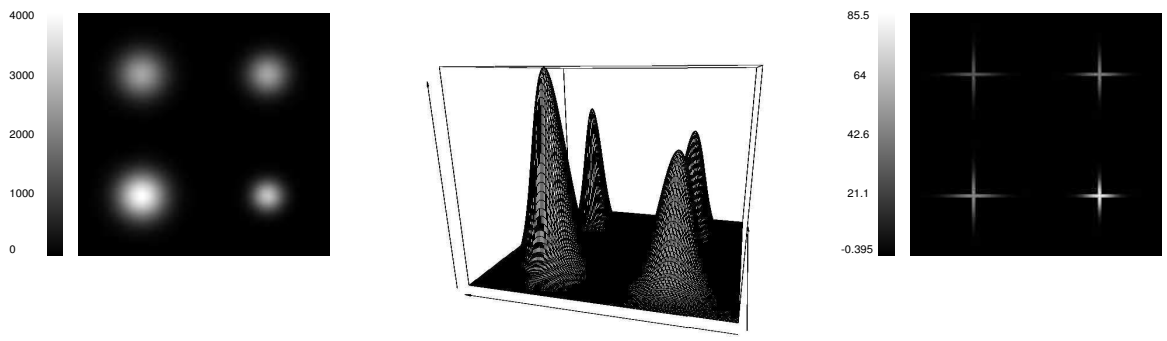
As demonstrated, this method uses the s-median operator to smooth images. Other window operators can be considered, however they result in different residual image implications. For example, if a mean cross smoother (i.e. "s-mean") is used on the Gaussian mountain in Figure 13(a) rather than a median smoother, the residual image does not reveal the shape of the mountain; see, e.g., Figure 13(b). Further, the shape of the smoothing window is also a critical component of consideration. Figure 13(c) displays the results when a median smoother with a grid or "box" shaped window sequence is used. Here, we now obtain a residual image that looks like a starburst instead of a cross. As a result, the spot center is now potentially more difficult to identify. The shape of the smoothing window (cross vs. box) and the summary statistic used (median vs. mean) thus affect the R image and the ability to detect the mountains in an image.

The issue of rotational invariance is an important concept within mathematical morphology operators used in image detection. Rotational invariance implies that the resultant image does not change when arbitrary rotations are applied to its input argument. In general, our spot finding method is rotational invariant for the Gaussian spots with zero correlation, e.g., spots of the type shown in Figure 1. Interestingly, if we induce any nonzero correlation in the spot, the spot finding method is no longer rotationally invariant. Figure 14(a) displays a bivariate normal density with a correlation of 0.50 between the two variables. Figure 14 (b) is the residual image from our proposed method. Meanwhile, Figure 14(c) is the result when employing a rotated version (45 degrees) of the structuring element used in Figure 14(b). Similarly, Figure 14(d) is the rotated version (90 degree) of Figure 14(a) with the corresponding R images shown in Figures 14(e) - (f). Our proposed spot finding method is not rotational invariant since the images in Figure 14(b) and Figure 14(e) are clearly different. Although our proposed method is not rotational invariant, it is possible to rotate our structuring element (cross) to align with the major and minor axes of a correlated spot as in Figures 14(c) and (f). Both versions of the residual images clearly show a cross shape and provide utility in terms of locating the spots in the image. Future work will further explore the characteristics of the cross in each residual image in order to detect spots in correlated images. Note, however, in our biological applications (e.g. 2D-DIGE), it is reasonable to assume that there is negligible correlation within a spot. For example in a DIGE image, the spots are created by electrophoresis in two dimensions where the electrophoresis for each dimension is performed separately.

When using the s-median operator for spot finding, the major consideration is the arm-length size c associated with the smoothing window, or alternatively the number of pixels

included in the smoothing window (structuring element). The s-median smoother naturally removes noise from I, hence the size of the smoothing window essentially decides the amount of smoothing to apply to the dataset. From Figures 4 and 9, the choice of c is critical, since choosing c too large will oversmooth the image and blend spots together, while choosing c too small will undersmooth the image and cause spurious spots due to noise to appear as real spots. Since the choice of c is essentially choosing a smoothing parameter, there are several available methods to consider when choosing an optimal value for c. The general methods for smoothing parameters include cross validation method and bootstrap algorithms. Future work will examine the data driven cross validation schemes for choosing an optimal value of c for specific image applications.

In conclusion, this manuscript develops a new method for spot finding and illustrates the technique's great utility within several biological settings such as mass spectrometry spectra, gel electrophoresis images, and microarray images. Further, this method can be easily extended to mountains in k dimensions and has applications in other domains, including economics, finance, astronomy, physics, chemistry, etc.



(a) (b) (c)

Figure 1: **Bump Hunting**: (a) A 200 pixel × 200 pixel image consisting of four Gaussian spots with different locations and scales. (b) The associated perspective plot. (c) The "rough" residual image ($R_{2,4}$ image) after running a specialized median smoother (s-median). Clearly a cross is present at the location of the respective spot centers associated with the spots shown in (a).

1	2	3
4	5	6
7	8	9

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

(a) (b)

Figure 2: **Examples of Window Shapes in 2D**: Window shape, often called the structuring element in morphology, refers to the pixels used to compute the median for smoothing. Note that, in this figure, the window shapes are shaded in grey. (a) A 3×3 s-median cross window consisting of 5 pixels. This s-median replaces the intensity of pixel 5 with the median intensity of the grey pixels (2,4,5,6,8). (b) A 5×5 s-median cross window consisting of 9 pixels. This s-median replaces the intensity of pixel 13 with the median intensity of the grey pixels (3,8,11,12,13,14,15,18,23).

