Mixture Models for Estimating Maximum Blood Flow Velocity

Caren Marzban\textsuperscript{1,2,*}, Wenxiao Gu\textsuperscript{2}, Pierre D. Mourad\textsuperscript{3,4}

\textsuperscript{1} Applied Physics Laboratory, \textsuperscript{2} Department of Statistics
Univ. of Washington, Seattle, WA, 98195 USA

\textsuperscript{3} Department of Neurological Surgery, University of Washington, Seattle, WA 98195

\textsuperscript{4} Department of Bioengineering, University of Washington, Seattle, WA 98195

Running Title: Envelope Estimation

Category: Original Research

\*Corresponding Author: marzban@stat.washington.edu
Abstract

**Objective:** In a recent article a method was put forth for estimating the maximum blood flow velocity derived from transcranial Doppler ultrasound data. The method was based on a gaussian mixture model for the distribution of blood flow. Mixture models, in general, are advantageous in that they allow for a statistically rigorous measure of “maximum” blood flow velocity. But gaussian mixture models are restrictive in that the distribution of blood flow is assumed to be gaussian. The objective of the current work is to develop mixture models which do not invoke the gaussian assumption.

**Methods:** Gaussian mixture models have been shown to be useful for estimating maximum blood flow velocity. But the gaussian assumption is, in fact, not satisfied by the data. Here, two generalizations are examined: 1) skewed gaussian and 2) non-gaussian mixture models. Both models are assessed on 59 patients.

**Results:** It is found that skewed-gaussian mixture models provide better models of the data both in terms of the quality of the fit to the data, and in terms of the resulting estimate of maximum blood flow velocity. Non-gaussian mixture models are less “robust” in that their quality is not consistent across patients.

**Conclusions:** Skewed-gaussian mixture models are superior to gaussian mixture models both in terms of providing a better fit to the data and for estimating maximum blood flow velocity, for all 59 patients examined in this study. Non-gaussian mixture models demonstrate a potential for even higher-quality assessment of blood flow, but further development is called for.

Keywords: blood flow, brain, head injury, noninvasive, transcranial Doppler
1 Introduction

The importance of estimating blood flow in major cerebral arteries has been well documented.\textsuperscript{1−5} The reduction of blood flow through a major cerebral artery - cerebral vasospasm - can lead to a wide range of disorders, and therefore, persistently monitoring blood flow has important clinical consequences\textsuperscript{6}. Transcranial Doppler (TCD) is one of the methods for assaying blood flow.\textsuperscript{7,8} Vasospasm leads to reduced blood flow but increased blood flow velocity (FV), and therefore, estimating the maximum blood flow velocity is of particular importance.\textsuperscript{9}

The time series of the histogram of FV is called a spectrogram, and the time series of maximum FV is called an envelope. Many envelope estimation algorithms have been proposed\textsuperscript{10−15}; but the one currently in use in many TCD devices is the Modified Geometric Method.\textsuperscript{10} That method essentially searches for a “kink” in the cumulative histogram of FV, and defines the maximum FV as the FV at which the kink occurs. It provides reliable envelope estimates, but lacks the important feature of quantifying what is meant by maximum FV. By contrast, the mixture model described in Marzban et al.\textsuperscript{15} allows one to quantify FV in terms of percentiles. For example, one may examine the envelope corresponding to the 90\textsuperscript{th}, 95\textsuperscript{th}, 99\textsuperscript{th} percentiles of FV. Each of these provides a unique and well-defined measure of “maximum” FV, and the ability to do is the main advantage of mixture models. Furthermore, it is reassuring that the “maximum” FV as defined by the Modified Geometric Method is in fact in close agreement with that produced by the 95\textsuperscript{th} percentile envelope in a mixture model.\textsuperscript{15} As such, mixture models enjoy two important features: 1) They allow for producing envelopes corresponding to multiple percentiles of FV, and 2) one of the percentiles (i.e., the
95th) produces an envelope which is approximately equal to envelope based on the Modified Geometric Method.

