Statistics and Its Interface Volume 3 (2010) 91–101

Using scientifically and statistically sufficient statistics in comparing image segmentations

Yueh-Yun Chi* and Keith E. Muller

Automatic computer segmentation in three dimensions creates opportunity to reduce the cost of three-dimensional treatment planning of radiotherapy for cancer treatment. Comparisons between human and computer accuracy in segmenting kidneys in CT scans generate distance values far larger in number than the number of CT scans. Such high dimension, low sample size (HDLSS) data present a grand challenge to statisticians: how do we find good estimates and make credible inference? We recommend discovering and using scientifically and statistically sufficient statistics as an additional strategy for overcoming the curse of dimensionality. First, we reduced the three-dimensional array of distances for each image comparison to a histogram to be modeled individually. Second, we used non-parametric kernel density estimation to explore distributional patterns and assess multi-modality. Third, a systematic exploratory search for parametric distributions and truncated variations led to choosing a Gaussian form as approximating the distribution of a cube root transformation of distance. Fourth, representing each histogram by an individually estimated distribution eliminated the HDLSS problem by reducing on average 26,000 distances per histogram to just 2 parameter estimates. In the fifth and final step we used classical statistical methods to demonstrate that the two human observers disagreed significantly less with each other than with the computer segmentation. Nevertheless, the size of all disagreements was clinically unimportant relative to the size of a kidney. The hierarchal modeling approach to object-oriented data created response variables deemed sufficient by both the scientists and statisticians. We believe the same strategy provides a useful addition to the imaging toolkit and will succeed with many other high throughput technologies in genetics, metabolomics and chemical analysis.

Keywords and phrases: Curse of dimensionality, Genomics, Metabolomics, Microarray.

1. INTRODUCTION

Segmentation partitions an image into several constituent components and aims to outline the anatomic structures and tumor-related objects. It provides an important means for clinical diagnosis and radiation therapy treatment planning. Segmenting one or more volume images, such as computerized tomographic (CT) and magnetic resonance images (MRI), help localize and display objects of interest, position the isocenters of the treatment beams, shape the radiation beams to conform to the outline of the target volume, and avoid nearby sensitive tissues. The process allows comparing competing treatment plans. Unfortunately, current segmentation practice is inherently expensive and requires slice-by-slice contouring tools and well-trained users to achieve acceptable results.

Automatic computer segmentation in three dimensions would greatly reduce the cost of three dimensional treatment planning for radiotherapy for cancer treatment. The kidney segmentation study, detailed in Section 2, compared the segmentations of two humans and one computer program over 12 CT images. After rigid alignment, surface comparisons generated distances between any pair of segmentations at roughly 20,000–30,000 surface points. Random variations between segmenters disallows determining correspondence of points and allows only overall comparisons. The statistical challenge arose from an average of 26,000 surface point distances for each of 24 kidney images from 12 people. An additional repeated measures dimension came from $\binom{3}{2} = 3$ pairwise differences among one computer and two human observers.

Such high dimension, low sample size (HDLSS) data dominate medical imaging, genetic, and metabolic research. The explosive increase in variable dimension has far outstripped the development of statistical methods designed for the task. The associated increase in cost has worsened the problem by pushing down the number of independent sampling units, and in return, raising the ratio of the number of variables to sample size. Classic statistical methods may be performed, but some believe the revolution in data collection has left us doomed to declare either too many false positives or too many false negatives.

If the data collection process allowed determining correspondence between surface points, traditional univariate statistical tests, such as t test, could have been applied individually to each point. The approach requires strong adjustment for the increased type I error (false positive) rate caused by simultaneous testing. The simplest, though conservative, adjustment is the Bonferroni correction of requiring a significance level of α/N, for N the number of points.

*Corresponding author.

With a large number of points, the resulting type I error rate for each test becomes so small that rejecting the null hypothesis of no group difference on one of the variables requires a very large underlying true difference. For instance, if the overall allowed type I error rate is 0.05, simultaneous testing on each of the 26,000 surface point distances would result in an individual type I error rate as small as 0.00000192 (0.05/26,000). Consequently, an increase in the frequency of false negative inferences occurs.

While some complex statistical inference problems can be treated in suitable asymptotic setups, different ratios of variable dimension to sample size demand different asymptotic results. Hall et al. (2005), and Ahn et al. (2007), by considering asymptotically increasing dimensionality with a fixed sample size, established geometric representations of the data under some regularity conditions. They showed that the geometric structure of HDLSS data becomes deterministic with the only randomness remaining in rotations of a simplex.

Presumably, principal component analysis (PCA) would allow dimension reduction. Asymptotic studies on HDLSS sample covariance matrices by Baik and Silverstein (2006) indicated that when the ratio of variable dimension to sample size goes to a constant (greater than 1), the sample eigenvalues behave as if the underlying covariance were an identity matrix. Furthermore, PCA with HDLSS data typically proves unreliable (Johnstone and Lu, 2009).

Meinshausen and Bühlmann (2006) found that consistent selection of the mean model with HDLSS Gaussian data required a sparse covariance matrix among predictors. Similarly, Bickel and Levina (2008) reviewed and evaluated strategies for HDLSS covariance estimation, with all depending on some form of sparseness. In practice, the requirement amounts to saying the existing methods only work for easy problems.

With that in mind, we advocate here that dimension reduction of HDLSS data should be oriented not only statistically but also scientifically. Rather than applying a routine procedure blindly, scientific questions and knowledge integrated with statistical reasoning can lead to a reliable solution and valid analysis. We will illustrate the point in the evaluation of medical imaging segmentation. We seek to provide a template for analyzing other HDLSS data from a similar scientific perspective. In the situation of interest, the scientists were clear that they did not care about pixel location, and only cared about overall performance. The approach takes advantage of the indifference to pixel location to simplify the problem.

