University of Alberta

Methods for Automatic Heart Sound Identification

by

Michael Joya

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2012
Edmonton, Alberta

For my grandmother, Cora

Abstract

This thesis provides a description of the cardiac rhythm as a latent chain of heart sound arrivals which occur over time, where each arrival generates a fixed window of observable data that can be described with arbitrary feature functions. This description of the process produces tractable procedures for inference of timing parameters and estimation of the most likely chain of arrivals. It is shown that the central obstacle for accurate estimation is that the timing of the arrivals for a particular subject will often differ substantially from those of the pooled sample, often resulting in poor estimates. One of the theoretical contributions of this work is a method for estimating the unique timing parameters of the rhythm through the use of signal filtration applied directly to the observed data. This technique is effective at modeling the distribution of these parameters for recordings with repetitious patterns in the signal.

Preface

The overall goal of my research is to find effective ways of modeling long-term history, or rather to model processes which exhibit a strong *memory*.

In this investigation I take both theoretical and constructive approaches: a model is described, it is shown to exhibit certain essential properties, then it is run on real data and comparisons are made. My goal is to extend this path toward more theoretical justifications in the future, but for now I have included a few derivations that are part of learning a model that can represent correlations over time.

The bigger question of, "how does one model the *current state* of a history-dependent process?" is narrowed to, "how does one model the interval durations between distinct events in the process?". The answer is partially provided by *point process models* in which the intervals are represented by independent positive random variables. These questions typically turn up when a sound recording captures a sparse set of events that are either well-separated in the time domain or sufficiently easy to decorrelate from each other. This work's primary theoretical contribution is a method of describing the influence of a point process on a continuous signal, such as one recorded and sampled by a microphone or stethoscope.

The identification of heart sounds also demands such an answer. In this task, the raw signal sounds produced by heartbeats vary substantially within the population and are only so reliable at identifying the beats. An accurate sense of the overall rhythm is sometimes the most predictive element that can be used in the prediction of the individual beats. However, the two are inextricably linked - the overall rhythm is dependent on the positions of the beats, and the positions of the beats is easiest to estimate when one knows the overall rhythm. The question of how to model this rhythm and to capture its unique characteristics for individual subjects is thus the primary research question of the current study.

I learned quite a bit about research, about memory, and about living while writing this thesis. Only a small amount of that managed to make its way into the document itself. The rest I have reiterated endlessly to my mother,

father, sister, aunt, friends, and my two supervisors. If the key to memory is repetition then I have acquired a full ring.

Finally, I want to give my deepest gratitudes to Dr. Ian Adatia and to Stollery Children's Hospital both for the data used in this study and for hours of invaluable consultation given during the course of its development.

Acknowledgements

The author would like to thank Dr. Ian Adatia and the staff at Stollery Children's Hospital for contributing the sample of heart sounds used in this study.

Contents

List of Tables

List of Figures

List of Symbols

$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Real non-negative
$\mathbb{R}_{++}$	Real positive
$\mathbb{N}$	Natural numbers
$f(t)$	function over a continuous space
$f[t]$	function over a discrete space
$\lvert f(t)\rvert^2, \lvert f[t]\rvert^2$	signal energy
$h[n]$	discrete filter response
$\Phi[t, f]$	time-frequency representation function
$\mathbb{I}_{\{\text{expr}\}}$	Indicator function
ν	scaling threshold
$k_\sigma(\cdot, \cdot)$	Gaussian kernel comparison function with scale parameter σ
$\mathcal{P}$	probability distribution
$L(\theta_{ab})$	likelihood function of parameter θ_{ab}
$\mathcal{L}(\theta_{ab})$	log-likelihood function of parameter θ_{ab}
ω, Ω	sample, sample space
E	label set
$Q, \hat{Q}$	arrival sequence, predicted
L	length of a sound recording
q	arrival
$y_q, Y_q, \mathbf{Y}, \mathbf{Y}_{\boldsymbol{\mathcal{T}}_q}$	observed value, as r.v., as a set of r.v., as a set of r.v. influenced by q
$\overline{\mathbf{Y}}_{\boldsymbol{\mathcal{T}}}$	observed r.v. that are not influenced by state
$\overline{\mathbf{Y}}(q, q', w)$	observed r.v. between arrivals that are not influenced by state
$x_q, X_q, \mathbf{X}$	label of arrival, as r.v., as a set of r.v.
$t_q, T_q, \mathbf{T}$	time of arrival, as r.v., as a set of r.v.
z_q, Z_q	waiting time following event q, as r.v.
w	window radius
I	discrete-time indices captured by sampling
E	mark (label) set
M	number of frequency features for a given time step

Chapter 1

Introduction

This work concerns the problem of automatically identifying and labeling heart sounds in recordings collected by stethoscope from the four primary chest listening sites used in cardiac auscultation. The motivation for this problem comes from pediatric cardiology, where diagnosis of certain pathologies is more difficult for infant patients e.g. because of the difficulty of having infants sit motionless during cardiac MRI scans.

There are at two long-term goals in the larger scope of this project. The first concerns the identification of a gap width between the two valve closures that make up the second heartbeat, which normally beat synchronously and with little or no audible delay between the two. This gap width, when present, is often the consequence of a late pulmonary valve closure due to high blood pressure in the pulmonary artery [20, 31]. The second concerns the probabilistic modeling of the heart sound process in order to support general cardiac auscultation for pathologies that are known to produce anomalous sounds. In support of the long-term goal of providing machine-aided diagnosis tools for auscultation, this thesis offers a method for modeling the heart sound and for identifying the times of heartbeats within a stethoscope recording sequence.

The main contributions of this thesis are threefold.

- First, a method is provided for normalizing a signal or function so that its local maxima become bounded. This method is used to remove the large intensities that arise in certain feature map representations.
- Second, an approach is described for learning and inference with marked point processes and feature representations of observed data. This method produces an estimate of the chain of heartbeats and their arrival times.
- Third, a filtration approach is presented which can find reliable parameters describing a patient's unique heart sound rhythms without having to know the exact positions of the beats in a recording.

In addition, this method offers a means by which several subject recordings can be combined to produce a more reliable estimate of the distributions associated with the valve closure timings.

In the experimental portion of the thesis, two empirical evaluations are given. The first is a statistical analysis that examines the suitability of various distributional models for the intervals that lie between the beats. The second is an empirical evaluation of the set of predictors that are developed using the methods herein. The predictor variants are evaluated using measure(s) referred to as the *continuous precision-recall* scores. These measures provide the analogous precision-recall score for classification tasks in which a machine-labeled region or segment may exhibit partial overlap with the region containing the true label. In this case, the classifier's performance on continuous precision and recall will vary in a way that is analogous to standard precision and recall curves, and the results can be compared while varying some parameter of interest.

The experimental section and overall thesis then concludes by summarizing the main points learned about heart sound analysis with respect to the algorithms and methods presented, including some thoughts on future direction in automatic diagnosis.

Chapter 2

Background

The majority of heart sound research might be partitioned roughly into two main categories. One category, *heart sound identification* concerns the task of identifying locations or regions of a sound signal that are associated with distinct cardiac events such as the onset and duration of the first or second heart sound (hereafter abbreviated by S1 and S2), murmurs, the diastolic interval, or the clicks and pops associated with valvular defects. The second category, *heart sound representation* concerns the task of transforming the sound signal into a numerical representation (typically vector-based) that facilitates the previous identification task. The problem of identifying heart sounds is somewhat more demanding than representation because solution correctness can be defined more precisely for the former problem. An important consideration of these studies is that they operate almost exclusively with the output of a phonocardiograph (PCG) and a set of contact sensors which provide a very noise-free sound signal. Often this sensor technology is complemented with an electrocardiogram (ECG) and/or carotid pulse sensor that is used to "gate" the sound signal and to provide a segmentation aid for determining the cardiac phase.

One of the inconveniences of working with recordings made by electronic stethoscope is that most research in the area of analyzing heart sounds has been with the phonocardiograph, often with an ECG signal which provides a pulse and a means of registering the signal to the cardiac cycle. Because of this, much research with heart sounds [21, 49, 22, 43, 32, 48, 19, 7] is conducted with ECG, either as an input feature or as a method of labeling the training data. Relatively few use the sound itself as the only data [33, 25, 38, 15] and these predominantly use the phonocardiograph to collect input. The problem of identifying the sounds becomes easier with ECG because this signal provides a landmark with which to register the cardiac cycle. The R-wave[1] peaks of the ECG signal denote the onset of the systolic interval and arrive shortly before the S1 beat [19]. Given the ECG signal and in particular the peak of

[1]This is usually the dominant peak in the ECG signal in a healthy patient.

the R-wave, identifying S1 from S2 becomes a trivial task. That said, there are very few existing methods which identify S1 and S2 for a wide variety of pathologies.

Some of the earliest works in the area of heart sound analysis examined the representation of the signal from a purely visual perspective (see Obaidat [37]). This type of study is still of interest as it is becoming more common for clinicians to visually inspect the signal during diagnosis (see Kudriavtsev [29]). Indeed, the rise in availability of electronic stethoscopes should make it more common for the physician to use a computer in diagnosis.