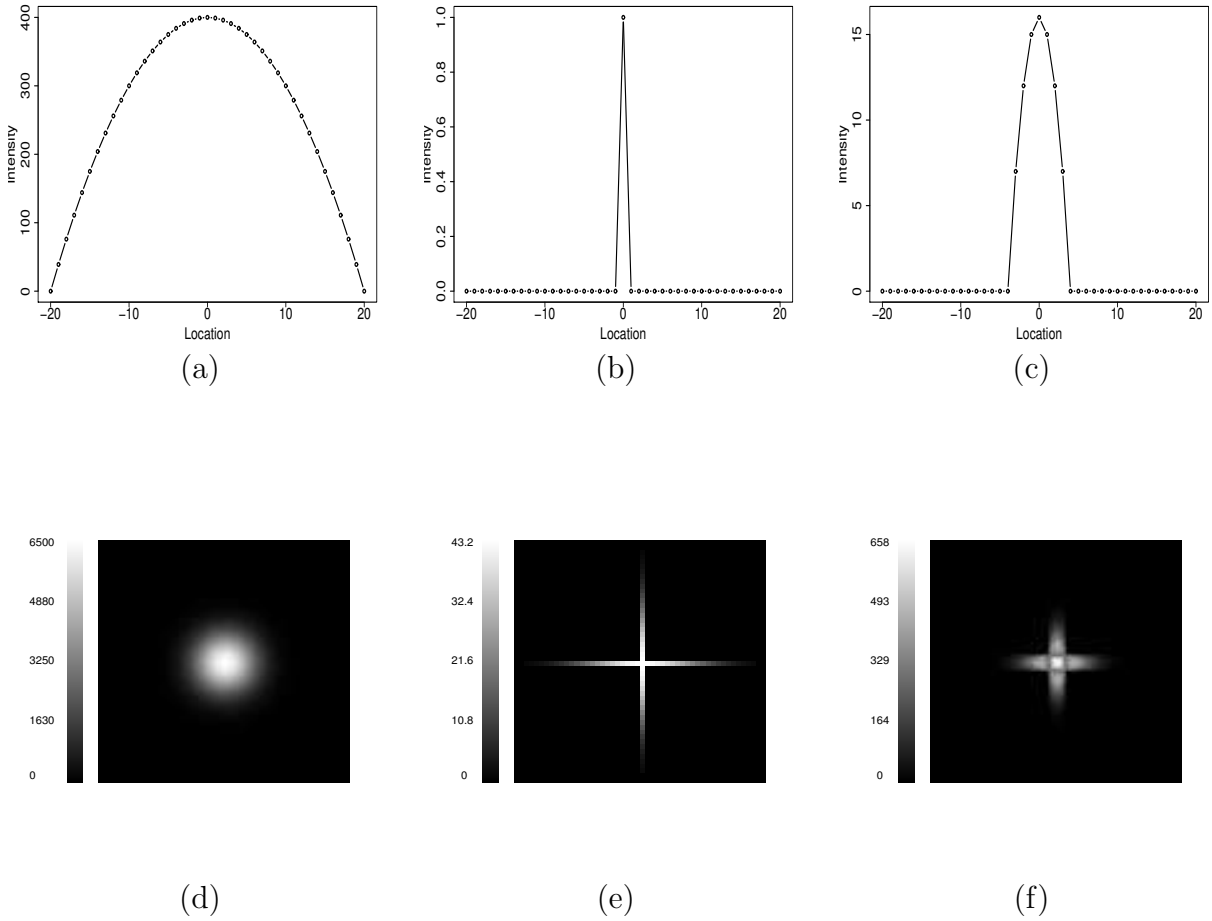
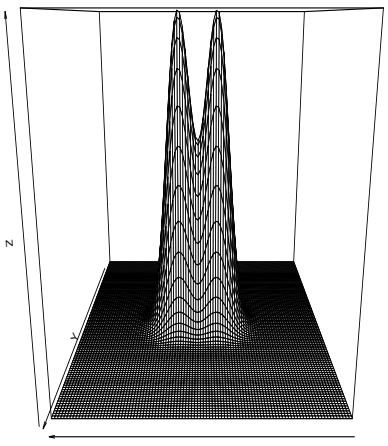
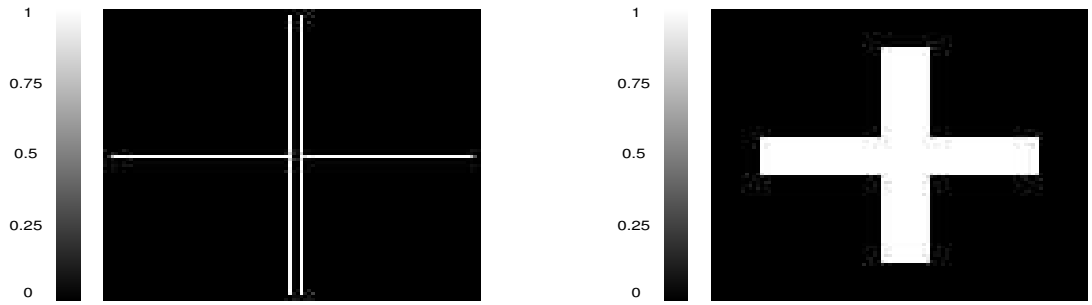


Figure 3: $R_{k,c}$ **for an image** I **of a single mountain**: (a) An image of a 1D mountain generated by $I_1(x) = -x^2 + 400$, where x consists of the integers between -20 and 20 (i.e. 41 points). (b) $R_{1,1}$, i.e. the residual image associated with I_1 when applying a smoothing window of 3 points ($c = 1$). Note that all of the residuals equal 0 except for the residual intensity at the center of the mountain. (c) $R_{1,7}$, i.e. the residual image associated with I_1 when applying a smoothing window of 15 points ($c = 7$). The residuals are greater than 0 at the location of the top of the mountain and several adjacent locations. (d) A 100 × 100 pixel image of a two-dimensional mountain generated by $I_2(\mathbf{x}) = 6500 * \exp\left(-\frac{(x_1-50)^2+(x_2-50)^2}{150}\right)$. (e) $R_{2,2}$, i.e. the residual image associated with I_2 using a 5×5 s-median where the smoothing window consists of 9 pixels ($c = 2$). The characteristic "cross" is clearly present in this image. (f) $R_{2,7}$, i.e. the residual image associated with I_2 using a 15×15 s-median where the smoothing window contains 29 pixels ($c = 7$). Clearly the "cross" is wider than that appearing in $R_{2,2}$, but the arms are not as long.



(a)



(b) (c)

Figure 4: **Two Nearby Mountains**: (a) Perspective plot showing two relatively close mountains. (b) The $R_{2,5}$ operator image associated with the image in (a). The two crosses indicate the presence of two relative maxima in the image. (c) The $R_{2,27}$ image obtained from the image in (a). In this situation, the two mountains are "blurred" into one cross.

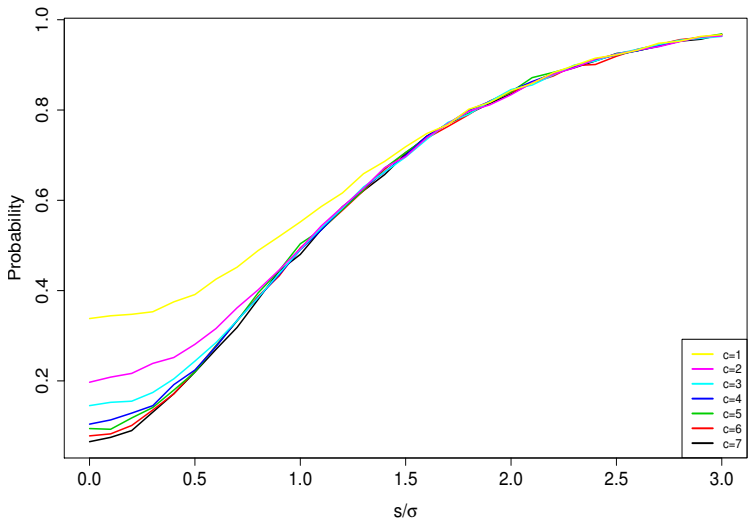


Figure 5: **Estimated Probability of $R = 0$**: For different values of c, the size of the smoothing window, the y axis is the estimated probability that the R image is zero over a monotone signal of step size s contaminated with i.i.d. normally distributed noise with standard deviation σ. The x axis is the signal-to-noise ratio.