However, the specific type of mixture model developed by Marzban et al.\textsuperscript{15} assumes that the distribution of FV, at any point in time, as measured by a transcranial ultrasound device, is gaussian. The agreement between an envelope produced by such a Gaussian Mixture Model (GMM) and that produced by the Modified Geometric Method suggests that either the distribution of FV truly is gaussian, or that the envelope is not sensitive to violations of that assumption. As shown below, the distribution of FV varies with time; at certain times it is gaussian (or at least, near-gaussian), while at other times it is not. The deviations from normality manifest themselves mostly as a skew. For this reason, one of the models described here assumes that the underlying distributions of FV are skewed gaussian. Data are employed to estimate the mean, the variance, and the shape parameters of the distribution. Here, this model is referred to as a Skewed-Gaussian Mixture Model (Skewed-GMM).\textsuperscript{16–18} Another model discussed (briefly) does not assume a parametric form for the distribution of FV at all; instead the distribution is inferred using kernel estimates.\textsuperscript{19–21} Such Kernel Mixture Models (KMM) are flexible in that they accommodate a wide range of distributions (including gaussian). However, kernel methods involve a \textit{smoothing parameter} whose value is \textit{a priori} unknown. Generally, extreme values of the parameter correspond to either a smooth distribution (e.g. a gaussian), or a highly irregular distribution. One often employs some criterion (such as the maximization of likelihood) to optimize the smoothing parameter.

In this paper, the Skewed-GMM and KMM are developed for estimating the envelope
for each of 59 patients. In previous work it was shown the GMM is generally superior to the Modified Geometric Method. Here, it is shown that Skewed-GMM leads to envelopes that are generally superior to those produced by the GMM, for the majority of the 59 patients. For the remaining patients the envelopes are comparable (i.e., no worse) than those from GMM. The KMM leads to envelopes that are moderately superior to those based on Skewed-GMM, but not for all patients; for this reason, the KMM is discussed only briefly, and only for the purpose of discussing future work.

2 Materials and Method

2.1 Data

The data for this work were collected from a pre-clinical study involving multiple hospitals in the USA. Details of this data set can be found in Marzban et al.\textsuperscript{22} and Marzban et al.\textsuperscript{15}. In the former, this data was used to develop a model of arterial blood pressure, which can in turn be used to predict intracranial pressure. The latter employed only the ultrasound-derived Transcranial Doppler spectra for the purpose of developing the aforementioned gaussian mixture models for envelope estimation. In accordance with the institutional review board for each hospital, informed consent was obtained from all patients or their families.

Using clinically approved TCD units blood flow velocity (FV) in the middle cerebral artery was measured with Doppler ultrasound. Data were collected for periods of time from 5 to 30 minutes. All retrospective data processing and analysis was conducted at the Applied
Physics Laboratory (APL), University of Washington. Although the Doppler spectral time series were initially sampled at 125 Hz, they were down-sampled at 40 Hz. This resolution was sufficient to resolve each patient's systolic rise, diachrotic notch, and diastolic minimum. For each patient, envelopes were produced over a fixed duration (118 time steps, or about 3 minutes). This duration is sufficiently long to include several cardiac cycles, while allowing details of the envelopes to be visually evident.

2.2 Data Processing and Statistical Analysis

The Modified Geometric Method is commonly employed, and for this reason it will be presented here for the purpose of providing a comparison. In fact, since the GMM outperforms the Modified Geometric Method\textsuperscript{15}, the Skewed-GMM and the KMM are all compared to the GMM.

In a GMM, the distribution (more accurately, the probability density function) of FV, for each patient, and at any time, is assumed to be a linear combination of two gaussians. The coefficients - called mixing coefficients - as well as the parameters of the gaussians are estimated from data. In Marzban et al.\textsuperscript{15} it was shown that the two gaussians represent the “signal” and the “background” portions of the distribution, respectively. This, then, allows one to use the upper percentiles of the former to quantify maximum FV. The left panel in Figure 1 shows an instance of the signal gaussian (blue), the background gaussian (red), the 90\textsuperscript{th}, 95\textsuperscript{th}, 99\textsuperscript{th} percentiles of the former (blue vertical lines), and the maximum FV according to the Modified Geometric Method (black vertical line). The data (black circles) are from
one patient and at a given time.