Early attempts to the identification problem in the 1990s relied on a scalar "energy" feature that is thresholded in order to identify regions containing the heartbeats. These methods often relied on hand-tuning of threshold values and fixed algorithms to perform identification. For example, Liang et al. [33] perform PCG heart sound identification using the normalized Shannon energy and thresholding to identify candidate beats, followed by a fixed rejection algorithm to identify certain beats as false positives. Later, Haghighi et al. [21] used the power spectral density of a target sub-band in an autoregression model of the raw PCG signal. This method is notable as one of the first to "learn" parameters (here, of the autoregression model) in order to solve the identification problem.

Another attempt to solve the identification problem with no help from ECG came from Hebden and Torry [22] who used a neural network to identify the start and endpoints of S1 and S2 within a recording sequence. That is, their method treated the problem as a segmentation task. Although their method relied on ECG to produce correct labelings for training data, it adopted the strange practice of using performance on test data to determine the stopping criterion of the training procedure. Still, this attempt stands out as one of the few that has looked exclusively at the task of identifying S1 and S2. Despite the innovations made in this work, the feature set used to represent the input was seemingly ad-hoc and not motivated by the physics or physiology of the sound. Additionally, the authors do not reveal the topological structure or configuration of their segmentation network, making the method irreproducible.

It should be noted that most solutions to the identification to this point take the form of segmentation tasks, where the S1 and S2 beats are represented by bounds which surround regions of variable length.

As computing power grew in the early 2000s, identification methods based on *time-frequency* (TF) representations flourished, bringing with them higher dimensionality and a need for numerical efficiency that could not be satisfied

with hand-tuned algorithms. It is in this setting that learning and optimization techniques started to become more popular for heart sound analysis. For example, the Morlet wavelet representation was used by Rajan et al. [38] as an input representation to a simple perceptron that was trained to identify the two major heartbeats as well as clicks and pops and other types of stethoscope noise. The TF representation allowed the authors to make explicit use of frequency information to distinguish the first and second heart sounds from other noises such as murmurs. Further exploration of TF representations was done by Wang et al. [48] while looking specifically at the S1 heart sound and using Mallat and Zhang's matching pursuit algorithm [34]. This is perhaps the earliest attempt to "learn" a numerical representation of the heart sound rather than supply it with hand-tuned features or those computed directly from the signal via filter bank. Another major stride was a pair of studies by Xu et al. [50, 51] which use a TF representation of the S2 heart sound as a raw signal, but characterize the source of the observed signal as a parameterized chirp whose phase function is modeled using a high-order polynomial and whose amplitude is found by solving a least-squares problem. It is during this time that researchers began to study the time-frequency "signatures" of specific cardiac events as a separate research goal, and indeed the last four papers in this category were structured around this approach.

The Xu et al. articles highlight a major research step in the realization that the relatively complex TF representations could be viewed as the "observed result" of a much simpler yet hidden source model. This view motivated the study of generative probabilistic models and was followed by Gamero and Watrous' study [19] using hidden Markov models (HMM) to model the PCG output represented by its Mel-frequency Cepstrum Components [14]. Although this study found a very low rate of error, their method is applied to a relatively noise-free PCG signal and uses ECG gating to produce reliable predictions. It is also noteworthy that commercialization of heart sound identification software started to become more visible at this time, and indeed the former authors submitted their work on behalf of a corporation[2] rather than an academic research institution.

Although many of these studies report high rates of success, no controlled study has been done comparing the methods on a common data set. Furthermore, nearly all of the work done in heart sound analysis has been done as a form of phonocardiography (which technically includes stethoscope sounds but classically refers to the relatively less noisy output of the phonocardiogram). Less work has been done on the subject of automated *auscultation*, which is the act of diagnosing a patient using the sounds heard via stethoscope. The modern digital stethoscope uses a single microphone and produces a relatively noisy signal compared to the set of contact sensors that are used by PCG.

[2]Zargis corporation.

The lower signal to noise ratio makes the identification problem slightly more difficult in this setting, and also further motivates probabilistic methods for the purpose of making robust predictions. Of the studies mentioned above, only the Gamero & Watrous study attempted to model the inputs using probabilistic methods (e.g.: using a density function and/or distribution function).

Despite the shortcomings of the stethoscope as a measurement device, it is still more commonly used than PCG as a diagnostic tool due to its ease of use, low cost, and portability. Although one might argue that identification of heart sounds via PCG is a solved problem, (and even more so with the aid of ECG to gate the cardiac phase) the presence of noise and the reliance on a single sensor make the identification problem still somewhat difficult when using the digital stethoscope.

This work contrasts with previous works by focusing exclusively on sounds collected by digital stethoscope rather than PCG, and does not rely on ECG gating to obtain markers of the cardiac phase. The model is also more advanced than earlier attempts in heart sound modeling in that it attempts to account for individualistic variation in the input patterns. There are two major theoretical foundations for the current work. From the literature in stochastic processes, there exists a class of models known as *point processes* which are well-suited to modeling "instantaneous" sequences of sparse events in the time domain. From applied harmonic analysis, one can obtain a wide variety of harmonic representations that provide a time-frequency feature appropriate for detecting the subtle changes in the tone and pitch of a sound recording. A quadratic filter known as the *pseudo-Wigner Ville distribution* (PWVD) has been selected on the basis of past application to heart sounds [29].

2.1 Point processes

Where automated diagnoses are concerned, there is some value to modeling predictions of health or biological "state" with probability theory. The ability to determine a level of confidence is especially important in these cases; and it is unavoidable that the signals collected contain some amount of noise. It is therefore beneficial for a diagnostic program to be able to emit its diagnoses in a way that communicates the uncertainty associated with the prediction. This is necessary in order for physicians to position the automated diagnosis in the context of partial information and classical clinical protocols.

The probability model used here to model the unknown times of the heartbeats is known as a *point process* [13]. This type of process describes a set of instantaneous points embedded in time called *arrivals*, each of which occurs at a precise instant known as the *arrival time*. The model provides a means

to model the time intervals between the arrivals explicitly using positive distributions. Here, we assume that the time intervals between the arrival times (known as *waiting times*) are independent and identically distributed.

The use of explicit potentials to model these intervals makes point processes distinct from the related *Bernoulli-Gaussian processes* (BGPs) [27]. BGPs treat observable real-valued observations as a Gaussian process conditionally dependent on a sequence of discrete labels which are latent. BGPs use a state space model that is based on random impulses, though these impulses are modeled as i.i.d. for every time step in the sequence and the times between the impulses are not modeled explicitly. Variants exist in which the observable sequence is a Gaussian mixture model whose component is determined by the hidden label [17]. Other related models include:

- hidden Semi-Markov models (HSMMs) [35, 54]: the hidden state is a discrete label whose value is sustained over a sequence of time steps of random length.
- jump-Markov linear systems (JMLSs) [44, 16]: the hidden state is a discrete label and the observed variable is determined by a linear system whose parameters are selected by the hidden label.

Subtleties of the state space make these models inappropriate for heart sound detection. The JMLSs use a real-valued state space that is influenced by a Bernoulli sequence. This formulation does not explicitly model the interval times between the "impulses", which are treated as i.i.d. and are drawn from a distribution over positive numbers. Under the i.i.d. assumption, these intervals must be exponentially distributed and so they do not accurately represent the dynamics of the cardiac cycle. The HSMMs do allow for explicit modeling of the *sojourn time*, or the time spent in a discrete hidden state. Here, it is difficult to specify an observational model for heart sounds that is structured around discrete "states" and can accomodate non-independent sequential data of arbitrary length.

A *simple point process* (SPP) can be thought of as a sequence of *arrivals* $Q = \langle q_i \rangle_i$ with q denoting an arbitrary element in the set. Each arrival is associated with a sequence of positive real-valued scalar random variables representing the *arrival times* $\mathbf{T} = (T_1, T_2, ...)$ which are monotonically increasing. An assignment of the random variables in a point process (whether considered full or partial) is referred to as a *configuration*.

Formally, a general point process can be thought of as the space of all measurable sequences defined over a more fundamental measure space. For an elegant treatment of general point processes in algebraic terms, see [5]. A more concise set of notation is given here. Let S be an arbitrary set, so that

$(S, \mathcal{B}, \mu)$ is a measure space. Then $(\mathbb{S}, \mathcal{F}, \mu^*)$ is the measure space of finite sequences of elements of S, with:

$$\mathbb{S} \;=\; \{t_1, ..., t_K \mid \forall i : t_i \in S,\; K \geq 0\} \tag{2.1}$$

The ordering of the elements within a subset is normally not relevant and one can define processes over joint time and space domains where S is multivariate. However for point processes in the time domain i.e. $S = \mathbb{R}_{++}$, it is appropriate to impose the monotonicity constraint $t_i < t_{i+1}$ for all i. Normalization of the measure λ via $1/\lambda(\mathbb{S})$ yields a probability space over finite configurations of points. Then the following conditions apply, taken from [24]:

$$\mathcal{P}(0 < T_1 \leq T_2 \leq \cdots < T_q) \;=\; 1 \tag{2.2}$$

$$\mathcal{P}(T_q < T_{q+1}, T_q < \infty) \;=\; P(T_q < \infty) \tag{2.3}$$

$$\mathcal{P}\left(\lim_{q\to\infty} T_q \;=\; \infty\right) \;=\; 1 \tag{2.4}$$

I.e.: the arrival times are treated as positive random variables, with eqn. 2.4 stipulating that only finitely many arrivals can occur within a finite time interval. (To see this, try assuming that an infinite number of events occur within a finite time interval; one can then immediately derive a contradiction by showing that the T_q is finite in the limit). The waiting times are denoted $Z_q = T_q - T_{q-1}$.