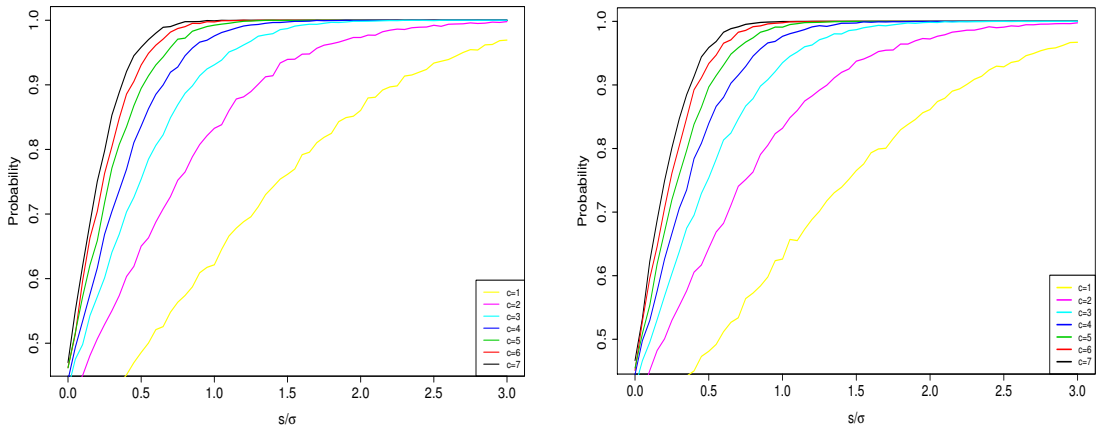
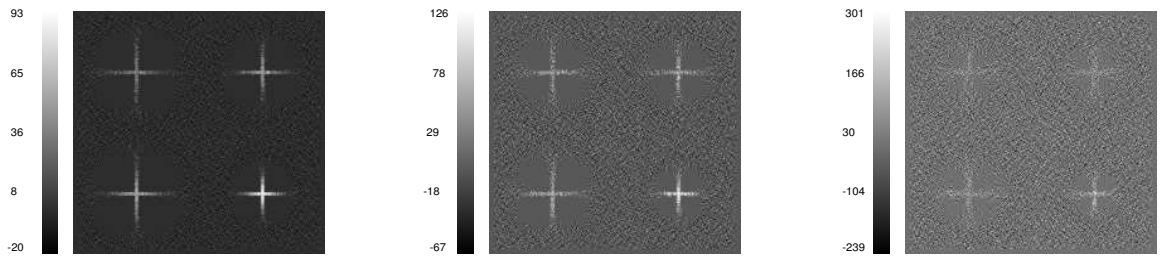
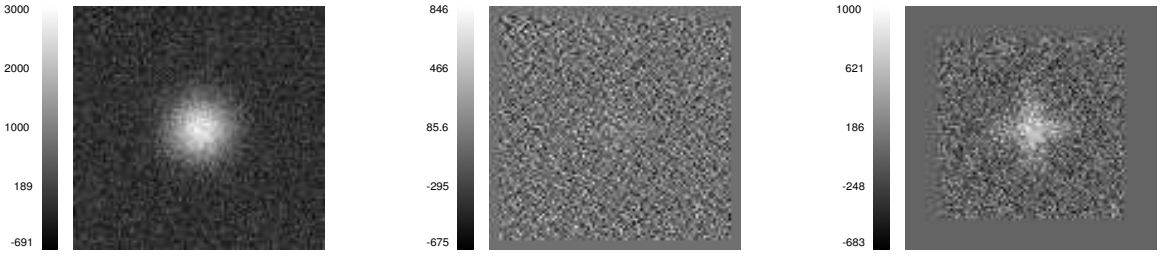


Figure 6: **Estimated Probability of $R > 0$ at a local maximum in 1D and 2D**: (a) $Pr(R_{1,c}(p)) > 0$, where p is the location of the absolute maximum in a 1D image. (b) $Pr(R_{2,c}(p_1, p_2)) > 0$ for a 2D image, where (p_1, p_2) represents the location of the absolute maximum in the 2D image.



(a) (b) (c)

Figure 7: **Spot Finding as Noise Increases**: (a) $R_{2,9}$ image associated with Figure 1, where i.i.d. normally distributed noise (with mean 0 and standard deviation 5, i.e. N(0,5)) was added at each point. (b) $R_{2,9}$ image when the noise in Figure 1 is N(0,$\sigma = 15$). (c) $R_{2,9}$ image when the noise in Figure 1 is N(0,$\sigma = 50$). As the standard deviation of the noise increases, the ability to detect the cross at each spot decreases.



(a) (b) (c)

Figure 8: **Mountains with Noise Added**: (a) A 100 $\times$ 100 image of a 2D mountain where noise is normally distributed with mean of 0 and standard deviation of 200. (b) The associated $R_{2,9}$ image; the characteristic "cross" is very difficult to discern by eye. (c) The $R_{2,27}$ image; the characteristic "cross" is still difficult to detect because of the noise, however it is more detectable than the cross in (b).