We proposed an alternate path for analyzing the kidney segmentation data in two steps. In the first step, consultation with the scientists led to representing the information as objects, specifically histograms of distances, one per image and observer pair. For the histogram to avoid losing information requires the deviation of the computer-generated surface from the true surface to have no relationship to location on the surface. The scientists derived the computer segmentations from a sophisticated and accurate model of shape. Furthermore the human observers in the experiment of interest were very well-trained and conscientious, as required by appropriate standards of medical care. Hence the scientists were adamant in saying they only cared about the sizes of the deviations, and not the locations.

In the second step, parametric fitting of the distribution of differences between the shape model and human values for each object led to discovering a statistical characterization which requires only two parameters per histogram. The two parameters were compared and tested against the null hypothesis of no difference across rater pairs and the side of the body. The two stages strategically separated the task of estimation and inference. The framework can also be cast as a special case of hierarchical modeling.

The hierarchy was imposed to help overcome the curse of dimensionality. We assumed a common but latent distributional form for each of the 72 (24×3) histograms of distances. The distributional parameters were allowed to differ across histograms in order to allow for the possibility that they are affected by factors of study design or intrinsic features. A systematic search for the underlying distributional form was conducted over each individual object and evaluated by the overall empirical goodness of fit. The preferred model was chosen to satisfy both scientific and statistical criteria of goodness of fit and sufficiency. Each histogram was then summarized by its corresponding sufficient statistics for the final model. The primary comparison among histograms of thousands of distances reduced to the comparison among the resulting sufficient statistics. The reduction in dimensions allowed using classical statistical methods with known good properties.

The rest of the paper is organized as the following. In Section 2, we present the kidney segmentation data. In Section 3, we first examine the distributional characteristics of the distance data, paying special attention to the possibility of multi-modality. In Sections 4 and 5, we search for a suitable parametric form to fit the distributions. We evaluate the goodness of fit, and obtain sufficient statistics of the model with the best fit. In Section 6, we conduct a multivariate repeated measures ANOVA analysis on each sufficient statistic to draw inferences about the differences between automatic and manual segmentation. We conclude with a brief discussion in Section 7.

All computations were done with SAS software (SAS Institute, 1999). The KDE procedure was used for exploring the data with kernel density estimation. The UNIVARIATE procedure provided maximum likelihood fitting of the gamma and Gaussian parametric density models discussed in Sections 4 and 5. The DATA procedure was used to compute method of moment estimates and goodness of fit statistics for the truncated Gaussian model discussed in Section 5.

2. KIDNEY SEGMENTATION STUDY

The scientists sought to compare a computer program based on a medial model, called m-rep (Pizer et al., 2003),

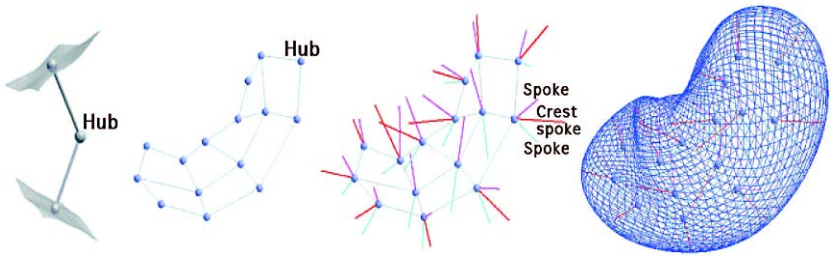

Figure 1. Frame 1: Medial atom with two equal-length spokes that define object width at the location of the atom. Frame 2: The medial sheet of a kidney represented as a 5×3 grid of medial atoms. Frame 3: Medial grid with spokes displayed. Frame 4: Wire-frame rendering of the surface implied by the medial sheet. Distance values are computed from the kidney surface in Frame 4 to the surface selected by human observers.

with experienced humans. In the study, a total of 12 treatment-planning CT images (24 kidneys) were used. Two trained humans (referred to as segmenter A and B, respectively) defined the target kidneys slice by slice on the original image data using interactive region filling together with pixel-painting editing tools for fine sculpting (Traction et al., 1994). The method forces the users to make pixel-level decisions at every location on the boundary. The work was performed without time constraints over multiple sessions scheduled at the convenience of the segmenters. The automatic segmentation (referred to as segmenter C) was provided by m-rep. Rao et al. (2005) detailed the structure and training of the m-rep model. By considering kidney parenchyma and pelvis as a single figure, a 5×3 grid of atoms was selected to best capture the full range of shape variability over the target population (Styner et al., 2003), as shown in Fig. 1. Frame 4 of Fig. 1 displays an example of the computer-generated surfaces which were compared to human-defined surfaces.

Surface comparisons between two human segmenters (A and B) and automatic m-rep (C) were performed using tools in Valmet (Gerig et al., 2001). After a rigid alignment, distances between a pair of segmentations were computed at roughly 20,000-30,000 surface points. The number and positions of surface points differed across pairs of segmentations. For each calculation, the disagreement was defined by the shortest distance between a point on the target surface to the nearest point on the reference surface. The measure was asymmetric due to the lack of point correspondence between the two compared surfaces. For instance, as illustrated in Fig. 2, the distance from a point on the target surface to the nearest point on the reference surface is not the same when measured in reverse.

In consultation with the image scientists, the size of the distances was of major interest as opposed to the spatial structure of the surface points. The three-dimensional array of distances for each image comparison was then reduced to a histogram for all subsequent analyses. With the asymmetry of all surface point distances, histograms were built twice between each pair of surfaces, with the role

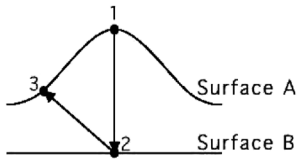

Figure 2. Illustration of the lack of symmetry when computing the minimum distance between two kidney surfaces in the study. The minimum distance to surface B from point 1 on surface A is defined by the line connecting points 1 and 2. However, the minimum distance to surface A from point 2 is defined by the line connecting points 2 and 3.