When a simple point process is defined on a temporal domain such as $\mathbb{R}, \mathbb{R}_{++}$, or $\mathbb{N}$, an alternative representation can be given in terms of its *waiting times* $Z_q = T_{q+1} - T_q$ and the time of the first arrival T_1, i.e.: $(T_1, Z_1, Z_2, ...)$. The probability space associated with the point process can be defined by providing distributions for these variables, which are defined over the positive real numbers. One of the simplest and most commonly found definitions given for such a distribution treats these variables as independently and identically distributed exponential random variables; a Poisson process.

Other representations exist which provide greater flexibility over the probability of witnessing a single (unique) arrival conditioned upon arbitrarily many other arrivals within the process or in parallel processes. For example, it is possible to define an *intensity function* $\lambda(t)$ as follows:

$$\lambda(t) \;=\; \lim_{h\to 0} \mathcal{P}(\; \exists! q : t < T_q \leq t + h \;)/h \tag{2.5}$$

Note that the uniqueness of this event precludes the possibility that two arrivals occur in the period $(t, t + h)$. This notation characterizes the intensity function as a predictor of a single arrival rather than than as a predictor for the time course that contains it. The intensity function representation is useful when the times of the arrivals is conditioned on self-excitation of the process

by previous arrivals [42, 4]. This is particularly true when the point process is expressed in terms of its *filtrations*.[3] Roughly speaking, filtrations are historical segments of the original process defined on left-connected subsets[4] of the original fundamental measure space. That is, they provide a view of the history of some process relative to an arbitrary point t. In this context, the intensity function can be used to predict the probability associated with seeing a future event after time t for very small distances into the process' future, conditioned on the history preceding t.

The intensity function can also be defined in terms of another point process [3], or in terms of points scattered in another space [45]. Both of these approaches would be suitable for future investigations into modeling heart sounds. For example, if one considers the other constituent sounds that make up the heart sound recording (murmurs, clicks & pops, fluid flow) as well as other processes which excite or mitigate the cardiac cycle (respiration) it is readily apparent that heart sound recordings taken from a clinical population might benefit from the application of such complex models. For simplicity, only a single process is employed to represent the cardiac rhythm in the current study.

The use of a single SPP has been used recently to model the "instantaneous" heart rate of a patient using ECG inputs [1, 8, 7, 10] and previously using the electromyogram [41]. In their application, the heartbeat times are known and the inference task is to determine time-dependent parameters of the interval distribution throughout the recorded sequence. Here, the inference task is somewhat simplified by the presence of known heartbeat times; the parameters of the point process can be computed directly by taking the waiting times from a labeled recording as i.i.d. samples from the related distributions.

One extension to the above is the *marked point process* (MPPs) [30]. An MPP is a simple point process augmented with a set of random variables for each arrival $\mathbf{X} = (X_1, X_2, ...)$ called *marks* that are members of an arbitrary set $X_q \in E$. The mark space E is completely arbitrary, e.g.: $\mathbb{R}, \mathbb{N}$, or a finite set of labels. In the classic treatment of the MPP, the mark set is augmented with a special mark called the *irrelevant mark* denoted ∇, so that $\bar{E} = E \cup \{\nabla\}$. This extra member allows Q to be countably infinite, with:

$$\mathcal{P}(T_q = +\infty, X_q = \nabla) = \mathcal{P}(T_q = +\infty) \tag{2.6}$$

so that the probability space can still represent sequences of finite marked events. In the following work, the mark space E is simply treated as a finite

[3]This use of the word *filtration* concerns a concept defined more precisely in the stochastic process literature. See [24] for more information. We do not examine in more detail here.

[4]A left-connected subset is a connected subset that contains the infimum value of the original set. The set functions for the sets used in filtrations are "càdlàg" (continue à droit, limite à gauche). They are continuous on the right and contain their limits on the left.

set of labels (the two heart sound types) and the arrival times are strictly increasing $T_{q+1} > T_q$.

The typical use of the MPP is to characterize heterogenous processes that occur sparsely in some domain [6, 9]. Our application to heart sound identification fits this description: there is a small, finite set of beats which can be assigned to the mark set, the observed data are sampled at a high rate, and the duration of each beat can be characterized as an instantaneous point which is the "center" of the sound generated by the heart valve closures for that beat. What remains is to describe the influence of these marked points on the observed stethoscope recording. For this, there exist contemporary signal processing methods which can be integrated with a point process model to produce a coherent probability model for heart sound analysis.

2.2 The pseudo Wigner-Ville distribution

In order to estimate the locations of the heartbeats, a feature representation must be used that can express variation in the frequencies present in the signal in a time-varying manner. For this purpose, a variety of harmonic representations have been considered for the phonocardiogram which may serve in the analysis of stethoscope recordings. These include the short-time Fourier transform [15, 39, 40, 52], a variety of wavelet transforms [25, 43, 52], and the Wigner transforms [39, 40, 49]. It is a frequently cited fact in the above studies that the short-time Fourier transform does not possess sufficient resolution to reveal the pair of valve closures that make up each heart sound as distinct events. In the case of the Wigner transform, proper resolution of the valve closure sounds often depends on correct setting of a scale parameter, although this value can vary from patient to patient, between recordings, or even between cardiac cycles themselves [32].

The pseudo Wigner-Ville distribution (PWVD) is a member of Cohen's [11, 12] class of distributions. It provides a time-frequency representation in the form of a complex-valued feature map. This feature map provides an intensity value for every time, frequency pair in a compact plane.

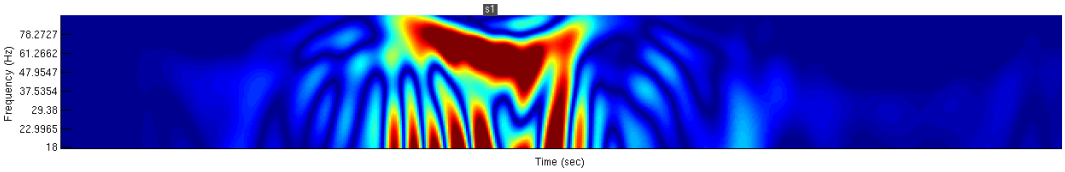
$$\Phi(t, f) \;=\; \int x(t-\tau)^* x(t+\tau)\, h\left(\frac{\tau}{\sigma}\right) \exp\left\{-2\pi i f \tau\right\} d\tau \tag{2.7}$$

The window function $h(\tau)$ is a scaled Gaussian function with scale parameter σ. The distribution is usually calculated for a uniformly discretized grid over a region of finite length and frequency range. Typically, the modulus of the result is used as a feature input and the phase component is discarded: $|\Phi(t, f)| \in \mathbb{R}_+$. This provides a representation that is real-valued and non-negative.

The strength of the PWVD is to characterize the distribution of frequency components "contained within" the energy of the signal. The semantics of the term "distribution" here are intended to reflect a semblance to the disproportionate allotment of mass in a probability distribution, but does not carry the mathematical denotation associated with its use in probability theory.

In heart sound analysis, the largest values of the PWVD representation tend to be well separated from mean values; this behaviour is characteristic of sounds captured by stethoscope. If these values were to be used as features, the peak values would tend to dominate in certain calculations, i.e.: convolutions and filtration operations that rely heavily on summation. In heart sound data, the most interesting parts of the time-frequency representation are the beats, which tend to produce localized maxima and loosely connected regions that contrast sharply with the ambient noise floor. In a given recording, these peaks do not tend to remain in the same intensity range. Rather, the local peaks in the time-frequency representation can fluctuate wildly and it is not uncommon to see a dominant beat type whose signal energy is one hundred times greater than that of the more diminutive beat type. At the same time, the general shape or *spectral signature* of a heartbeat type tends to remain somewhat consistent throughout the recording.

The PWVD has been proposed for use in auscultation in the past [29]. Its primary advantage is high time-frequency resolution, though the filter tends to produce a distinctive pattern of rippling artifacts as consequences of the quadratic filter response. With the correct choice of scaling parameter, the PWVD provides high contrast between regions of relatively high and low signal intensity.



The ripples are a consequence of the quadratic filter's differential response to the phases of the basis function for each frequency band.

A central difficulty involved with the use of time-frequency representations such as the PWVD is that of achieving high resolution of a pure signal. When a pure sine wave at frequency f is converted into any particular time-frequency representation, the PWVD responds at nearby frequencies $f \pm \delta$ due to the similarity among the neighbouring frequences. This phenomena occurs across a wide variety of harmonic representation classes, and is dubbed the Gabor *uncertainty principle* [18]. Significant research effort has been devoted to circumventing this limitation. While it is known that the PWVD is not able to

resolve the split in the gap between the valve closures in the first or second heart sound [36], it suffices for identifying one heart sound from the other.

Before using the PWVD representation as input to an algorithm, its values are scaled using a normalization method developed in Section 3. This method bounds the values of the features and eliminates some of the widely varying peak intensities that can occur in the course of a heart sound recording, while preserving the contrast necessary to identify each beat.