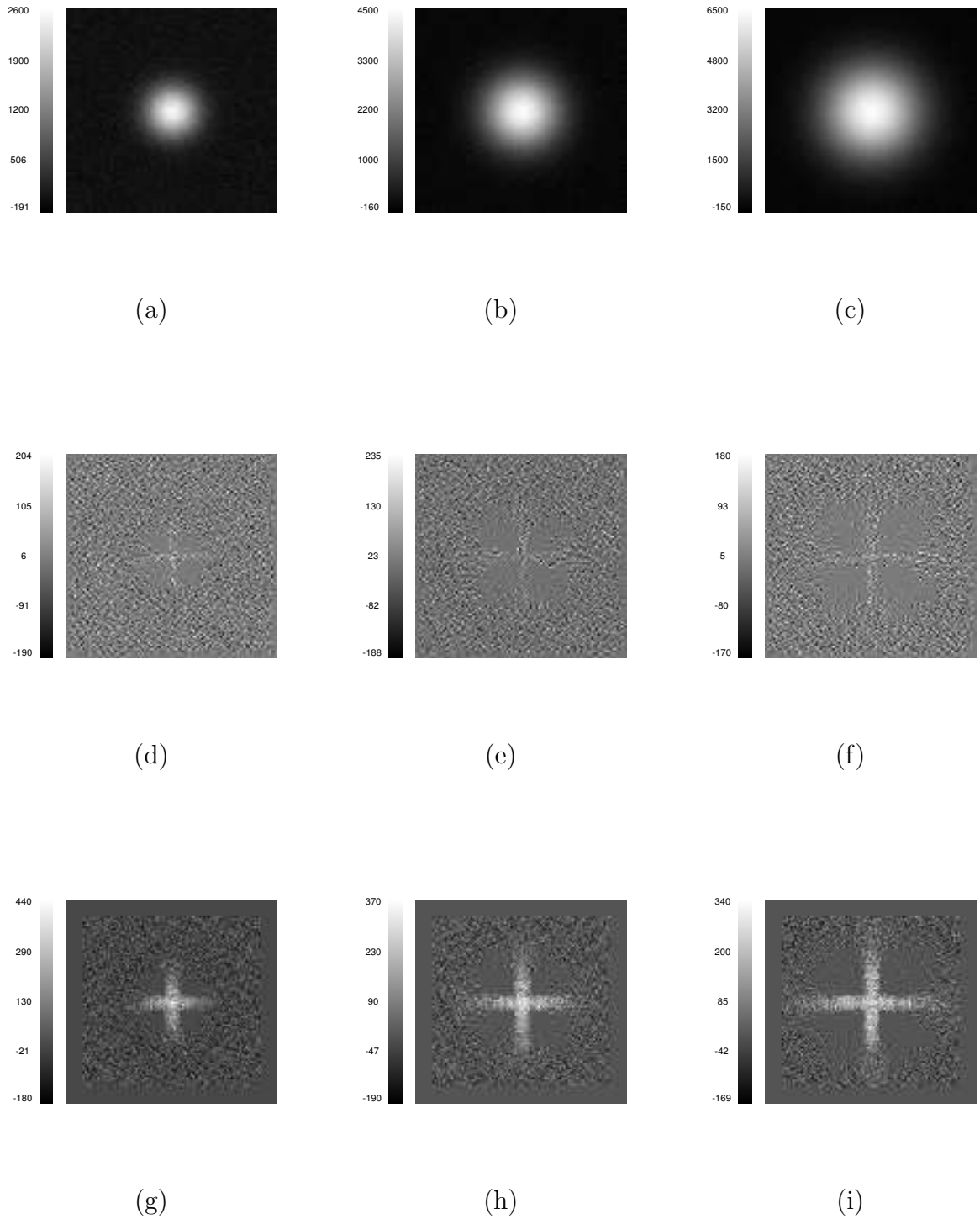
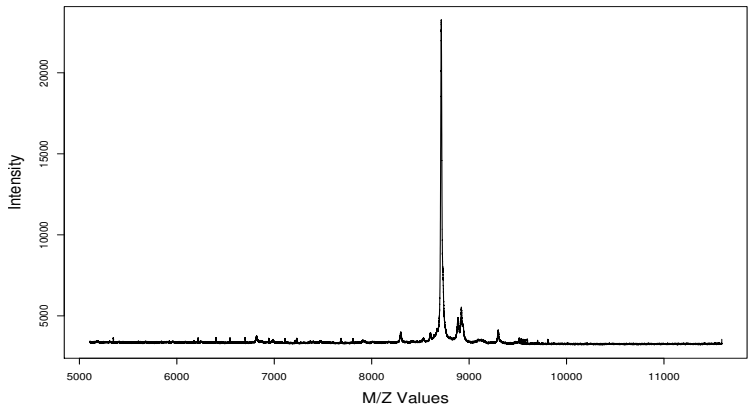
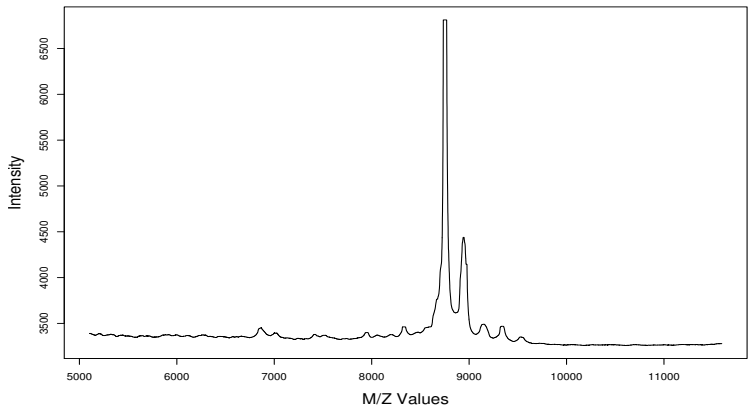


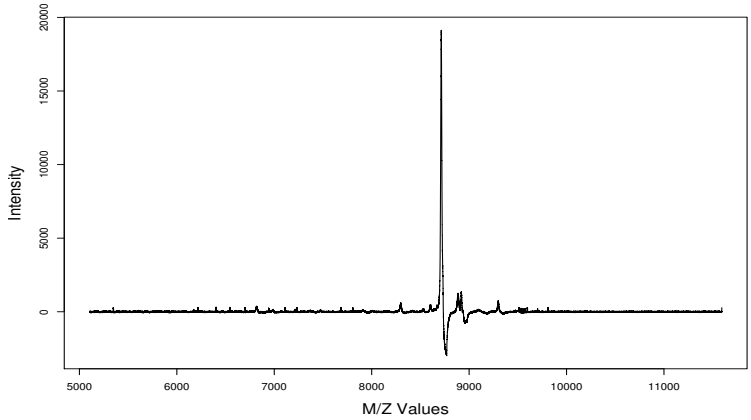
Figure 9: **Choice of Smoothing Operator**: A series of Gaussian mountains of increasing size and normally distributed noise of mean 0 and standard deviation of 48 are shown in (a), (b) and (c). The residual images from applying an $R_{2,2}$ operator to (a), (b), and (c) are shown in (d), (e), and (f), respectively. The residual images from applying an $R_{2,3}$ operator to (a), (b), and (c) are shown in (g), (h), and (i), respectively. The spots are more easy to distinguish using the $R_{2,3}$ operator (row 3) rather than the $R_{2,2}$ operator (row 2).



(a)

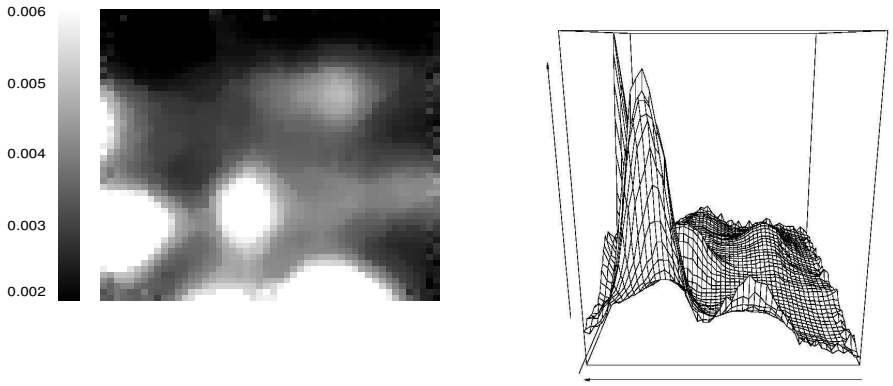


(b)

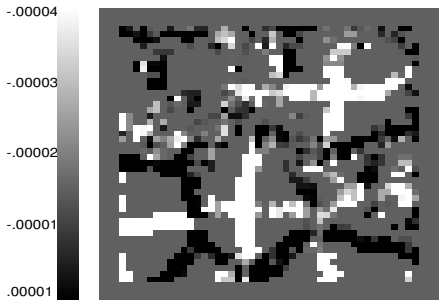


(c)

Figure 10: **Mass Spectrometry**: (a) MALDI spectrum based on a tumor sample. (b) S image based on a smoothing window of 95 m/z. (c) Associated R image. Note that the R image contains spikes at each local maximum in the original image I.