Table 1. Summary statistics for the six pairwise distance measures (in cm)

Pair	N	Mean	Std. Dev.	Min	Max
AB	3307	0.466	0.362	0.005	1.645
BA	2711	0.351	0.255	0.005	1.235
AC	4424	0.556	0.408	0.005	2.195
CA	3594	0.448	0.339	0.005	1.805
BC	3751	0.518	0.385	0.005	1.685
CB	3164	0.443	0.351	0.005	1.755

of target and reference exchanged. Given a set of 24 kidneys, 144 (24×6) histograms were generated, each with about 20-30,000 distances. The results presented in Rao et al. (2005) were based on pooled histograms by summing counts in individual distance bins, and on comparisons in mean and Hausdorff (maximum) distance separation. Using pooled histograms in quantifying the disagreement between a pair of segmentations may lose information embedded in the nature of the distance asymmetry, e.g. local curvature.

Table 1 gives summaries of the six pairwise distance measurements. The three pairs AB, AC and BC differed on average about 1 mm more than their respective counterparts, BA, CA and CB. In the following development, we focused on the comparison for a set of pre-specified pairs, namely AB, AC and BC. For each pair, the first segmenter corresponds to the target surface while the second segmenter

corresponds to the reference surface in the calculation of distances. Given the general lack of local curvature of a kidney, our imaging scientists were comfortable with a focused analysis on the selected pairs. The small size of the errors and discrepancies among alternate pairs relative to the size of a kidney also contributed to the comfort with working with a subset. The data included an average of 26,000 distances per image pair with 72 (24×3) pairs. The lack of correspondence in distance measures between image pairs disallows point-wise comparisons of any type of multiple comparison adjustments.

The study design of segmentation accuracy has a number of limitations which could be rectified in subsequent research. Segmentation of images of phantoms would provide an objective standard for quantifying the absolute accuracy of human and computer segmenters. Of course previous studies of human segmentation accuracy could be cited to defend the design. The inclusion of two or more human segmenters would increase the credibility of the estimates of human performance, and in turn, enhance the power of evaluating differences between human and computer segmentations. However, we ask the reader to set aside any concerns they may have about the design of the empirical study and focus on the data analysis approach we propose. We recommend the approach and leave the evaluation of the specific research to other venues.

Rao et al. (2005) compared segmentations in terms of two summary metrics, mean and Hausdorff distance. In contrast in the present paper we describe how to model the entire histogram by applying classical statistical methods of curve fitting to the family of histograms. Finding a common functional form of a density that allows adequately fitting each separate histogram by varying the parameter estimates allows representing the information in the entire curve by the parameter estimates, the sufficient statistics. Using the hierarchical process allows overcoming the high dimensionality of the data and providing accurate inference in a small sample.

The process we propose depends on a number of assumptions in order to succeed. Most importantly, the underlying distribution must be fully characterized by a relatively small number of parameters. Obviously the analysis process must include the appropriate distribution as a candidate for the process to succeed. Furthermore, the data must suffice for stable parameter estimation and to distinguish the appropriate distribution from all other plausible choices.

3. NONPARAMETRIC EXPLORATION OF HISTOGRAMS

We first explored the data graphically in order to gain insight about the distributions. We examined all 72 histograms in groups of 6, with the group including left and right kidney segmentations for each of three segmenter pairs. In an attempt to capture any dominant features embedded in thousands of distances with noise, we used nonparametric kernel density estimation with a variety of bandwidths to smooth each histogram. In examining the smoothed histograms of distances we focused especially on assessing the possibility of multiple modes.

Kernel density estimation is a nonparametric technique in which a known density function (the kernel) provides local weights for the observed data points to create a smooth approximation. A thorough review and discussion can be found in Silverman (1986). For the kidney data, we applied a Gaussian kernel and systematically varied the bandwidth in order to examine sets of smoothed histograms. We selected a bandwidth based on Silverman's rule of thumb and systematically applied a bandwidth multiplier of size five. The multiplier was used to strike a balance between under- and over-smoothing.

Figure 3 displays the 72 histograms and kernel density fits, with rows corresponding to patients and columns corresponding in sequence to AC/L, BC/L, AB/L, AC/R, BC/R, and AB/R comparisons. All kernel density estimates show unimodality and moderate to strong positive skewness of the histograms. Means ranged roughly from 0.12 cm to 0.21 cm, with standard deviations from 0.04 cm to 0.07 cm. For comparisons between the two human raters A and B, both left and right kidney histograms revealed prominent peaks near zero, reflecting the fact that most target points from rater A were close to the surface points from rater B. In contrast, distances between human and m-rep segmentations on average deviated more from zero, with an average 95th percentile of 0.55 cm, as opposed to an average of 0.34 cm between the two human segmentations. Both averages are considered clinically unimportant differences for radiotherapy treatment planning.

4. GAMMA DENSITY FITS

The apparent lack of a pattern of multi-modality in the kernel density estimates led us to consider fitting unimodal parametric distributions with positive support. The discussion of families of distributions in Johnson, Kotz, and Balakrishnan (1994, Section 12.4) provided a useful conceptual framework. The goal was to find a model with a reasonable fit, and then to reduce the data to the sufficient statistics of the model selected. We first considered a gamma model, which arises naturally in describing the distribution of a sum of squared independent random variables. With Gaussian variation of unit variance in each of three dimensions, the squared distance between two objects follows a chi-square distribution, a special case of a gamma distribution. The actual distance value would follow a chi distribution. A gamma density for a distance, $f_X(x) = \beta^{-\alpha}\Gamma(\alpha)x^{\alpha-1}e^{-x/\beta}$, depends on the shape ($\alpha > 0$) and the scale parameter ($\beta > 0$). Changes in α and β allow a wide range of shapes, including a monotone decreasing and unimodal form. Johnson, Kotz, and Balakrishnan (1994, chapter 17) provided a detailed discussion of the gamma distribution.