Chapter 3

Signal Scaling

For signal data that are acquired by recording natural sounds (e.g.: speech, music, heart sounds) the signal energy $|s(t)|$ typically obeys a heavy-tailed distribution[1]. When filtration methods are used to transform the signal from its waveform representation to a new representation (e.g. a time-invariant functional or time-frequency feature map) these large amplitudes are often transferred to the new representation. As a result, a number of undesired consequences can occur. For example, the maximal values in the new representation can mask or overshadow each other if they occur in close proximity to one another and yet have differing magnitude.

In order to deal with the influence of large changes in intensity that are produced by heart sound recordings, a simple method is offered for scaling a function that preserves the structure of its local maxima while bounding the range of the map without introducing discontinuities (e.g. from less elegant strategies such as truncation). The main idea behind this method is that the feature map is differentially scaled so that its local maxima attain the value $f(x^*) = 1$ while the curvature, contrast, and overall shape of the function are approximately preserved.

Intuitively, the distance from a maximal point should help determine how the function is scaled; in this way, the maxima themselves can be scaled in relation to one another should they occur in close proximity. Our solution in algorithm 1 will therefore provide an answer to the question of what constitutes locality when considering a maximal point, and how should neighbouring maxima be viewed when they are close to one another.

Consider an arbitrary feature function $\phi[x]$ defined over the joint feature-time space, where x represents an instant in time and $\phi[x] \in \mathbb{R}_+^M$. The maximal

[1]Informally, this means that the corresponding density function converges to zero "slowly". As a result, a larger portion of the probability mass is located away from the modes of the distribution and by a larger distance than a distribution that is not heavy-tailed.

points of the map are used in conjunction with a symmetric kernel $k(x, x')$ (e.g. a Gaussian kernel). Note that it is desirable to have $k(x, x') \leq 1$ and $k(x, x) = 1$ in order to produce the condition given below in eqn. (3.1). In the context of the PWVD features discussed in the previous section, the notation can be interpreted to mean that $\phi[x]$ and $h[x]$ represent vector quantities, e.g.: $\phi[x] = \Phi[x, :]$. (Here we use the MATLAB-like notation ":" to denote all indices in the second dimension of the array Φ). In this case, the scaling is performed component-wise.

Algorithm 1 $[\ \bar{\phi}\ ]$ = `relative_scaling`$(\ \phi, \nu, k(\cdot,\cdot)\ )$

for all $x \in \text{domain}(\phi)$ **do**
  $h[x] \leftarrow \nu$
end for
$M \leftarrow$ local maxima of $\phi[\cdot]$
for all $x' \in M,\ x \in \text{domain}(F)$ **do**
  $h[x] \leftarrow \max(k(x, x') \cdot \phi[x'], h[x])$
end for
for all $x \in \text{domain}(f)$ **do**
  $\bar{\phi}[x] \leftarrow \phi[x]\ /\ h[x]$
end for

The algorithm makes use of the map h to store a scaling factor for every point in the map. This scaling factor is defined by multiplying each maximal value of the map by a Gaussian kernel function that is centered on that point. Initializing $h[\cdot]$ with a small value $\nu > 0$ prevents division by zero and ensures that only neighbourhoods with significant maxima (e.g. those above β) are considered. For example, a typical pre-processing step is to first scale the feature values so that their mean value across the entire recording is equal to 1, and correspondingly to set $\beta = 1$. In this way, "small" maxima are effectively ignored.

Since the scale factor for a maximal point x^* is its own value $\phi(x^*)$, maximal points will be reassigned the scale factor $\bar{\phi}(x^*) = 1$:

$$\bar{\phi}[x^*] = \frac{\phi[x^*]}{\max\limits_{x} \phi[x]} = 1 \qquad \text{by definition} \tag{3.1}$$

provided that no other maximal points are sufficiently close *and* sufficiently large. The kernel function k often has a horizontal scaling parameter which controls the tradeoff between closeness and function value. Here, a wider kernel function will produce a more discriminating comparison between any two maxima at the same distance, or the same level of discrimination at a wider distance.

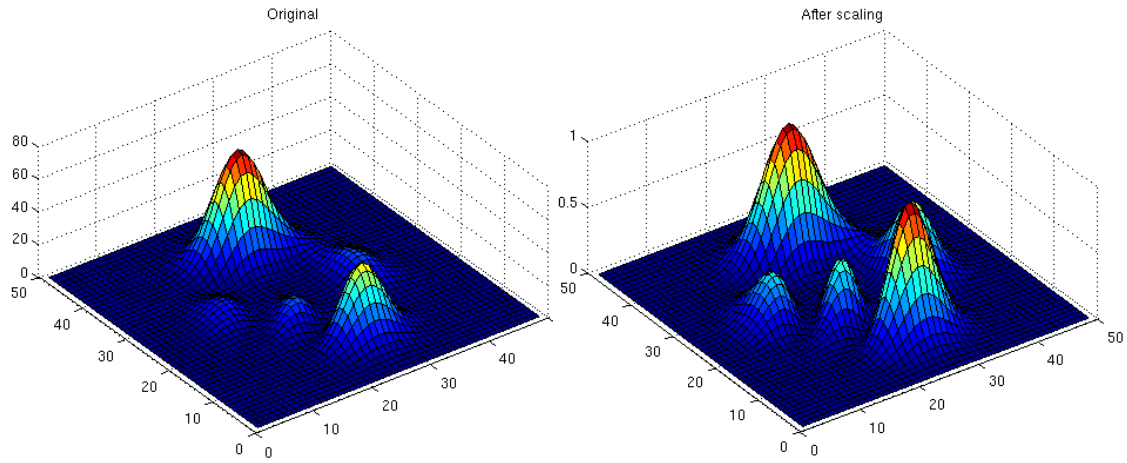


Figure 3.1: A 2-dimensional "signal" before and after using the relative scaling operation. Note that only the extremal values are actually bounded at $f(x^*) \leq 1.0$. It is still possible for the function to exceed this value after the transformation.

Chapter 4

A Modification to Marked Point Processes

The mathematical model for the following *modified marked point processes* can be described by three sets of variables denoted by the tuple $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$. The first set of variables $\mathbf{Y} = \langle Y_t \rangle_{t \in I}$ are a sequence of real-valued feature vectors subscripted by a time index. These values are the "observed" part of the model; the other values are latent and represent a marked point process. In heart sound modeling, these values represent time-dependent feature values that are a function of the input recording. The variables $\mathbf{T} = \langle T_q \rangle_{q \in Q}$ are the arrival times, and the variables $\mathbf{X} = \langle X_q \rangle_{q \in Q}$ are discrete labels (marks) that provide a hidden state representation.

The observed random variables Y_t are typically a vector of time-frequency features computed directly from a discrete signal, but in practice they can be any feature vectors as long as a conditional density can be provided that is conditioned on the time and type of an arrival. With heart sounds, the time-frequency representation can be sampled at a rate that is a lower multiple than the discrete signal. For example, for observed data $Y_t = y_t$, the vector of features is computed using $y_t = \Phi[\frac{t}{n}, :]$, where Φ is a suitable time-frequency representation such as the PWVD.

The indices $q \in Q$ refer to a single arrival in a chain of length $|Q|$ where $T_q \in I$. The subset of the observed variables that are influenced by these arrivals is $\mathbf{Y}_{\boldsymbol{\mathcal{T}}}$, and the subset of observed variables influenced by a single arrival is $\mathbf{Y}_{\boldsymbol{\mathcal{T}}_q} \subseteq \mathbf{Y}_{\boldsymbol{\mathcal{T}}}$ Although this definition subsumes a wide variety of generative observational models, the principal configuration used here is that $\mathbf{Y}_{\boldsymbol{\mathcal{T}}}$ refer to those variables that are *w-proximal* to some arrival in the chain. Formally, w-proximality is defined as:

$$\mathbf{Y}_{\boldsymbol{\mathcal{T}}} \quad = \quad \{y_t \mid \exists q : |T_q - t| \leq w\} \tag{4.1}$$

for arbitrarily chosen w. I.e.: $\mathcal{T}$ is a qualifier designating the set of indices q that meet the construction in eqn. (4.1) In order to maintain tractability,

it is assumed that *there is at most one arrival that influences an observation at any time step,* i.e.: $[q \neq q'] \Rightarrow [\mathbf{Y}_{\boldsymbol{\mathcal{T}}_q} \cap \mathbf{Y}_{\boldsymbol{\mathcal{T}}_{q'}} = \emptyset]$. This simplification makes the current approach different from classical treatments of the point process, where the generative view of the model stipulates that new arrivals are the product of an ambient intensity rather than presuming the existence of a deterministic chain whose arrivals are consequences of those before them. Without this simplification, the computation of arrival likelihood becomes a problem of deconvolution that becomes very difficult as one considers large numbers of overlapping events.

4.1 Probability model

The variables form a joint density/mass function that factorizes as follows:

$$\mathcal{P}(\mathbf{Y}, \mathbf{X}, \mathbf{T}) \;=\; \mathcal{P}(\mathbf{Y}|\mathbf{X}, \mathbf{T})\, \mathcal{P}(\mathbf{T}|\mathbf{X})\, \mathcal{P}(\mathbf{X}) \tag{4.2}$$

This factorization is valid under the assumptions:

- That a heartbeat's time and type is independent of all other heartbeats in the chain given only the time and type of the beat preceding it.
- That the observations are conditionally dependent on the times and types of all the heartbeats.
- That the sequence of heartbeat types is entirely deterministic.