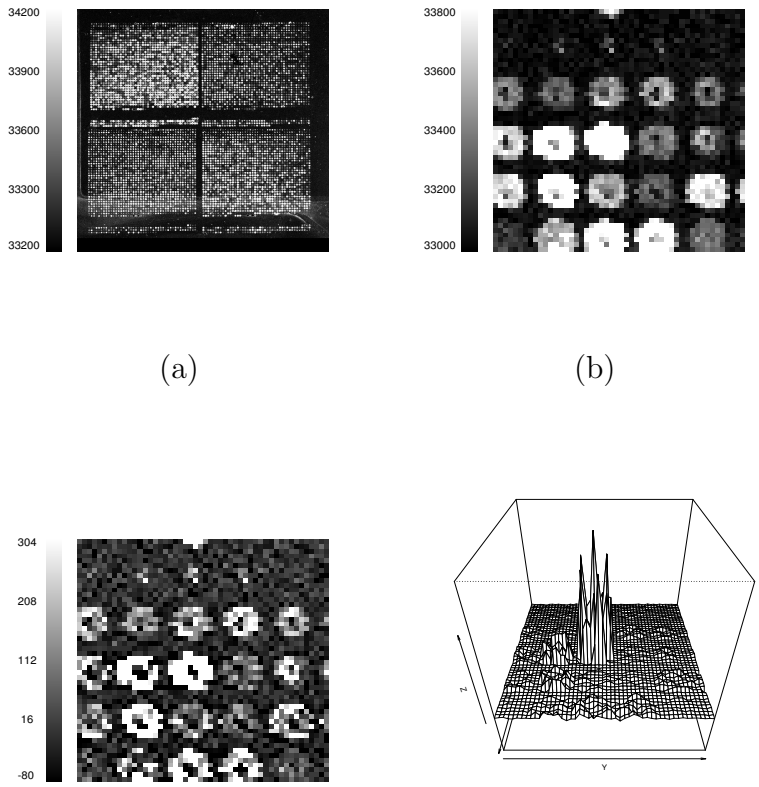


(a) (b)



(c)

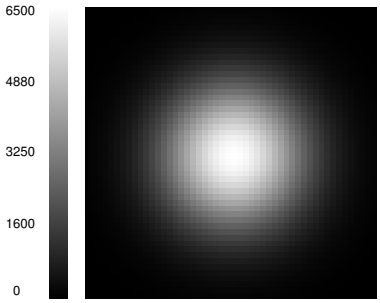
Figure 11: **2D-DIGE Images**: (a) Subset of 2D-DIGE image. This 50 pixel $\times$ 50 pixel image is a subimage of the *Drosophilia* proteome gel, normalized according to the model in Sellers et al. (2007). (b) The associated perspective plot for the data in (a). (c) The associated $R_{2,6}$ image. There are several "crosses" apparent, which indicate protein spots in the image. There is also "speckled" black and white noise pattern present in the image with a black outline around several of the spots.



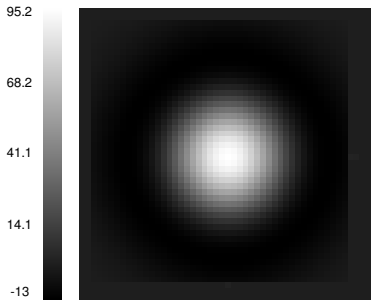
(a) (b)

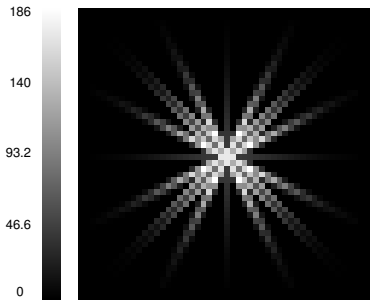
(c) (d)

Figure 12: **Gene Microarray**: (a) Full image of a microarray chip containing roughly 4000 spots. (b) Microarray subimage showing roughly 25 spots. (c) $R_{2,4}$ image corresponding to s-median smoother applied to (b). Note the black pixels at the center of each spot indicate local minima. Common phenomena of pin microarray images are local minima due to the lack of probe material caused by the pin containing the probe material impacting against the glass slide. This impact causes a "donut" pattern in the perspective plot of the microarray image. This "donut" pattern is clearly recognized in (c). (d) A perspective plot of the data shown in (c) further detailing the "donut" pattern.



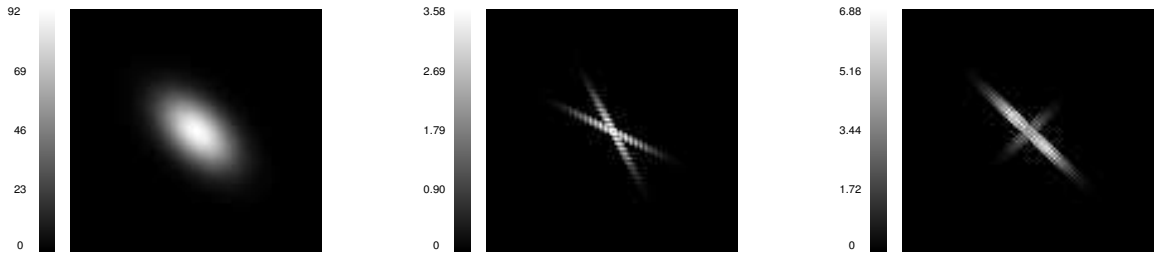
(a)



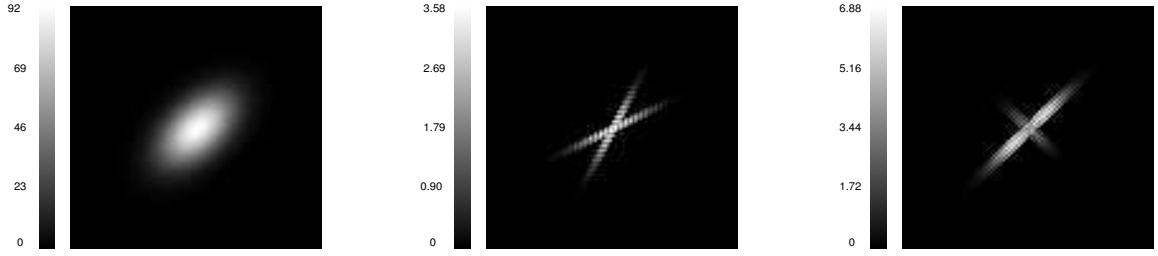


(b) (c)

Figure 13: **Other measures**: The results from different variations on the s-median. (a) A Gaussian spot image, I, of dimension 50×50. (b) The $R_{2,2}$ image obtained from using an s-mean rather than s-median on the image in (a). (c) Associated residual image obtained from (a) using an s-median, where the window sequence is a 5 pixel $\times$ 5 pixel box shape containing all 25 pixels. It is not immediately clear if there is a utility in this image.



(a) (b) (c)



(d) (e) (f)

Figure 14: **Rotational Invariance**: (a) A scaled bivariate normal density with a correlation of 0.50. (b) The resulting residual image using a $R_{2,4}$ operator. (c) The resulting residual image when the structuring element in the residual operator used in (b) is rotated 45 degreed to align with the major axis of the spot in (a). (d) a 90 degree rotation of the spot in (a), i.e. a spot with a correlation of -0.50. (e) The residual image using a $R_{2,4}$ operator. The resulting residual image when the structuring element used in (e) is rotated 45 degrees.

SUPPLEMENTAL MATERIALS

Theory for spot finding Supporting theory for the spot finding method.