The generalization from GMM to Skewed-GMM is relatively straightforward in that each gaussian in the mixture model is replaced by a skewed gaussian. This introduces only one more parameter, namely the shape-parameter which effects the skew of the gaussian. The middle panel in Figure 1 shows the various ingredients. The data are for the same patient and the same time as in the left panel. Note the evident skew in the signal and the background components. The Skewed-GMM algorithm is implemented by the R package called mixsmsn, available online from the Comprehensive R Archive Network (CRAN).

In KMM the distribution of FV is assumed to be a linear combination of two distributions, but unlike GMM and Skewed-GMM, each distribution is estimated using kernel methods. The kernel method employed here uses the Expectation Maximization algorithm for fitting multivariate non-parametric mixtures with completely unspecified component densities, except for a conditional independence assumption described by Benaglia et al. The kernel is the standard normal density function. The algorithm is implemented in an R package (R Development Core Team, 2007) called mixtools available at CRAN.

The right panel in Figure 1 shows the resulting distributions. Note that the distributions are now more irregular, capturing the non-smooth structure evident in the data (black circles). The Appendix explains the manner in which the smoothing parameter is set.
3 Results

The spectrogram and the envelopes for GMM, Skewed-GMM, and KMM for one patient are shown in Figure 2. (The corresponding patient is different than that for Figure 1; the reason different patients are used for the different figures is to illustrate different characteristics of the various methods.) For this particular patient, it can be seen that the three percentiles according to the GMM (left panel) are close to one another, and generally under-estimate the FVs. By contrast, the three envelopes based on the Skewed-GMM are more widely spread, and more consistent with the spectrogram (middle panel). The KMM’s envelopes (right panel) have similarities and differences with the GMM and the Skewed-GMM. For example, they under-estimate the FVs (like GMM), but are relatively spread out (like Skewed-GMM). These patterns are generally true across many patients. Also, note the complete disagreement between the spectrogram and the 99th percentile envelope at the time just prior to 20. This type of failure is characteristic of the KMM, and it is discussed further in the Discussion section.

Another noteworthy feature in Figure 2 is that GMM under-estimates the highest FV values in the spectrogram. This is evident in the manner the envelope is “below” the highest FVs in the spectrogram but only when FVs are highest. The Skewed-GMM solves that problem. This is a consequence of the fact that low FV values are consistent with gaussian distributions, but for high FV values the distributions are more skewed.

The time-dependence of the shape of the distributions can be confirmed by examining the values of the shape-parameter as a function of time. Figure 3 shows the time series for the
estimated shape parameter for the signal (left panel) and for the background (right panel). It can be seen that for the majority of the time series, the values of the shape parameters are approximately zero, i.e., the distribution of the FV is gaussian. However, there are also times at which the shape parameters are unambiguously nonzero. The times correspond to regions of the spectrogram where FVs take their highest values. Note that the shape parameters are nonzero, they are positive for the signal, and negative for the background components of the mixture model. (A positive value of the shape parameter corresponds to a right-skewed gaussian.)

As seen for the patient corresponding to Figure 3, shape parameter are mostly zero. But that is not true for all patients; Figure 4 shows the histogram of the estimated shape parameters for a different patient. The left mode corresponds to the background, while the right mode is for the signal. It follows that the background component is consistently left-skewed (with shape parameters in the -30 range), and the signal component is either gaussian (corresponding to the peak at zero) or right-skewed (with shape parameters between 0 and +30).

Figures such as those in Figure 2 are useful for assessing the quality of an envelope. But they are impractical for addressing the quality of the envelopes for all 59 patients. Consequently, although we have performed this visual inspection for all patients, the results are not shown here. Instead, one can rely on the goodness-of-fit measure itself - not the envelope - to compare the various models developed here. There are numerous quantities that can be used for measuring how well a model fits the data. One commonly-employed measure is the Log-Likelihood (LL). Given that the fitting of the model is performed at each
time, there exists a time-series for LL as well (not shown). For the purpose of comparing two models, however, it is sufficient to examine the difference between LL values of the two models. Therefore, for each patient the histogram, or boxplot, of this quantity provides a visual assessment of the relative performance of the models. If two models are comparable in terms of goodness-of-fit, then the boxplot will be centered on the number zero. Otherwise the two models can be deemed different.