Under the assumption that the event types X_q form a Markov chain, eqn. (4.2) can be factorized further. Numbering the arrivals from 1 to $|Q|$, and treating X_0 and T_0 (i.e.: $q = 0$) as placeholder notation for an arrival that preceded the first arrival the chain, the mark distributions can be written:

$$\mathcal{P}(\mathbf{X}) \;=\; \prod_{q \in Q} \mathcal{P}(X_q \,\big|\, X_{q-1}) \tag{4.3}$$

...where $\mathcal{P}(X_1|X_0)$ can be either an "initial label" distribution over the label types, or it can be a steady-state distribution that is consistent with the stochastic matrix that describes $\mathcal{P}(X_q \,\big|\, X_{q-1})$.

Under the assumption that the waiting times are i.i.d., the arrival times can be factored as:

$$\mathcal{P}(\mathbf{T}\big|\mathbf{X}) \;=\; \prod_{q \in Q} \mathcal{P}(T_q \,\big|\, T_{q-1}, X_q, X_{q-1}) \tag{4.4}$$

Note that the arrival times are easiest to model via the waiting times $Z_q = T_{q+1} - T_q$. This density can be represented by a function over a positive variable

that represents the interval between the two arrivals, i.e.: for some pair of label types $X_{q+1} = b, X_q = a$. For $q > 1$,

$$\begin{aligned} \mathcal{P}(T_q|T_{q-1}, X_q, X_{q-1}) &\sim \mathcal{P}(T_q - T_{q-1}|X_q, X_{q-1}) \\ &= \mathcal{P}(Z_{q-1}|X_q, X_{q-1}) \end{aligned} \tag{4.5}$$

This motivates the following density function, which captures the waiting time in terms of neighbouring arrival times:

$$\begin{aligned} f_{ab}(z) = f_{ab}(t' - t) &= \mathcal{P}(T_q = t' \,\big|\, T_{q-1} = t, X_{q-1} = a, X_q = b) \\ \text{i.e. } f_{ab} &: \mathbb{R}_{++} \to [0, 1] \end{aligned} \tag{4.6}$$

Here, f_{ab} can be used to define the probability for any positive length interval between successive arrivals with labels a and b.

Recall the observed variables that depend on an arrival q are $\mathbf{Y}_{\boldsymbol{\mathcal{T}}_q}$ and that these are by requirement w-proximal to T_q. Also consider that since a given observation can depend on at most one arrival, the arrivals must be spaced apart by at least $2w$. This implies that the chosen value of w will shape these density functionals; specifically for $x \leq 2w$, $f_{ab}(x) = 0$ in order to prevent the arrivals from influencing the same observed variables Y_t. An obvious solution is to apply a non-negative distribution that is shifted horizontally by $+2w$, meeting the above requirement.

In practice, the source of the arrivals producing the observable data may extend beyond the start and end of the recorded signal. In this case, extra factors must be introduced to account for the gaps between the first and last arrival near the edges of the recording. The need for these factors is illustrated in figure 4.1.

Recall that the joint probability model captures only a finite chain of variables even though the chain of arrivals extends backward and forward in time past the bounds of the observed variables. One might presume an arrival time $T_0 < 0$ for some precursor arrival that occurred before the start of the recorded signal, and similarly an arrival $T_{|Q|+1} > L$ that arrives after its end. So, the first factor in eqn.(4.4) models the probability of seeing the first visible arrival at time t given that some arrival preceded it and occurred before the start $(t' = 0)$ of the sequence.

To express these factors, a distribution function is created by conditioning on $T_0 < 0$ and integrating the appropriate density function for the waiting times:

$$\mathcal{P}(T_1 = t \,\big|\, T_0 < 0, X_0 = a, X_1 = b) = \int_t^\infty f_{ab}(x)dx \tag{4.7}$$

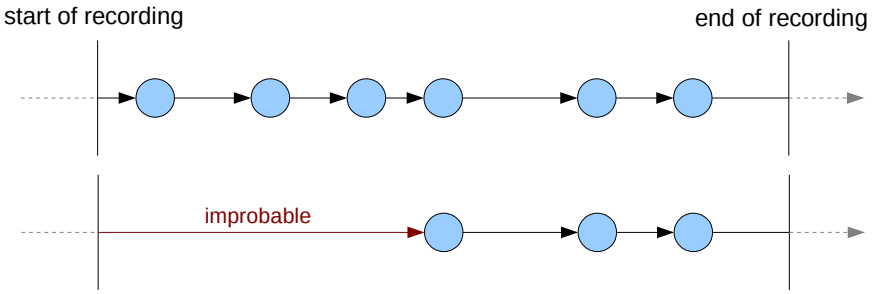


Figure 4.1: The configuration on the top shows a typical recording sequence whose start and end fall in between some pair of arrivals in an infinitely long chain. What is needed is a way to account for the waiting time between any unseen arrivals occuring outside the recording and those producing observable patterns within it. Without taking this waiting time into consideration, it is possible to assign a high probability to a chain with unlikely waiting times near the start and end of the sequence. This would be the typical outcome of inference about the best chain of arrivals since a smaller chain of arrivals would in general produce a smaller total log-likelihood values.

This allows us to marginalize over X_0, T_0 instead of conditioning on these variables:

$$\begin{aligned} \text{e.g. for} \quad & \mathcal{P}(T_1 = t \;\big|\; X_1 = b) \qquad (4.8) \\ &= \sum_a \mathcal{P}(T_1 = t \;\big|\; T_0 < 0, X_0 = a, X_1 = b)\, \mathcal{P}(X_0 = a) \\ &= \sum_a \left[1 - \int_0^t f_{ab}(x)dx\right] \mathcal{P}(X_0 = a) \\ &\equiv f_{0b}(t) \qquad (4.9) \end{aligned}$$

and the substitution of $\mathcal{P}(T_1 \;|\; T_0, X_0, X_1)$ in eqn.(4.4) by $\mathcal{P}(T_1 \;|\; X_1)$.

A similar approach can be used to account for the end-gap, although this factor cannot be hidden by introducing separate semantics for X_0, T_0 in $\mathcal{P}(T_1|\cdots)$. Instead, they must be supplied separately from the product in eqn.(4.4), leading to a revised expression for the arrival time model:

$$\begin{aligned} &\text{let } e \;=\; |Q|, \quad e' \;=\; |Q|+1 \\ &\mathcal{P}(T_{e'} \geq L \;\big|\; T_e, X_e = a, X_{e'} = b) \;=\; \int_t^\infty f_{ab}(x)dx \qquad (4.10) \\ &\mathcal{P}(\mathbf{T}\big|\mathbf{X}) \;=\; \sum_b \mathcal{P}(T_{e'} \geq L \;\big|\; T_e, X_e = a, X_{e'} = b)\, \mathcal{P}(X_{e'} = b \;\big|\; X_e = a) \cdot \\ &\qquad \prod_{q \in Q} \mathcal{P}(T_q \;\big|\; T_{q-1}, X_q, X_{q-1}) \qquad (4.11) \end{aligned}$$

...where L is the length of the observed sequence. This revision adds one extra factor for a supposed arrival e' after the end of the recording, and avoids modeling observations for this arrival altogether by summing out the time and label for this arrival. Eqn.(4.11) can now be used in place of eqn.(4.4) in order to correct for the start-gap and end-gap phenomenon.

The observation model can also be factorized according to the arrival chain. The set $\boldsymbol{\mathcal{T}}_q$ refers to the set of points in the discretized sample space that are influenced by some arrival q. The factorization is:

$$\mathcal{P}(\mathbf{Y}|\mathbf{X},\mathbf{T}) \;=\; \mathcal{P}(\overline{\mathbf{Y}_{\boldsymbol{\mathcal{T}}}}) \prod_q \mathcal{P}(\mathbf{Y}_{\boldsymbol{\mathcal{T}}_q} \;\big|\; X_q, T_q) \tag{4.12}$$

where $Y_{\boldsymbol{\mathcal{T}}_q}$ denotes the set of observations that are w-proximal to the arrival time T_q, and $\overline{\mathbf{Y}_{\boldsymbol{\mathcal{T}}}}$ denotes those observations that are not influenced by any arrival. A distinctive aspect of this factorization is that the observation variables exhibit *context-specific independence*. Specifically, this means that the values of the arrival times T_q determine the conditional dependence relationships of the variables in $\mathbf{Y}$ variables via the set $\mathbf{Y}_{\boldsymbol{\mathcal{T}}}$. See [26] for a more thorough review.

This brings the full factorization of the model to:

$$\begin{aligned}\mathcal{P}(\mathbf{Y},\mathbf{X},\mathbf{T}) \;=\; & \mathcal{P}(\overline{\mathbf{Y}_{\boldsymbol{\mathcal{T}}}}) \cdot \\ & \prod_{q\in Q} \mathcal{P}(X_q \;\big|\; X_{q-1})\, \mathcal{P}(T_q \;\big|\; T_{q-1}, X_q, X_{q-1})\, \mathcal{P}(Y_{\boldsymbol{\mathcal{T}}_q} \;\big|\; X_q, T_q) \cdot \\ & \sum_b \mathcal{P}(T_{e'} \geq L \;\big|\; T_e, X_e, X_{e'} = b)\, \mathcal{P}(X_{e'} = b)\end{aligned} \tag{4.13}$$

All the model factors and their independence relationships can be visualized as a graphical model in a dynamic Bayes network (DBN). An illustration of the model is given in figure 4.2. In this illustration, the special notation $\overline{\mathbf{Y}}(t_1, t_2, w)$ refers to the observations that lie between two events.