References

Agard, D., Steinberg, R., Stroud, R., 1981. Quantitative analysis of electrophoretograms: A mathematical approach to super-resolution. Analytical Biochemistry 111, 257–268.

Beaton, A., Tukey, J., 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 20 (11), 147–185.

Bertin, E., Arnouts, S., 1996. Sextractor: software for source extraction. Astronomy and Astrophysics 14 (4).

Cutler, P., Heald, G., White, I. R., Ruan, J., 2003. A novel approach to spot detection for two-dimensional gel electrohporesis images using pixel value collection. Proteomics 3, 392–401.

Diggle, P., 1990. Time series: a biostatistical introduction. Oxford University Press.

Fink, G., Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B., 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell 9 (12), 3273–3297.

Gavrila, D., Giebel, J., Perception, M., Res, D., Ulm, G., 2002. Shape-based pedestrian detection and tracking. In: IEEE Intelligent Vehicle Symposium, 2002. Vol. 1.

Gentleman, R., 2005. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer.

Gong, L., Puri, M., Unlu, M., Young, M., Robertson, K., Viswanathan, S., Krishnaswamy, A., Dowd, S., Minden, J., 2004. Drosophila ventral furrow morphogenesis: a proteomic analysis. Development 131 (3), 643–656.

Guo, B., Damper, R. I., Gunn, S. R., Nelson, J. D. B., 2008. A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. Pattern Recognition 41 (5), 1670–1679.

Hoaglin, D., Mosteller, F., Tukey, J., 1991. Fundamentals of Exploratory Analysis of Variance. Wiley-Interscience.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2), 153–158.

Justusson, B., 1981. Median Filtering: Statistical Properties. Two-Dimensional Digital Signal Processing II: Transforms and Median Filters, 161.

Karas, M., Bahr, U., Ingendoh, A., Nordhoff, E., Stahl, B., Strupat, K., Hillenkamp, F., 1990. Principles and applications of matrix-assisted UV-laser desorption/ionization mass spectrometry. Anal. Chim. Acta 241, 175–185.

Lalush, D. S., 2003. Effects of spot and background defects on quantitative data from spotted microarrays. Proceedings of the 25th Annual International Conference of the IEEE, 3563–3566.

Lindeberg, T., 1998. Feature detection with automatic scale selection. International Journal of Computer Vision 30 (2), 77–116.

Maragos, P., 1987. Tutorial on advances in morphological image processing and analysis. Opt. Eng. 26 (7), 623–632.

Mergliano, J., Minden, J., 2003. Caspase-independent cell engulfment mirrors cell death pattern in Drosophila embryos. Development 130 (23), 5779–5789.

Polzehl, J., Spokoiny, V. G., 2000. Adaptive weights smoothing with applications to image restoration. Journal of the Royal Statistical Society. Series B 62 (2), 335–354.

Saber, E., Murat Tekalp, A., 1998. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. Pattern Recognition Letters 19 (8), 669–680.

Schena, M., Shalon, D., Davis, R., Brown, P., 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270 (5235), 467.

Sellers, K., Miecznikowski, J., Viswanathan, S., Minden, J., Eddy, W., 2007. Lights, camera, action: Quantitative analysis of systematic variation in two-dimensional difference gel electrophoresis. Electrophoresis 28 (18), 3324–3332.

Serra, J., 1986. Image analysis and mathematical morphology. 1982. New York: Academic Press. IS. Zhuang, X. and RM Haralick, Morphological Structuring Element Decomposition. Computer Vision, Graphics, and Image Processing 35, 370–382.

Soille, P., 2003. Morphological image analysis: principles and applications. Springer-Verlag New York, Inc. Secaucus, NJ, USA.

Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M., 1995. Novelty detection for the identification of masses in mammograms. In: Artificial Neural Networks, 1995., Fourth International Conference on. pp. 442–447.

Tibshirani, R., Wang, P., 2007. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics.

Tukey, J., 1977. Exploratory Data Analysis. Addison-Wesley, NY.

Tukey, J., Mosteller, F., 1977. Data analysis and regression. Addison-Wesley Series in Behavioral Science: Quantitative Methods, Addison-Wesley, Reading, MA.

Velleman, P., 1980. Definition and comparison of robust nonlinear data smoothing algorithms. JASA 75 (371), 609–615.

Velleman, P., Hoaglin, D., 1981. Applications, Basics, and Computing of Exploratory Data Analysis. Cornell Cooperative Extension.

Wainer, H., Thissen, D., 1975. Multivariate semi-metric smoothing in multiple prediction. Journal of the American Statistical Association 75 (371), 568–573.

Wang, Y., Chua, C., Ho, Y., 2002. Facial feature detection and face recognition from 2D and 3D images. Pattern Recognition Letters 23 (10), 1191–1202.

Supplemental Material for Multidimensional Median Filters for Finding Bumps

August 2009

Overview

The following sections detail the s-median computations for the 1D continuous function and the 1D and 2D discrete functions.

S-median Computation on Continuous Functions

In this section, we develop the s-median smoother and thus the $R_{k,c}$ image in the context of continuous functions. The s-median is the median value corresponding to a spatially-structured sample space. Thus, before calculating the s-median (or even the median), we must characterize the distribution of the output values from a function when given a distribution of input values. Note that, here, we assume the function to be continuous. Once we obtain the distribution for the output values, then s-median estimation is straightforward (i.e. computing $R_{k,c}(\mathbf{x}) = I_k(\mathbf{x}) - S_{k,c}(\mathbf{x})$).

One-dimensional Continuous Functions

The distribution of the random variable $G(X)$, and thus the median of $G(X)$, is straightforward to compute when G is a monotone function over the sample space for X (Wasserman, 2004). Let $\Phi_X(x)$ and $\phi_X(x)$ denote the cumulative density function (cdf) and probability density function (pdf), respectively, for the arbitrary random variable X evaluated at the point x; analogously, we denote the cdf and pdf for a random variable Y at point y. Let $G(x) : \mathcal{X} \to \mathcal{Y}$ be a function that defines a mapping from the support set $\mathcal{X}$ (for the random variable X) to the support set $\mathcal{Y}$ (for the random variable Y). The goal is to obtain an expression for $\Phi_Y(y)$ [and thus $\phi_Y(y)$] which then allows for a straightforward calculation of M_Y, the median of Y (i.e. $\Phi_Y(M_Y) = 0.5$).

Monotone Case

Consider the case where the function $Y = G(X)$ is strictly monotone on the interval (x_0, x_n). Then, for G increasing,

$$\Phi_Y(y) = \begin{cases} 1 & y \geq G(x_n) \\ \frac{G^{-1}(y) - x_0}{x_n - x_0} & G(x_0) \leq y \leq G(x_n) \\ 0 & y \leq G(x_0), \end{cases}$$

while, for G decreasing,

$$\Phi_Y(y) = \begin{cases} 1 & y \geq G(x_0) \\ \frac{x_n - G^{-1}(y)}{x_n - x_0} & G(x_n) \leq y \leq G(x_0) \\ 0 & y \leq G(x_n). \end{cases}$$

These computations allow us to determine $\Phi_Y(y)$, and thus take its derivative to obtain $\phi_Y(y)$. Strict monotonicity in G implies its invertibility, i.e. $\Phi_Y(y) = \Phi_X(G^{-1}(y))$ for any y; in particular, by definition of M_Y, we have $\Phi_Y(M_Y) = 1/2 = \Phi_X(G^{-1}(M_Y))$. Hence, in the monotone case, $M_Y \equiv G(M_X)$, where M_X denotes the median associated with the random variable X, and $\equiv$ denotes equivalence of the statistics as defined in Lehmann and Casella (1998), p.36. Thus, the median for Y is equivalent to the function evaluated at the median for X. The remainder of this section assumes that X follows a Uniform distribution on (x_0, x_n), i.e. $X \sim U(x_0, x_n)$. Hence, in the monotone case, $M_Y \equiv G\left(\frac{x_n+x_0}{2}\right)$. The next section extends these ideas to consider a piecewise monotone function, G.