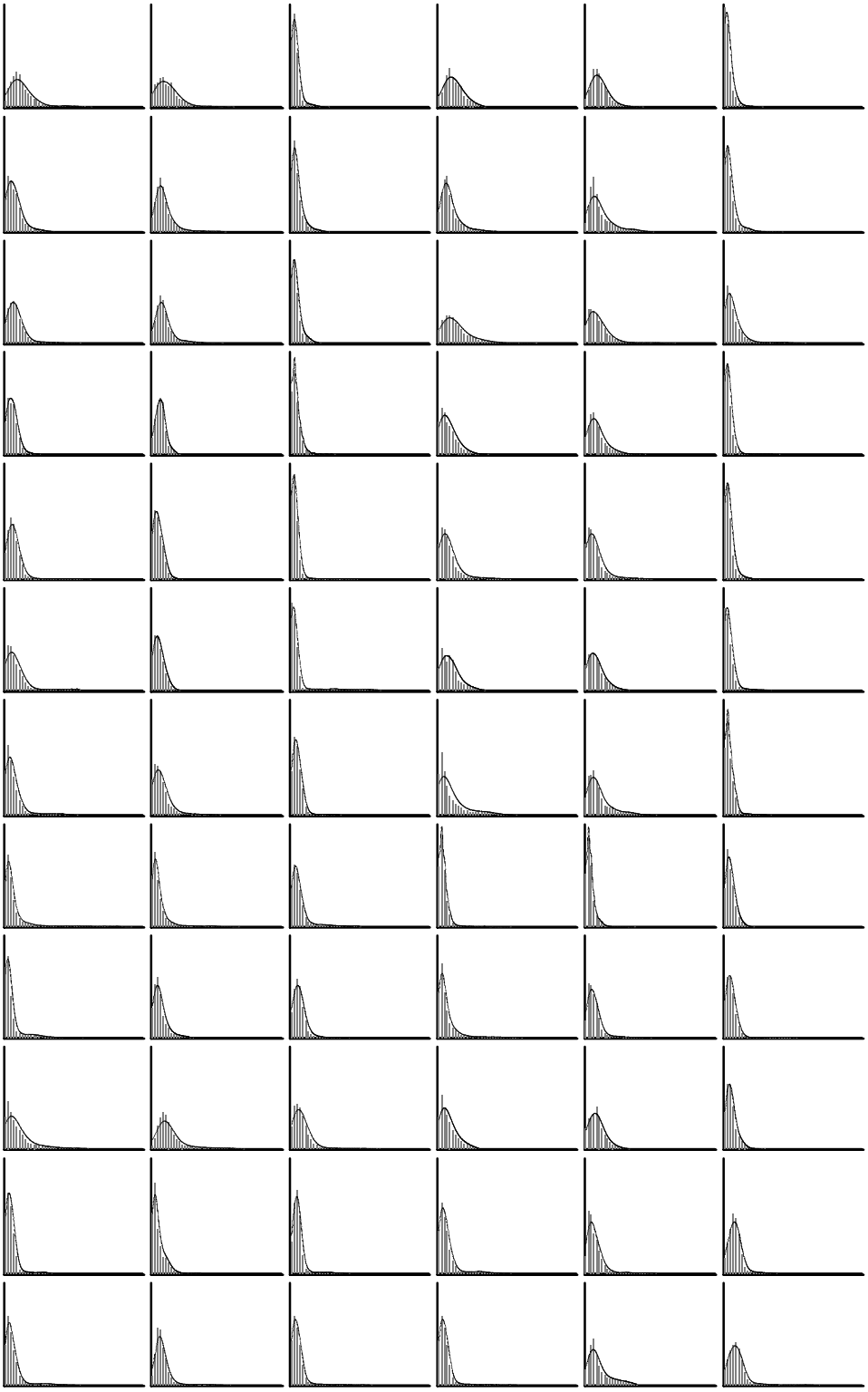

Figure 3. Kernel density fit for 12 CT images, 1 per row. Columns are (left to right) AC/Left, BC/Left, AB/Left, AC/Right, BC/Right, AB/Right. The horizontal axes all span 0–2.4 cm and the vertical axes all span 0–0.4.

We fitted the gamma and chi model separately for each of the 72 histograms of distances. We used the Kolmogorov-Smirnov D statistic to evaluate the goodness of fit of the gamma model. The statistic ranges from 0 (perfect fit) to 1 (worst possible fit) and equals the largest absolute discrepancy between the empirical and fitted distribution functions, namely $D = \sup_x |F_e(x) - F_f(x)|$. Figure 4 shows the fit of the gamma model, which was better than the fit of the chi model. When distance data piled up against zero (Left and Right AB pair), the fitted gamma model failed to closely capture the shape of the distribution. The small corresponding p-values for the Kolmogorov-Smirnov D statistic were not surprising given the large number of distances involved in the evaluation. Furthermore the p-values considered should be discounted to a certain extent to account for the multiple-testing issue. Consequently, the magnitudes of the D statistics themselves, as a scale-free measure of goodness of fit, should be the center of attention. The average Kolmogorov-Smirnov D statistic over 72 comparisons was 0.071 with standard deviation 0.024 for the gamma model and 0.323 with standard deviation 0.047 for the chi model (Table 2).

5. GAUSSIAN DENSITY FITS, TRUNCATED AND NOT TRUNCATED

A Gaussian model was our third choice. The positive nature of the distance data first led to the log-transformed Gaussian form, which is suitable for modeling unimodal and positively skewed distributions. We visually inspected the fit over the 72 histograms and computed the Kolmogorov-Smirnov statistic to assess the goodness of fit. A Gaussian model of distance was ruled out due to its lack of fit as compared to the gamma model.

We proceeded by conducting Box-Cox power transformations in the search of the form that least violates the Gaussian assumption. The cube root transformation of the distances outperformed other choices, a result consistent with the slight lack of fit from the log-transformed Gaussian model. We computed the cube root of all distances, and fitted both Gaussian and left truncated Gaussian models (Johnson, Kotz, and Balakrishnan, 1994, Chapter 13) to the transformed histograms. The Wilson-Hilferty (1931) approximation allows a chi-square random variable to be well-approximated by a function of a cube root Gaussian. The cumulative distribution function is approximated by

$$F_{\chi^2}(x;\nu) \approx \Phi\left\{\sqrt{\frac{9\nu}{2}}\left[\left(\frac{x}{\nu}\right)^{1/3} - 1 + \frac{2}{9\nu}\right]\right\},$$

with $\Phi(t)$ representing the standard Gaussian cumulative distribution function.