4.2 Estimation of parameters

Inference of parameters for the conditional density and mass functions described above can be conducted with a training set consisting of *point-labeled* signal recordings denoted S_{train}. An arbitrary member of this set is denoted s. Associated with each member is: a) a sequence of observed data y_t^s sampled at a known uniform sampling rate, b) a sparse set of labels $x_q^s \in [1, N]$ for a set of known typed arrivals Q_{train}, and c) a corresponding sparse set of times $t_q^s \in \mathbb{R}_+$ representing the arrival times. When not referring to a specific recording in S_{train}, the superscript s on the associated data is dropped. These labeled recordings provide values for the random variables of the model described in section 4.

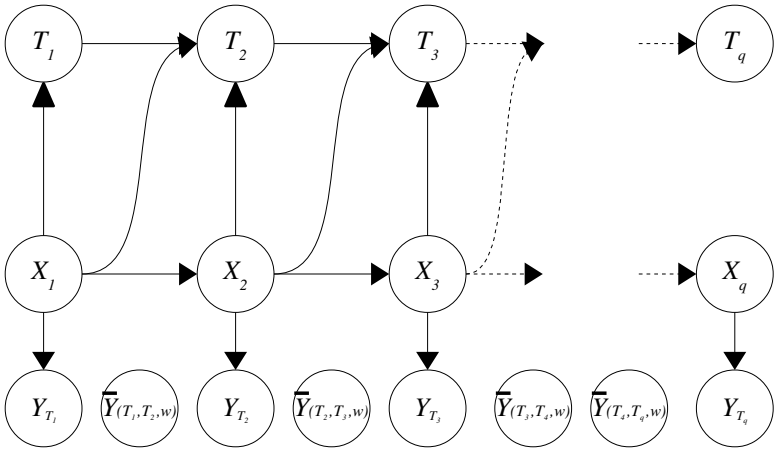


Figure 4.2: A dynamic Bayes network illustration of the probability model. Observed variables with the overline $\overline{\mathbf{Y}_{\boldsymbol{\mathcal{T}}}}(T_1, T_2, w)$ are not influenced by the arrival sequence and are independent of the rest of the graph.

One of the benefits of the model factorization described above is that when given complete data for X, Y and T, the parameter estimation problems for each of the groups of densities becomes fully independent of one another. This is because for each distribution in the factorization, data are available for both the dependent variables and their parents. These data can be treated as i.i.d. samples from stable distributions. The parameters of each model can then be treated as independent of the rest of the model given the labeled data for their specific distribution. This means that estimation can be achieved by computing separate estimators for each part.

One generic assumption is used throughout the parameter estimation part, and that is that if maximum likelihood estimators are available for a particular waiting time distribution, then we can compute an estimator of the parameter for the entire population that is unbiased by the number of beats in any particular individual. This "length free" estimator ignores the effect of having larger numbers of beats in recordings with faster heart rates. A revised score is computed as the average of the likelihood of each waiting time or each observation in the training data which is then divided by the total number of heart sounds. The motivation for this method comes from the assumption that the individual patient's identity is independent from the parameter that generated the data.

4.2.1 Parameters of the waiting time model

There are a wide variety of distributional models that can implement $\mathcal{P}(\mathbf{T}|\mathbf{X})$. The only prerequisite is that these density functions have support in the positive real numbers. Recall that the radius parameter w requires

that these density functions have zero density over the domain $[0, 2w]$. If traditional positive-support parameteric models are employed, (Poisson, Log-Normal, Gamma, etc.) they must be shifted horizontally by $+2w$ in order to prevent the condition whereby two arrivals occur within range of influencing the same observation. In the subsequent discussion, we restrict attention to parameteric models. Although it is possible to use non-parameteric models here, the following framework is designed with the idea of using likelihood formulations for which these estimators exist and are computable.

It is also worth mentioning that many distributions over the entire real line can be adapted to the non-negative reals using the elementary transformation $Z' = \log(Z)$.

Only a generic description of the training process is provided here. In this case, the waiting times z_q derived from the training labels are sufficient statistics for the estimators. If the labeled recordings are independent of one another, the likelihood function will contain separate terms for each of the recordings in S_{train}. Under the assumption that each recording is i.i.d., each estimate can be computed using a weighted likelihood in the objective by averaging over the number of arrivals in the individual labeled recordings.

$$\begin{aligned} &\text{let } q' = q + 1 \qquad \text{(an example sequence pair)} \\ &\text{let } g \text{ be a likelihood function with parameter } \theta \\ &\text{if } \hat{\theta}^s_{ab} = \arg\max_{\theta} \sum_{(q,q') \in \Pi(Q^s)} \ln g(x_q, t_q, x_{q'}, t_{q'} \,;\, \theta) \qquad (4.14) \\ \text{then } \quad &\hat{\theta}_{ab} = \arg\max_{\theta} \sum_{s \in S} \frac{1}{|Q^s|} \sum_{(q,q') \in \Pi(Q^s)} \ln g(x_q, t_q, x_{q'}, t_{q'} \,;\, \theta) \qquad (4.15) \end{aligned}$$

where $\Pi(Q)$ denotes the sequential transition pairs in Q. Solving this maximization globally relies on a differentiable, convex (or quasiconvex) function $\ln g()$. If this condition is present, the maximization can be solved using gradient methods.

As an example, the training procedure is illustrated using the log-Normal distribution to capture the waiting times.

All waiting times between pairs of arrivals of types a and b are considered i.i.d. samples of a single *population-wide* distribution denoted ab, and that a latent wait time of this type can be modeled as a random variable Z_{ab}.

The assumptions here are that:

- All beats are drawn from a common, *population-wide* distribution.
- The beats are i.i.d. within a pair class ab.

- As a consequence of the first two assumptions, the number of beats in any given patient's recording is not a factor in the likelihood expression.

These assumptions do not hold for heart sounds.[1]

$$\begin{aligned} x_{q-1} = a \quad , \quad x_q &= b \quad , \quad Z = T_q - T_{q-1} \\ Z_{ab} &\sim \text{LogNormal}(\theta_{ab}) \qquad (4.16) \\ \theta_{ab} &= \langle \mu_{ab}, \sigma^2_{ab} \rangle \qquad (4.17) \end{aligned}$$

The MLE parameters for θ_{ab} can be computed by maximizing the log-likelihood expression:

$$\mathcal{L}(\theta_{ab}) \;=\; \sum_{s \in S} \frac{1}{|Q^s|} \sum_{z \in \mathcal{Z}_{ab}} \ln\left(\frac{1}{\sqrt{2\pi}\sigma_{ab}}\right) - \frac{\left[\ln(z) - \mu_{ab}\right]^2}{2\sigma^2_{ab}} \qquad (4.18)$$

Here, $\mathcal{Z}_{ab}$ denotes a set of waiting times constructed from the t_q variables in the training data.

$$\mathcal{Z}_{ab} \;=\; \left\{ t_{q'} - t_q \;\middle|\; (q, q') \in P(Q)\,,\, x_q = a\,,\, x_{q'} = b \right\} \qquad (4.19)$$

It is now straightforward to maximize the sample-weighted likelihood using eqns.(4.14,4.15,4.18):

$$\hat{\theta}_{ab} \;=\; \arg\max_{\theta} \sum_{s \in S} \sum_{(q,q') \in P(Q^s)} \ln\left(\frac{1}{\sqrt{2\pi}\sigma_{ab}}\right) - \frac{\left[\ln(t_{q'} - t_q) - \mu_{ab}\right]^2}{2\sigma^2_{ab}} \qquad (4.20)$$

A note on the use of non-parameteric densities. There is some advantage to using non-parameteric estimators to model the waiting time densities. Consider the type of transition where an arrival of type b follows an arrival of type a only at some small integer multiples, but with some small noise that permits slight deviations from these modes. Classical unimodal parameteric models cannot capture this behaviour without resorting to the use of mixture models, and the number of mixtures can be difficult to know *a priori*. Non-parameteric models do capture multi-modal densities effectively, with one caveat: because the density is defined over positive numbers, it is easy to learn a density that puts too much mass near zero due to the selection of kernel functions and bandwidth parameter. Indeed, choosing the bandwidth parameter is one of the central challenges in constructing these estimators, but it is made more acute when the random variable is bounded on one side (i.e.: at zero).