Piecewise monotone case

We define $\mathcal{X} = \{x : \phi_X(x) > 0\}$ as an open interval (x_0, x_n) with $n \in \mathbb{Z}^+$. Let $A_i = (x_i, x_{i+1})$, $i = 0, \cdots, n-1$, be the smallest collection of disjoint open intervals such that G is strictly monotone on each A_i. By definition, G is strictly monotone increasing on A_i if, for any two values $\{x_1, x_2\} \in A_i$ such that $x_1 < x_2$, $G(x_1) < G(x_2)$ holds. Analogously, G is strictly monotone decreasing if $G(x_1) > G(x_2)$ for $\{x_1, x_2\} \in A_i$ such that $x_1 < x_2$. Note that continuity may not be enough for the $\{A_i\}$ to be countable. Further, note that we assume strict monotonicity in the function G. Similarly, we define the interval

$$B_i = G(A_i) = \begin{cases} (G(x_i), G(x_{i+1})), & \text{for } G \text{ increasing on } A_i, \\ (G(x_{i+1}), G(x_i)), & \text{for } G \text{ decreasing on } A_i. \end{cases}$$

While the sequence $\{A_i\}$ partitions $\mathcal{X}$, $\{B_i\}$ does not necessarily partition $\mathcal{Y}$; e.g., see Figure 1, where $\mathcal{X} = (-1, 2)$ and $G(x) = x^3 - x$.

Let $G_i(x) = G(x)I_{A_i}(x)$; i.e., $G_i(x)$ denotes the A_ith decomposition of $G(x)$, and I defines the indicator function of A_i by $I_{A_i}(x) = I(x \in A_i) = 1$ (0) if x is in (not in) the interval (x_i, x_{i+1}). By definition, $G_i(x)$ is strictly monotone. Thus, for $x \in \mathcal{X}$, we have $G(x) = \sum_{i=0}^{n-1} G(x)I_{A_i}(x) = \sum_{i=1}^{n-1} G_i(x)$, implying that any function G can be decomposed into the sum of its strictly monotone components, G_i. With this notation established, we have the following for $\Phi_Y(y)$:

$$\begin{aligned}
\Phi_Y(y) &\equiv Pr(Y \leq y) = Pr(G(X) \leq y) \\
&= \sum_{i=0}^{n-1} Pr(G(X) \leq y \mid X \in A_i) Pr(X \in A_i) \\
&= \sum_{i=0}^{n-1} Pr(X \leq G_i^{-1}(y)) Pr(X \in A_i) \\
&= \sum_{i=0}^{n-1} \Phi_X(G_i^{-1}(y)) \cdot \left(\frac{x_{i+1} - x_i}{x_n - x_0} \right),
\end{aligned}$$

where, for G_i increasing on A_i,

$$\Phi_X(G_i^{-1}(y)) \equiv Pr(X \leq G_i^{-1}(y)) = \begin{cases} 1 & y \geq G(x_{i+1}) \\ \frac{G_i^{-1}(y) - x_i}{x_{i+1} - x_i} & y \in (G(x_i), G(x_{i+1})) \\ 0 & y \leq G(x_i), \end{cases}$$

and, for G_i decreasing on A_i,

$$\Phi_X(G_i^{-1}(y)) \equiv Pr(X \leq G_i^{-1}(y)) = \begin{cases} 1 & y \leq G(x_{i+1}) \\ \frac{G_i^{-1}(y) - x_i}{x_{i+1} - x_i} & y \in (G(x_i), G(x_{i+1})) \\ 0 & y \geq G(x_i). \end{cases}$$

For all but the most simple functions, there is no closed form solution by which to define M_Y. The above equations, at least, provide an explicit equation by which to determine it.

Shifting the function

There are a few generalizations for the median in the 1D case where we consider vertical and horizontal shifts of the function, G. For the following, we define $Y = G(X)$.

Lemma 0.1. *Let $Y' = Y + C = G(X) + C$. Then $M_{Y'} \equiv M_Y + C$, i.e. a vertical shift of size C in the function G implies a vertical shift in M_Y by the same amount, C.*

S-median Computation on Discrete Functions

Let $\tilde{G}(X)$ denote a "discrete" function, i.e. $\tilde{G}$ denotes discrete/countable realizations from the continuous function, $G(X)$.

One-dimensional Discrete Functions

Let $\Phi_X(x)$ and $\phi_X(x)$ denote the discrete cdf and probability mass function (pmf), respectively, for the arbitrary random variable X evaluated at the point x. Let $\tilde{G}(x) : \mathcal{X} \to \mathcal{Y}$ denote discrete/countable realizations from the continuous function $G(X)$, i.e. $\tilde{G}$ is a function that defines a mapping from the support set $\mathcal{X}$ (of the discrete random variable X) to the support set $\mathcal{Y}$ (for the discrete random variable Y). We want to obtain an expression for $\Phi_Y(y)$ to then calculate the s-median of Y.

In the discrete case, the calculation of the s-median will proceed by assuming a discrete uniform distribution for X (i.e. $X \sim$ Discrete Uniform(N) as defined in Casella and Berger

(1990)), and then computing $\Phi_Y(y)$ based on the function, $\tilde{G}(X)$. Thus, the machinery developed in the section for continuous functions can be applied to compute M_Y.

We can analogously represent $\Phi_Y(y)$ using discrete random variables X and Y as we did for the continuous case, namely

$$\Phi_Y(y) = \sum_{i=0}^{n-1} Pr(X \leq \tilde{G}_i^{-1}(y)) Pr(X \in A_i), \tag{1}$$

where assuming $X \sim$ Discrete Uniform(N) implies that

$$Pr(X \leq \tilde{G}_i^{-1}(y)) = \begin{cases} 0 & \tilde{G}_i^{-1}(y) < 1 \\ \frac{\lfloor \tilde{G}_i^{-1}(y) \rfloor}{N} & 1 \leq \tilde{G}_i^{-1}(y) \leq N \\ 1 & \tilde{G}_i^{-1}(y) \geq N, \end{cases} \tag{2}$$

and $Pr(X \in A_i) = \frac{\lceil x_i \rceil - \lfloor x_{i+1} \rfloor}{N}$. , where "$\lfloor \rfloor$" denotes the floor function and "$\lceil \rceil$" denotes the ceiling function. For the special case of a strict monotone discrete function $\tilde{G}$ on the full interval $\{1, \cdots, N\}$,

$$\Phi_Y(y) = \begin{cases} 0 & \tilde{G}_i^{-1}(y) \leq 1 \\ \frac{\lfloor \tilde{G}_i^{-1}(y) \rfloor}{N} & 1 \leq \tilde{G}_i^{-1}(y) \leq N \\ 1 & \tilde{G}_i^{-1}(y) \geq N; \end{cases}$$

note that this relation does not depend on the direction of monotonicity for $\tilde{G}$. These computations allow us to determine $\Phi_Y(y)$, and thus compute the s-median for Y (i.e. M_Y) in 1D.