To conform with the positive support of the transformed distances, we truncated the Gaussian distribution at zero. Rather than permitting values from minus infinity to positive infinity, the left truncated Gaussian model sets up a left bound to the distribution. With $\phi(x)$ representing the standard Gaussian density function, $\{\mu, \sigma^2\}$ the mean and variance of the untruncated Gaussian, and $\delta = \mu/\sigma$, a Gaussian variable left truncated at zero has mean

$$\mathrm{E}(X) = \mu + \frac{\phi(-\delta)}{1 - \Phi(-\delta)}\sigma,$$

with corresponding variance

$$\mathcal{V}(X) = \left\{1 + \frac{-\delta\phi(-\delta)}{1 - \Phi(-\delta)} - \left[\frac{\phi(-\delta)}{1 - \Phi(-\delta)}\right]^2\right\}\sigma^2.$$

Parameter estimation can be conducted by the method of moments or by maximum likelihood.

As for the evaluation of the gamma model, the Kolmogorov-Smirnov D statistic was computed for each histogram to assess the goodness of fit. Figure 5 displays the Gaussian fits for the cube root transformed distances. The vertical axis for each graph ranges from 0 to 0.3, and each graph was based on roughly 26,000 distances. Most Gaussian curves fitted very well, with the remaining showing only a modest lack of fit. The Gaussian quantile plots in Fig. 6 further support the claim of adequency of fit with a consistent pattern of the poorest fit at the highest quantiles. The average of the Kolmogorov-Smirnov D statistics over all 72 comparisons was 0.065 with a standard deviation of 0.022 (Gamma gave a mean of 0.071 with standard deviation 0.024). In addition to examining discrepancies between fitted and observed curves via Kolmogorov-Smirnov D statistics, corresponding Gamma and Gaussian fits were compared visually by all members of the research group. A consensus favored the Gaussian fit as far better than good enough. Please note that the vertical axes differ in scale between Figs 4 and 5 (but not within each) in order to maximize resolution in each separately.

From Table 2, the goodness of fit for a left truncated Gaussian model, fitted by maximum likelihood, was very similar to the fit for a Gaussian model (average D of 0.064 and a standard deviation of 0.021 for truncated versus 0.065 and a standard deviation of 0.022 for not truncated). Given the similar fits and greater complexity of a truncated model, we selected the Gaussian as our final model. We proceeded

Table 2. Summary of the Kolmogorov-Smirnov D statistics of different parametric models

Model	N	Mean D	Std. Dev.	Min	Max
Gamma	72	0.071	0.024	0.028	0.183
Chi	72	0.323	0.047	0.203	0.399
Gaussian	72	0.065	0.022	0.025	0.161
Truncated Gaussian (Method of Moments)	72	0.072	0.029	0.027	0.176
Truncated Gaussian (MLE)	72	0.064	0.021	0.025	0.150

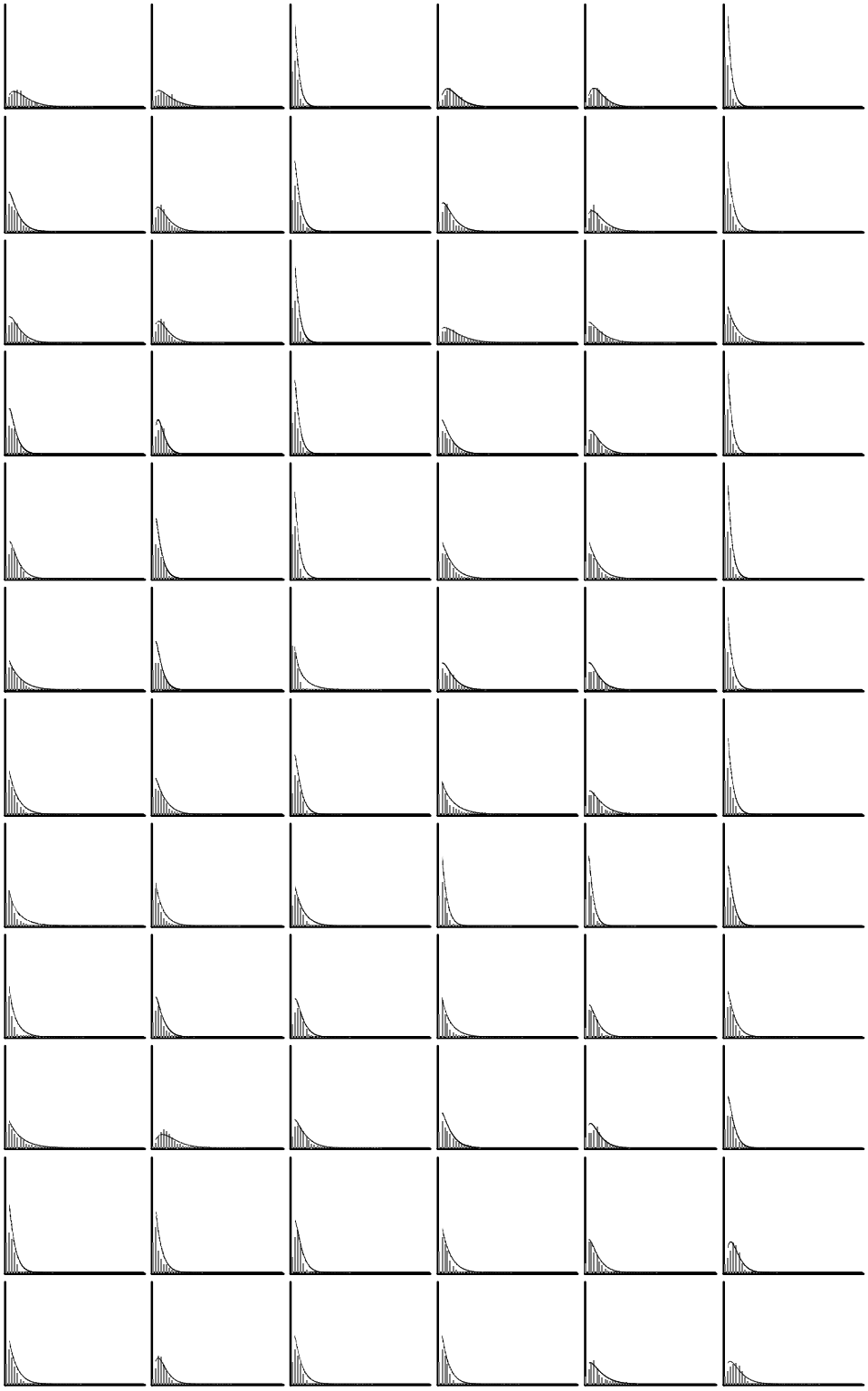

Figure 4. Gamma density fit for 12 CT images, 1 per row. Columns are (left to right) AC/Left, BC/Left, AB/Left, AC/Right, BC/Right, AB/Right. The horizontal axes all span 0–2.4 cm and the vertical axes span 0–0.8.