4.2.2 Parameters of the label distribution model

It is notable that the chain of labels and their respective conditional probability functions can be thought of as a Markov chain. The full set of conditional

[1] A patient's waiting times generally vary according to a more precise distribution. This issue is revisited in section 5.

distributions can be specified in an $N \times N$ *stochastic matrix* of parameters denoted $\Gamma = \langle \Gamma_{ab} \rangle_{a,b \in E}$. Formally,

$$\Gamma_{ab} \;=\; \mathcal{P}(X_q = a \,\big|\, X_{q-1} = b) \tag{4.21}$$

The choice of initial distribution or steady-state distribution to use for the marginals is arbitrary. The parameters for this distribution can be stored in a vector $\tau = \langle \tau_a \rangle_{a \in E]}$ whose length is $|E|$.

$$\tau_a \;=\; \mathcal{P}(X_q = a) \tag{4.22}$$

The estimates $\hat{\Gamma}^s_{ab}$ and $\hat{\tau}^s_a$ for a single labeled recording s are built from counts of the labels and the paired transitions that are present in the training data but considered latent in test data:

$$\eta^s_a \;=\; \sum_{q=1}^{|Q^s|-1} \mathbb{I}_{\{x_q = a\}} \tag{4.23}$$

$$\hat{\tau}^s_a \;=\; \frac{\eta_a}{|Q^s| - 1} \tag{4.24}$$

$$\hat{\Gamma}^s_{ab} \;=\; \frac{1}{\eta_a} \sum_{q=2}^{|Q^s|} \mathbb{I}_{\{x_{q-1} = a\}} \mathbb{I}_{\{x_q - b\}} \tag{4.25}$$

The intermediate normalizations in (4.24) and (4.25) re-weight the estimates in order to prevent training samples with large $|Q^s|$ from dominating. The pooled sample estimates of these parameters can be found by taking the means of the computed estimates of all the labeled sequences in the training set:

$$\hat{\tau}_a \;=\; \frac{1}{|S|} \sum_{s \in S} \hat{\tau}^s_a \tag{4.26}$$

$$\hat{\Gamma}_{ab} \;=\; \frac{1}{|S|} \sum_{s \in S} \hat{\Gamma}^s_{ab} \tag{4.27}$$

4.2.3 Parameters of the observation model

The probability density used for the observation model depends greatly on the nature of the observed data itself. If the variables Y_t are categorical and an arrival can only influence the observations in the time step in which it occurs, then a stochastic matrix can be used to implement $\mathcal{P}(\mathbf{Y}|\mathbf{X}, \mathbf{T})$ in the same way that it is used to implement the conditional probability tables of a hidden Markov model. Since the planned application of this framework is to learn to label signal data, we concentrate on situations where the original input is a continuous waveform captured by uniform sampling, and is transformed via one of many "direct" approaches[2]. The outputs of this transformation are

[2]Making use of convolution, linear filtration, or other computations that avoid sampling and/or optimization.

typically the coefficients of a time-frequency transform such as the windowed Fourier transform. It is natural to use the absolute value or the product of each coefficient with its conjugate in order to obtain a feature that is real-valued. In both of the previous cases, the resulting features will be positive $Y_t \in \mathbb{R}_+^M$.

Training of the observation model is shown with the multivariate exponential distribution to describe the conditional density of the observed values on an arrival's label value. This distributional model should be considered for illustration purposes only - it has several shortcomings when applied to feature representations derived from signals. Specifically, the multivariate exponential treats each feature as an independent exponential random variable with its own parameter θ_i. These parameters are then organized into a vector θ. This is an unrealistic modeling assumption given the fact that the features for neighbouring frequencies do tend to correlate strongly in a harmonic representation, and especially if the signal exhibits a true "source" frequency that falls between two of the neighbouring frequency values that were used to construct the basis. Better choices for the observation model include the log-Normal distribution and the asymmetric Laplace distribution, although the best choice of observational model will be domain dependent.

In the experiments detailed in section 6, the asymmetric Laplace distribution of Kotz [28] is used to model $\mathcal{P}(\mathbf{Y}_{\mathcal{T}_q}|\mathbf{T}_q, \mathbf{X}_q)$.

The parameters for the multivariate exponential can be learned separately for each label. Consider all arrivals with a given label value, say a, and for each arrival consider a window of arbitrary size $2w$ centered at the arrival time. The window of observed values is referred to by $y_{\boldsymbol{\mathcal{T}}_q} \in \mathbb{R}_+^{2w \times M}$ and is embedded in the time-feature space. The log-likelihood function for $|Q|$ events is written:

$$\begin{aligned} &\text{for } \theta \in \mathbb{R}_+^{2wM} \qquad \text{let } \lambda_q = \triangle(y_{\boldsymbol{\mathcal{T}}_q}) \\ &\mathcal{L}(\theta) = \sum_{q \in Q} \left[-\sum_{i}^{2wM} \ln(\theta_i) \right] - \left(\frac{1}{\theta} \cdot \lambda_q \right) \end{aligned} \tag{4.28}$$

where $\triangle(y_{\boldsymbol{\mathcal{T}}_q})$ is the vectorization of a window matrix. Inside the exponent, the dot product reflects the product of several exponential functions of each independent variable in λ_q.

Again, we invoke the sample-weighting likelihood technique of eqn.(4.15) and maximize the log-likelihood according to eqn.(4.28) with respect to the parameter vector θ:

$$\hat{\theta}_a = \arg\max_{\theta} \sum_{s \in S} \frac{1}{|Q^s|} \sum_{q \in Q} -\ln(\theta_i) - \frac{1}{\theta} \cdot \lambda_q \tag{4.29}$$

The treatment of the multivariate log-Normal log-density follows similarly but has two parameters instead of one. Again, there is one set of such parameters for each label type. (i.e.: $x_q = a$)

$$\langle\ \theta\ \in \mathbb{R}^M\ ,\quad \Sigma \in \mathbb{R}^{M\times M}\ \rangle := \text{ the mean and covariance parameters}$$

$$\mathcal{L}(\theta, \Sigma) = \sum_{q\in Q} -\ln(\mathcal{Z}) - \frac{1}{2}(\ln\lambda_q - \theta)^\top \Sigma^{-1}(\ln\lambda_q - \theta) \tag{4.30}$$

$$\langle\hat{\theta}_a, \hat{\Sigma}_a\rangle = \arg\max_{\theta,\Sigma} \sum_{s\in S} \frac{1}{|Q^s|} \sum_{q\in Q^s} -\ln(\mathcal{Z}) - \frac{1}{2}(\ln\lambda_q - \theta)^\top \Sigma^{-1}(\ln\lambda_q - \theta) \tag{4.31}$$

$$\text{where } \mathcal{Z} = \sqrt{2\pi}|\Sigma|^{1/2}$$

A note on the multivariate exponential and deterministic label sequences. The multivariate exponential can exhibit unwanted effects if not used with appropriate forethought. For example, consider computing the likelihood of some arrival whose label is known to be $x_q = a$. Assume that there are two observational windows for which to compute the likelihood, one with medium-large feature values, and another in which all the feature values are zero or close to zero. In this case, the latter will tend to produce a higher likelihood for an arrival of any given type because most of the probability mass for this density is concentrated close to zero. Note that this is true *regardless of the value of the density's scale parameter.* The multivariate exponential will always produce large likelihood values for time steps with relative silence because of the large probability mass near zero for this density class. This means that, in selecting from a large number of time steps for which an arrival may have occurred, an estimation procedure may tend to favour time steps with low feature values - which is often not the intended result.

In general, the multivariate log-Normal is recommended in favour of the multivariate exponential for reasons given above.

4.3 Inference of the true point configuration

The intended use case for the overall framework is to produce point labels for new recordings in order to match the timings of the training set point labels and their match to the observed signal data in time-feature space. The formal goal is to accept a sequence of observed feature data and to produce a finite chain of labels Q which has the format $\langle x_q, t_q\rangle$ with the same semantics as the point-labeled recordings used for training data. Below, a method is provided that produces the labeling with the maximum probability over the space of all possible label chains. This labeling can be considered the maximum likelihood estimate of the latent chain, or as the maximum a posteriori estimate of the latent chain given an initial arrival distribution which acts as a prior.

4.3.1 Maximization of the arrival chain probability

The optimization solved by chain estimation is:

$$\arg\max_{Q,x_Q,t_Q} \mathcal{P}(\mathbf{X}, \mathbf{T} \mid \mathbf{Y}) \tag{4.32}$$

where x_Q, t_Q denote a full set of marks and arrival times for Q, which is a sequence of arrivals of unobserved length. The law of conditional probability shows that the same solution can be found by dropping the partition function since $\mathcal{P}(\mathbf{Y})$ is constant in the expression:

$$\mathcal{P}(\mathbf{X}, \mathbf{T} \mid \mathbf{Y}) = \frac{\mathcal{P}(\mathbf{X}, \mathbf{T}, \mathbf{Y})}{\mathcal{P}(\mathbf{Y})} \tag{4.33}$$

$$\arg\max_{x_Q,t_Q} \mathcal{P}(\mathbf{X}, \mathbf{T} \mid \mathbf{Y}) = \arg\max_{x_Q,t_Q} \mathcal{P}(\mathbf{X}, \mathbf{T}, \mathbf{Y}) \tag{4.34}$$

The problem can be stated in additive terms by instead maximizing the logarithm of the probability in eqn.(4.34).

The steps to computing the maximum probability chain for a sample of length L and M features for $|E|$ mark types:

1. Calculate the sample mean vector $\hat{\mu} = \langle \frac{1}{D} \sum_{k=1}^{D} x_{i,k} \rangle_i$
2. Calculate the biased sample covariance matrix:
   $\hat{S} = \langle \frac{1}{D} \sum_{k=1}^{D} (x_{i,k} - \hat{\mu}_i)(x_{j,k} - \hat{\mu}_j) \rangle_{ij}$
3. Compute the inverse of the sample covariance matrix $\hat{S}^{-1}$. This takes $O((2wM)^3)$
4. Compute the likelihood of each of $|E|$ label types for every time step in a sequence of length L. Note that the observed feature values are of size $2wM$ and that, under the Log-Normal model, all pairs of these variables must be multiplied. This step takes $O((2wM)^2 L|E|)$ time.
5. Construct a memoization array $\Phi[t, a]$ upon which to implement a dynamic programming solution. (The semantics of this array are discussed in section 4.3.3). The size of the array is in $O(L|E|)$. In order to compute the score for each time step and label, one must search backward over some segment of the history preceding a given point to find the last most likely label. This search phase is computed in $O(HL|E|)$ time. Here, H is the length of the history in which to look back for the previous arrival.
6. Find the optimal chain by tracing the pointers in the dynamic programming array backward to the start of the sequence. This involves first search over a history window near the end of the recording for the best-scoring endpoint of the chain, then following the backward links from this point in order to recover the estimate of the best chain. This is computed in $O(L + H|E|)$ time.