Figure 5 shows the boxplot of LL(Skewed-GMM) - LL(GMM) for all 59 patients. (Recall that the variability of these boxplots is due to time.) The fact that the majority of boxplots are “above” the horizontal line at zero implies that for the majority of the patients the LL of the skewed-GMM is generally higher than that of the GMM across time. In other words, on the average (across time) the Skewed-GMM is a better model of the data than GMM. A few exceptions are patients 19, 41, and 53 for whom the boxplots are mostly centered around zero, and as such there is no significant difference between GMM and Skewed-GMM. To assess whether Skewed-GMM has higher LL values than GMM, a 1-sided, t-test is performed. All of the p-values are found to be less than 0.05, except for those of patients 19, 41, and 53 (shaded boxplots in Figure 5). (The assumption of normality for the t-test is confirmed by a visual inspection of the normal qq-plot.)

The comparison between KMM and GMM can be done in the same fashion. The boxplots for the difference between their LL is shown in Figure 6. Given the systematic presence of the boxplots above the horizontal line at LL = 0, it is evident that KMM has a generally higher LL than GMM. (All of the p-values are below 0.05.) It is important to emphasize that this conclusion is not unexpected. After all, KMM allows for more flexible distributions,
and consequently, leads to higher LL values than GMM. The more important question is if KMM also leads to better envelopes than GMM; and the answer to that is in the negative. A visual examination of the KMM envelopes for all 59 patients suggests that the resulting envelopes are no better than those of the GMM, and are often comparable to those produced by Skewed-GMM. KMM is also not as consistent (across time) as GMM or Skewed-GMM, as seen by the “peak” in the right panel of Figure 2. Therefore, whereas there is sufficient evidence to advocate the use the of Skewed-GMM, that is not the case for KMM. Further discussion of the KMM is provided in the next section.

4 Discussion

In an earlier study it was shown that GMM has numerous advantages over the Modified Geometric Method for estimating the envelope. Here, it has been shown that a Skewed-GMM has the same qualitative advantages as GMM (e.g., allowing for a percentile-based notion of maximum FV), but also outperforms it in terms of both the quality of the envelope (assessed through a visual comparison of the spectrogram and the envelopes) and a quantitative comparison (in terms of the goodness-of-fit).

One may wonder if the higher values of goodness-of-fit for Skewed-GMM, as compared with that of GMM) may be a consequence of overfitting. This concern is unwarranted, because the Skewed-GMM has only two additional parameters - the shape parameters for the signal and the background components. As such, the total number of independent parameters to be estimated is only seven (mean, standard deviation, and shape parameters
for each of the signal and background components, plus one for the mixing coefficient). The data employed for estimating these parameters number in the hundreds, and therefore, overfitting is not a concern.

An issue that arises in the mixture-model approach to envelope estimation is the identifiability of the two components with signal and background. More specifically, a mixture model with two components will identify two components in the distribution of FV values; however, an additional criterion must be introduced to decide which of the two components should be identified as the signal. This is important, because it is the percentiles of the signal component which lead to the envelope. Visually, the signal component can be identified as the one on the left, but quantifying what is meant by “on the left” is nontrivial. The approach adopted here for identifying the signal and background components of the mixture model is outlined in the Appendix. That proposal is somewhat ad hoc, and so it would be desirable to develop a more statistically sound method for identifying the signal component of mixture models.

As described above, the smoothing parameter of the KMM determines the shape of the kernel estimate of the probability density functions. Here, an algorithmic criterion is used to set the values of that parameter, but the criterion optimizes the goodness-of-fit, i.e. LL. As such, the criterion does not automatically lead to improved envelopes. An instance of this is demonstrated in Figure 2 where the KMM envelope displays an abrupt “peak” (at time \( \sim 15 \)), clearly inconsistent with the observed spectrogram. This is one of the main reasons why KMM is not whole-heartedly advocated here. Another reason is more computational: The procedure for estimating kernels is rather computer intensive, and therefore slow. In
fact, for the current study, KMM is developed on data where the lowest 5% of the FV values have been trimmed (deleted). This significantly increases computational speed, but introduces an *ad hoc* parameter (i.e., the 5%). In short, more work is required before one can benefit from the flexibility of KMM as a useful envelope estimation method.