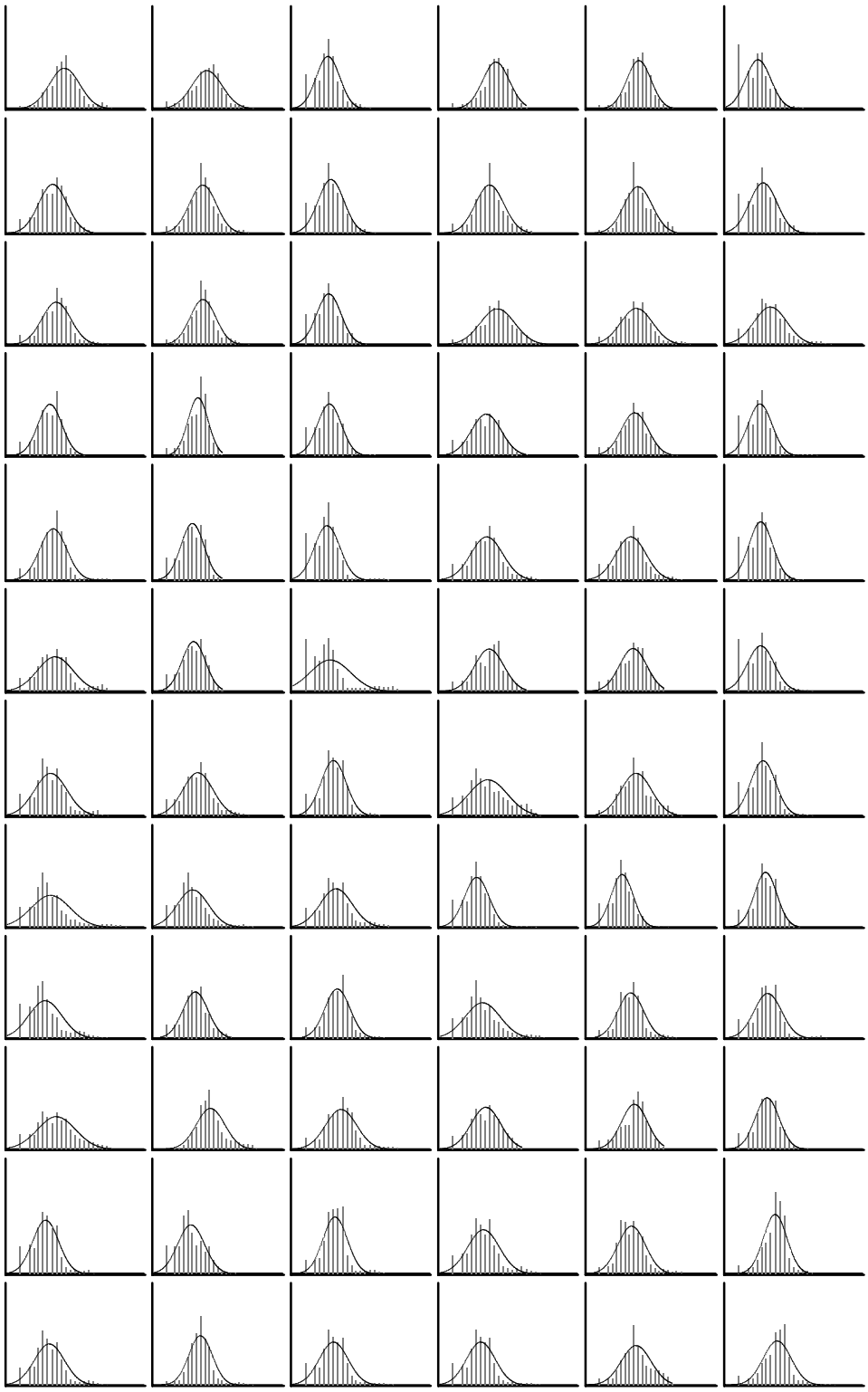

Figure 5. Gaussian fit on $\sqrt[3]{\text{distance}}$ for 12 images (rows). Columns are (left to right) AC/Left, BC/Left, AB/Left, AC/Right, BC/Right, AB/Right. The horizontal axes all span 0–$\sqrt[3]{3.375}$ (cm), and the vertical axes all span 0–0.3.