Two-dimensional Discrete Functions

As in the continuous case, we must be precise in how we define the sample space Ω for X and Y in order to compute the s-median for $Z = G(X, Y)$. For the 2D discrete case, we have the following definition for the sample space:

Definition 0.2. *Let X be a discrete uniform on $(1, N)$, Y be discrete uniform on $(1, N)$, and (x^*, y^*) be a fixed point such that x^* and $y^* \in \{1, 2, ..., N\}$. Then let Ω be of the form:*

$$\Omega = \{(x, y) | \{(x^*, y) \text{ where } y \in \{1, 2, ..., N\}\} \cup \{(x, y^*) \text{ where } x \in \{1, 2, ..., N\}\}\}$$

Figure 2 illustrates the associated sample space. With this definition, the derivation of the s-median for Z follows analogously to the 2D continuous case (which is analogous to the 1D case).

Let $G : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ define a mapping from the support sets $\mathcal{X}$ and $\mathcal{Y}$, of the random variables X and Y respectively, to the support set $\mathcal{Z}$, for the random variable Z.

We define the functions, $G_{x^*}(y) = G(x^*, y)$ and $G_{y^*}(x) = G(x, y^*)$, such that $G_{x^*}(y) = G_{x^*}(y) \sum_{j=0}^{k-1} I_{B_j}(y) = \sum_{j=0}^{k-1} G_{x^*}(y) I_{B_j}(y) \equiv \sum_{j=0}^{k-1} G_{x^*,j}(y)$, and $G_{y^*}(x) = \sum_{i=0}^{n-1} G_{y^*}(x) I_{A_i}(x)$ $\equiv \sum_{i=0}^{n-1} G_{y^*,i}(x)$, where $A_i = (x_i, x_{i+1})$ is the smallest set of disjoint open intervals such that $G_{y^*}(x)$ is strictly monotone on each A_i, $i = 0, \cdots, n-1$; and $B_j = (x_j, x_{j+1})$ is the smallest set of disjoint open intervals such that $G_{x^*}(y)$ is strictly monotone on each B_j, $j = 0, \cdots, k-1$.

In this setting,

$$
\begin{aligned}
\Phi_Z(z) &= Pr(G(X,Y) \leq z) \\
&= Pr(\{G_{y^*}(X) \leq z\} \cup \{G_{x^*}(Y) \leq z\}) \\
&= Pr(G_{y^*}(X) \leq z) + Pr(G_{x^*}(Y) \leq z) - \underbrace{Pr(\{G_{x^*}(Y) \leq z\} \cap \{G_{y^*}(X) \leq z\})}_{=0} \\
&= \sum_{i=0}^{n-1} Pr(G_{y^*}(X) \leq z \mid X \in A_i) Pr(X \in A_i) + \\
&\quad \sum_{j=0}^{k-1} Pr(G_{x^*}(Y) \leq z \mid Y \in B_i) Pr(Y \in B_i) \\
&= \sum_{i=0}^{n-1} \underbrace{Pr\left(X \leq G^{-1}_{y^*,i}(z)\right)}_{\text{known}} \underbrace{Pr(X \in A_i)}_{\text{known}} + \sum_{j=0}^{k-1} \underbrace{Pr\left(Y \leq G^{-1}_{x^*,j}(z)\right)}_{\text{known}} \underbrace{Pr(Y \in B_j)}_{\text{known}}
\end{aligned}
$$

defines the cdf of Z. Nicely, all of the above quantities can be computed since we specified the distributions for X and Y. Note the zero quantity in the third line is due to the intersection of the sets containing only the single point (x^*, y^*). The $Pr(Y \in B_j)$ and $Pr(X \in A_i)$ depend on the length of B_j and A_i, respectively.

References

Casella, G., Berger, R. L., 1990. Statistical Inference. Duxbury Press, Belmont, California.

Lehmann, E., Casella, G., 1998. Theory of Point Estimation. Springer, New York.

Wasserman, L., 2004. All of Statistics: A Concise Course in Statistical Inference. Springer, New York.

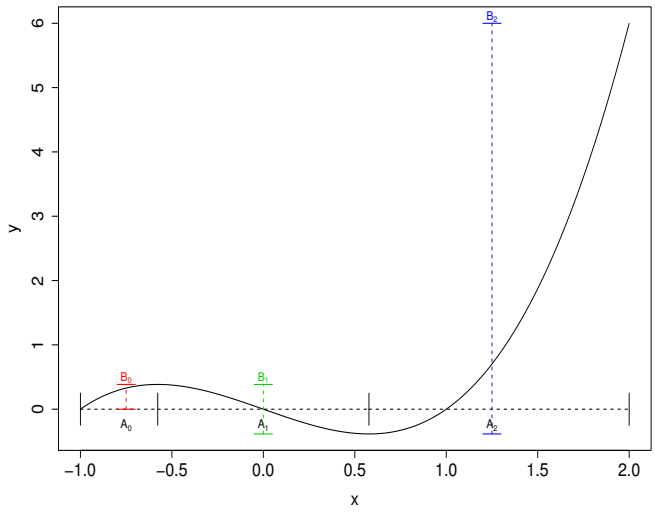


Figure 1: **Partition**: $G(x) = x^3 - x$, where $X \in (-1, 2)$. The dashed horizontal lines denote the $\{A_i\}$ partition of the x axis, while the dashed vertical lines denote the $\{B_i\}$ intervals on the y axis.

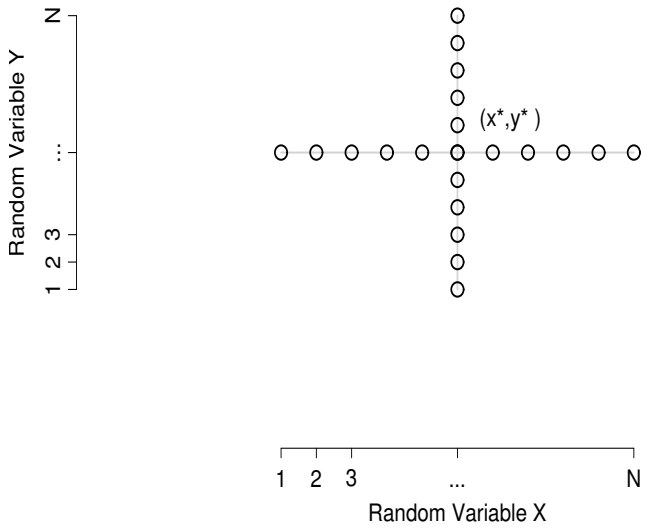


Figure 2: **2D discrete sample space**: The discrete uniform random variables in the 2D setting. The center point is at the coordinates (x^*, y^*).