5 Appendix

Kernel methods generally involve a smoothing parameter (often called *bandwidth*) whose value must be specified. For the current study, its value is determined as follows: The default choice of the bandwidth $h$ minimizes a quantity called Asymptotic Mean Integrated Squared Error\textsuperscript{24} and is given by

$$h = 0.9n^{-1/5}\min\{SD, \frac{IQR}{1.34}\}$$

(1)

where $SD$ and $IQR$ are the standard deviation and interquartile range of all $n$ observations, respectively. It has been shown that this estimate is somewhat smaller than the true optimal value\textsuperscript{20}, and so, the value chosen here is taken to be larger than the default value, which in turn leads to smoother fits. As a result, at each time, and for each patient, $h = 10$ is used, unless Silverman’s rule gives a larger value, in which case the larger value is selected. The fits are visually inspected to assure that the chosen value of the bandwidth does not lead to overfitting.

As described above, the mixture model identifies two components in the distribution of FV values. But the task of identifying which of the two components is signal (and which is background), is complex. Here, the following three criteria are employed to derive an
algorithm for the identification task: The signal component is more likely the one with 1) lower mean, 2) higher peak, and 3) larger standard deviation. Denote the means, peaks, and standard deviations of the two components as \((\mu_1, \mu_2), (p_1, p_2), (s_1, s_2)\), respectively. Define the ratios \(a = \frac{\mu_1}{\mu_2}, \ b = \frac{p_1}{p_2}, \ c = \frac{s_1}{s_2}\). As such, the signal component is more likely to be the component labeled “1” when \(a\) is small, and \(b\) and \(c\) are large. Therefore, if \(x = \frac{1}{a} \times b \times c\) is larger than 1, then the component labeled “1” should be identified as the signal. Similarly, \(x < 1\) indicates that the “1” component should be identified as the background.

Acknowledgments

This work has received support from National Institutes of Health (Grant R43NS46824-01A1); National Space Biomedical Research Institute (Grant SMS00701-2009-513); and PhysioSonics Incorporated. P. D. Mourad has a financial interest in PhysioSonics. The authors are grateful to P. R. Illian and D. Morison for contributions made during an early phase of this project.
References


Figure Legends

Figure 1. The distribution (more accurately, the probability density function) for the signal (in blue) and the background (in red), as determined by mixture models - GMM (left panel), Skewed-GMM (middle), and KMM (right). The blue vertical lines denote the 90th, 95th, 99th percentiles of the signal FV, and the black vertical line marks the maximum FV according to the Modified Geometric Method.

Figure 2. For one patient, the spectrogram (colored background), and the 90th, 95th, 99th percentile envelopes (in black) according to GMM (left), Skewed-GMM (middle), and KMM (right).

Figure 3. For one patient, the time series of the estimated shape parameter for the signal (left) and the background (right) distribution.

Figure 4. For one patient, the histogram of estimated shape parameter. The two modes correspond to the signal (right) and background (left) components of the mixture model.

Figure 5. For all 59 patients, the boxplot of LL for Skewed-GMM minus LL for GMM. The shaded boxplots indicate that the corresponding p-value is greater than 0.05.

Figure 6. same as Figure 5 but for LL(KMM) - LL(GMM).
Figure 1. The distribution (more accurately, the probability density function) for the signal (in blue) and the background (in red), as determined by mixture models - GMM (left panel), Skewed-GMM (middle), and KMM (right). The blue vertical lines denote the 90th, 95th, 99th percentiles of the signal FV, and the black vertical line marks the maximum FV according to the Modified Geometric Method.
Figure 2. For one patient, the spectrogram (colored background), and the 90\textsuperscript{th}, 95\textsuperscript{th}, 99\textsuperscript{th} percentile envelopes (in black) according to GMM (left), Skewed-GMM (middle), and KMM (right).
Figure 3. For one patient, the time series of the estimated shape parameter for the signal (left) and the background (right) distribution.
Figure 4. For one patient, the histogram of estimated shape parameter. The two modes correspond to the signal (right) and background (left) components of the mixture model.
Figure 5. For all 59 patients, the boxplot of LL for Skewed-GMM minus LL for GMM. The shaded boxplots indicate that the corresponding p-value is greater than 0.05.
Figure 6. same as Figure 5 but for LL(KMM) - LL(GMM).