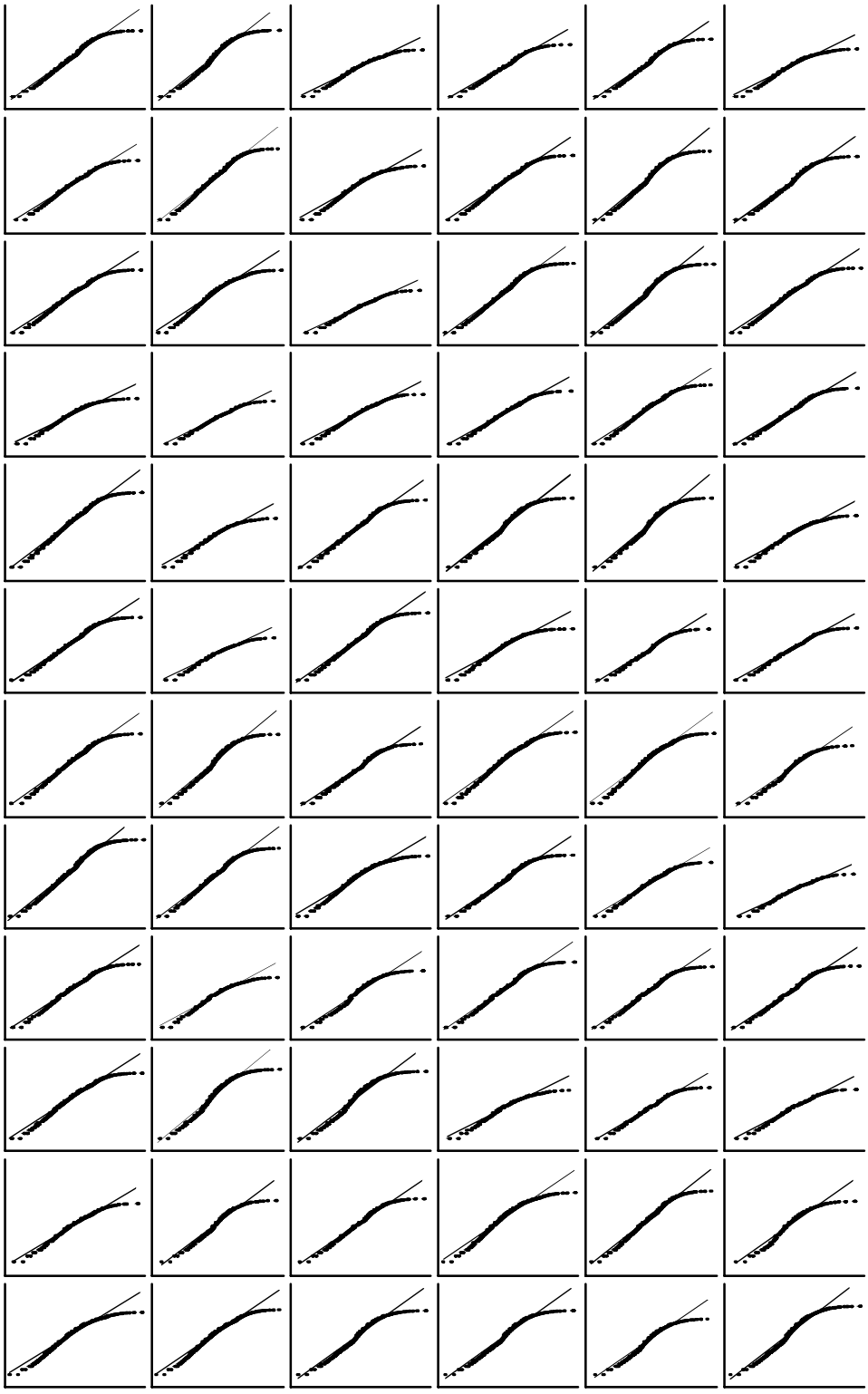

Figure 6. Quantile plots of Gaussian fit on $\sqrt[3]{\text{distance}}$ for 12 images (rows). Columns are (left to right) AC/Left, BC/Left, AB/Left, AC/Right, BC/Right, AB/Right. The horizontal axes are Gaussian quantiles spanning from -3 to 3 and the vertical axes span 0–$\sqrt[3]{3.375\ (cm)}$.

Table 3. Mean ± standard error of the three summary statistics of $\sqrt[3]{\text{distance (cm)}}$

	μ	$\mu+\sigma$	$\log(\sigma)$
AB	0.44 ± 0.01	0.59 ± 0.02	-1.93 ± 0.03
AC	0.51 ± 0.02	0.68 ± 0.02	-1.78 ± 0.02
BC	0.54 ± 0.02	0.69 ± 0.02	-1.88 ± 0.02

Table 4. Multivariate approach to repeated measures ANOVA test results (Hotelling-Lawley F test p-values)

Effect	Effect df	p-value		
		μ	$\mu+\sigma$	$\log(\sigma)$
1. Side×Pair interaction	2	0.4647	0.3121	0.2541
2. Side main effect	1	0.2830	0.4716	0.9029
3. Pair main effect	2	0.0033	0.0062	0.0009
AC vs. AB	1	0.0202	0.0030	0.0002
BC vs. AB	1	0.0014	0.0011	0.0799
BC vs. AC	1	0.0570	0.4530	0.0033

to further analyses by summarizing each histogram by its mean and standard deviation (the sufficient statistics for a Gaussian distribution) of cube root distance.

6. REPEATED MEASURES ANOVA

Obtaining a well-fitted parametric form allowed us to use classic statistical methods on the corresponding sufficient statistics. By choosing the Gaussian model for the cube root transformed distances, we effectively reduced the data dimension to its mean and standard deviation. In turn, three summary statistics were considered: mean, logarithm of the standard deviation, and the sum of mean and standard deviation. The logarithm transform of the standard deviation was motivated by the Gaussian assumption required in a repeated measures ANOVA analysis. Table 3 summarizes the three statistics for each segmenter pair.

The value of the mean plus one standard deviation ($\mu+\sigma$) has an elegant interpretation. It indicates the transition point at which the slope of a Gaussian cumulative distribution function stops rising and starts to decrease. The statistic has been used with success in studies of image processing in breast cancer detection (Pisano et al., 1997) because it possesses a simple interpretation as a threshold of certainty, from the perspective of psychophysics.

We conducted repeated measures ANOVA analyses on the three summary statistics to test the difference between human and automatic segmentation while accounting for the association between left and right kidneys and rater pairs. The resulting tests were equivalent to Wald tests from a mixed model, with a Kenward-Roger adjustment on degrees of freedom and all factors considered random. Muller and Stewart (2006; Ch. 3–5, 12, 17) gave general background while Edwards et al. (2008) discussed the specific equivalence. Appropriate residual analyses (Chapters 7, 10, and 11, Muller and Fetterman, 2002) were performed to verify the Gaussian assumption for repeated measures ANOVA. Given the validity of the data transformation, the three summary statistics would be expected to follow a Gaussian distribution based on being close to maximum likelihood estimates (MLEs) from thousands of values in each histogram.

Table 4 lists the result for repeated measures ANOVA using a nominal 0.01 type I error level to account for simultaneous tests on the three statistics. We considered the full model with the main effects of rater Pair and Side (Left, Right) of the comparison, and their interaction (the same saturated model was used implicitly for the random effects). Neither the Side×Pair interaction nor the Side main effect was deemed significant. The Pair main effect was significant for all three summary statistics. Stepdown comparisons were conducted to help evaluate Rater pair differences. Although significant differences in mean were found between pair BC and AB and between pair AC and AB, the mean difference in $\sqrt[3]{\text{distance (cm)}}$ was at most 0.1 $\text{cm}^{1/3}$. Given the size of a kidney and the nature of the radiation planning task, a disagreement of 0.001 cm was considered clinically negligible. As can be seen from results in Tables 3 and 4 similar patterns of results were found in comparing $\mu+\sigma$ between human and m-rep segmentations.

7. DISCUSSION

HDLSS data are becoming ever increasingly prevalent in many fields. Our successful experience in evaluating image segmentations highlights a strategy that can be used in other arenas to overcome the curse of dimensionality. The formation of distance histograms by ignoring the spatial structure was led by scientific objectives and repeated discussions with the scientists. Models that incorporated structured covariance between surface points might effectively improve the goodness of fit. In addition, accounting for the association between left and right organs of the same image might improve the fit.

Exhaustive model selection is not feasible given the infinite combinations of possible mean and covariance models. We achieved our results by considering a group of the most reasonable models for geometric distances, most belonging to an exponential family. Searching for a good enough model, let alone the "best" model requires a good understanding of the data in order to identify logical modeling options. As discussed at the end of Section 2, validity of the general process depends on a number of assumptions in order to succeed.

We emphasize that the transformation process employed in Section 5 was handled exactly like a Box-Cox residual analysis in the context of regression analysis. Hence two key features were maintained at all times. First, validity of the approach required that we control for possibly important predictors since the model assumed applies to deviations

conditional on image and rater pair. Second, no hypothesis tests or associated location estimate was computed until the transformation was chosen, and then an a priori analysis plan was followed. Although we were obviously aware of the previously published results for the same experiment, the radical changes in data reduction and analysis methods gave no promise of similar results.

Uncertainty about the validity and sensitivity of statistical analysis with HDLSS data, coupled with a lack of well-defined correspondence, drove our work. Point-wise comparisons of the human and computer-defined kidney surfaces create a substantial multiple-testing problem. Taking advantage of the statistical principle of sufficiency allowed successfully addressing the scientific questions with standard and powerful statistical tools. Although the approach taken here required traveling a longer road, the imaging scientists plan to continue using and promoting the method to others. Hence the work will continue to pay dividends in the future.

The approach described does not provide a universal solution, it simply gives a useful additional tool for the imaging toolbox. The method relies on an indifference to pixel location, an assumption that does not hold in other medical imaging settings, such as fMRI analysis. Such settings require different approaches that do account for pixel location.

Received 23 September 2009

REFERENCES

Ahn, J., Marron, J. S., Muller, K. E. and Chi, Y. Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766. MR2410023

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97** 1382–1408. MR2279680

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227. MR2387969

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F. and Schabenberger, O. (2008). An R-square statistic for fixed effects in the linear mixed model. *Statistics in Medicine* **27** 6137–6157.

Gerig, G., Jomier, M. and Chakos, M. (2001). *VALMET: A new validation tool for assessing and improving 3D object segmentation*. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 516–523.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B* **67** 427–444. MR2155347

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. **1**, 2nd ed. Wiley, New York. MR1299979

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal component analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics* **34** 1436–1462. MR2278363

Muller, K. E. and Fetterman, B. A. (2002). *Regression and ANOVA: An integrated approach using SAS software*. SAS Institute, Inc., Cary, NC. MR1991968

Muller, K. E. and Stewart, P. W. (2006). *Linear Model Theory; Univariate, Multivariate, and Mixed Models*. Wiley, New York, NY. MR2242366

Pisano, E. D., Chandramouli, J., Hemminger, B. M., Glueck, D. H., Johnston, R. E., Muller, K. E., Braeuning, M. P., Puff, D., Garrett, W. and Pizer, S. M. (1997). The effect of intensity windowing on the detection of simulated masses embedded in dense portions of digitized mammograms in a laboratory setting. *Journal of Digital Imaging* **10** 174–182.

Pizer, S. M., Fletcher, T., Fridman, Y., Fritsch, D. S., Gash, A. G., Glotzer, J. M., Joshi, S., Thall, A., Tracton, G., Yushkevich, P. and Chaney, E. L. (2003). Deformable m-reps for 3D medical image segmentation. *International Journal of Computer Vision* **55** 85–106.

Rao, M., Stough, J., Chi, Y. Y., Muller, K. E., Tracton, G. S., Pizer, S. M. and Chaney, E. L. (2005). Comparison of human and automatic segmentations of kidneys from CT images. *International Journal of Radiation Oncology, Biology and Physics* **61** 954–960.

SAS Institute (1999). *SAS/IML User's Guide*, Version 8. SAS Institute, Inc., Cary, NC.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. MR0848134

Styner, M., Gerig, G., Pizer, S. M. and Joshi, S. (2003). Automatic and robust computation of 3D medial models incorporating object variability. *International Journal of Computer Vision* **55** 107–22.

Tracton, G. S., Chaney, E. L., Rosenman, J. G. and Pizer, S. M. (1994). MASK: combining 2D and 3D segmentation methods to enhance functionality. *Mathematic Methods in Medical Imaging* **2299** 98–109.

Wilson, E. P. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Science* **17** 684–688.

Yueh-Yun Chi
Department of Epidemiology and Health Policy Research
University of Florida
1329 SW 16th Street, Room 5232, Gainesville
FL 32610, USA
E-mail address: ychi@biostat.ufl.edu

Keith E. Muller
Department of Epidemiology and Health Policy Research
University of Florida
1329 SW 16th Street, Room 5125, Gainesville
FL 32610, USA