The sample covariance matrix must be positive definite in order to be invertible. The conditions for positive definiteness are shown in the following derivation. Assume that Z is a table of real-valued data with i.i.d. samples in the columns and features in the rows. For simplicity, let the sample mean of the data be equal to zero. The sample covariance matrix is then $\hat{S} = \frac{1}{D}\sum_{k=1}^{D} ZZ^\top$. This matrix is positive definite if $\forall x \in \mathbb{R}^N : x^\top ZZ^\top x > 0$. This condition is satisfied if $\text{null}(Z^\top) = 0$.

Overall, steps 4-6 bear strong resemblence to the Viterbi [47] algorithm, though neither that model nor the current one takes into account the unaffected observed variables $\mathcal{P}(\mathbf{Y}_{\boldsymbol{\mathcal{T}}})$ which may fall between any pair of arrivals. The main difference between the Viterbi algorithm and the one given by steps 4-6 is that here the algorithm must search over a larger history of the sequence whereas the Viterbi algorithm only needs to search over a history of length H. However, incorporation of the unaffected/observed variables can be performed additional step. See the bottom of subsection 4.3.3.

4.3.2 Computation of likelihood for arrival times and label values

Let $\mathcal{L}[t, a]$ represent a table of size $(Ls) \times (N)$, where each cell computes the log-likelihood of there existing an arrival of type x at time t based on the trained observation model:

$$\begin{aligned} [\text{ Recall that: } \lambda_t &= \triangle(y_{\boldsymbol{\mathcal{T}}_t}) \text{]} \\ \mathcal{L}[t, a] &= \ln \mathcal{P}(\triangle(Y_{\boldsymbol{\mathcal{T}}_t}) = \lambda_t \;\big|\; \exists q : X_q = a, T_q = t) \end{aligned} \tag{4.35}$$

Consider the implications of calculating the log-likelihood for a single entry in this table using the log-Normal density function:

$$\mathcal{L}[t, a] = \ln\left(\frac{1}{\sqrt{2\pi}|\Sigma_a|^{1/2}}\right) - (\ln \lambda_t - \mu_a)^\top \Sigma_a^{-1} (\ln \lambda_t - \mu_a) \tag{4.36}$$

The log-Normal observation model requires every pair of elements in the vector $\ln(\lambda_t) \in \mathbb{R}_+^{2wM}$ to be multiplied in order to compute the factor inside the braces, requiring $O((2wM)^2)$ operations. If only the diagonal elements are used instead of training the full covariance matrix, one arrives at a "naïve" multivariate log-Normal model that can be computed in only $O(2wM)$ operations.

4.3.3 Dynamic programming

If the likelihood of each label type is computed for some[3] points in the time domain, it is possible to recover an estimate of the arrival chain using a

[3]It can be computed at every time step or only at selected candidate points. See the experimentation section for details on how these ideas were applied.

variation on the classical Viterbi algorithm [47] for decoding the states of a hidden Markov model.

Consider that the optimization problem in eqn.(4.34) can be decomposed so that a solution can be written as a solution to a subproblem plus some additional information. To see that this can provide a recursive solution to the optimization, consider that the likelihood derived from the full joint distribution in eqn.(4.13) can be converted to a log-likelihood expression which provides an additive objective function:

$$\Phi \;=\; \ln \mathcal{P}(\overline{\mathbf{Y}_{\boldsymbol{\mathcal{T}}}}) + \tag{4.37}$$

$$\left[\sum_{q\in Q} \ln \mathcal{P}(X_q \;\middle|\; X_{q-1}) \;+ \ln \mathcal{P}(T_q \;\middle|\; T_{q-1}, X_q, X_{q-1}) \;+ \ln \mathcal{P}(Y_{\boldsymbol{\mathcal{T}}_q} \;\middle|\; X_q, T_q)\right] \;+$$

$$\left(\sum_{b} \ln \mathcal{P}(T_{e'} > L \;\middle|\; T_e, X_e, X_{e'} = b) \;+ \ln \mathcal{P}(X_{e'} = b \,\middle|\, X_e = a)\right) \tag{4.38}$$

The goal is to find the best chain by solving $\Phi^* = \max\ \Phi$ over the space of configurations, $\mathbf{X}, \mathbf{T}$.

Consider that any configuration that optimizes the objective Φ for some observations over $[0, t)$ and whose final arrival occurs at $T_q = t$ must also be optimal for the observations over $[0, t')$ for the arrival subsequence that is one shorter, i.e.: $t > t' = T_{q-1}$. With this assumption in mind, the likelihood of the optimal chain over $[0, t)$ can be written in terms of that over $[0, t')$.

A recursive expansion of Φ can now be provided. Denote by $\Phi[t, b]$ the objective for the subproblem on the timespan $[0, t)$ with the assumption that there exists an arrival with label $X_q = b$. To solve the subproblem for a given time step and label, we maximize over all former arrivals of type $X_{q-1} = a$ in the subsequence.

$$\Phi[t, b] \;=\; \max_{t'<t, a} \Phi[t', a] \;+\; \Psi[t', a, t, b] \tag{4.39}$$

where $\Psi[t', a, t, b]$ is the incremental term linking the smaller subproblem with the larger, and represents the log-likelihood associated with the waiting time $t - t'$:

$$\begin{aligned}\Psi[t', a] = \ln \mathcal{P}(X_q = b \;\middle|\; X_{q-1} = a) \\ +\; \ln \mathcal{P}(T_q = t \;\middle|\; T_{q-1} = t', X_{q-1} = a, X_q = b) \\ +\; \ln \mathcal{P}(Y_{\boldsymbol{\mathcal{T}}_q} \;\middle|\; X_q = b, T_q = t) \;+ \ln \mathcal{P}(\overline{\mathbf{Y}}(t', t, w))\end{aligned} \tag{4.40}$$

There are two points to make here. First, eqn. (4.39) is different from the Viterbi algorithm in only one aspect: the previous step of the relevant sequence need not be $t' = t - 1$. Rather, this optimization allows t' to range over an arbitrarily long history prior to t. Second, the observations $\overline{\mathbf{Y}}(t', t, w)$ refer to those that arrive between t, t' and are not w-proximal to either point:

$$\overline{\mathbf{Y}}(t_1, t_2, w) \;=\; \{Y_t \;\big|\; t_1 + w < t < t_2 - w\} \tag{4.41}$$

The overall expression in eqn.(4.39) leaves out only one term (factor) from eqn.(4.37), which is the corrective factor for the end-gap from eqn.(4.10). This term is added to the end in the recursive definition of the total dynamic programming objective:

$$\Phi^* \;=\; \max_{t,a} \Phi[t, a] + \Upsilon^{(e)}[t, a] \tag{4.42}$$

$$\begin{aligned} \text{where}\;\; \Upsilon^{(e)}[t, a] \;&=\; \sum_{b}^{N} \ln \mathcal{P}(T_{e'} > L \;\big|\; T_e = t, X_e = a, X_{e'} = b) + \\ &\qquad \ln \mathcal{P}(X_{e'} = b \,\big|\, X_e = a) \end{aligned} \tag{4.43}$$

The log-likelihood table $\mathcal{L}[t, b]$ can be computed directly by taking the conditional probabilities of the data under the observation model. The table $\Phi[t, b]$ can be built incrementally starting at $t = 0$ for each label type. For some short span over the beginning of the sequence, solving eqn.(4.39) will be slightly different because the time interval term must be adjusted as per eqn.(4.8) in order to account for the start-gap. In plain English, when calculating the likelihood of an arrival near the start of the sequence, the search for maxima must also consider the possibility that the arrival in consideration is the first one that occurred after the start of the chain. Denote $\Upsilon^{(s)}[t, b]$ the log-likelihood explaining the existence of a start-gap preceding some first arrival at time t with label b.

$$\begin{aligned} \Upsilon^{(s)}[t, b] \;&=\; \left[\sum_{a}^{M} \ln \mathcal{P}(T_1 = t \;\big|\; T_0 < 0, X_0 = a, X_1 = b) \;+ \ln \mathcal{P}(X_0 = a)\right] \\ &\quad +\; \ln \mathcal{P}(Y_{\boldsymbol{\mathcal{T}}_1} \;\big|\; X_1 = b, T_1 = t) \;+\; \ln \mathcal{P}(\overline{\mathbf{Y}}(0, t, w)) \end{aligned} \tag{4.44}$$

Again, we provide a revised expression for solving the subproblem that takes into consideration this special behaviour of the maximization for small values of t:

$$\Phi[t, b] \;=\; \max\{\, \max_{t' < t, a} \Phi[t', a] + \Psi[t', a, t, b]\,,\; \Upsilon^{(s)}[t, b] \,\} \tag{4.45}$$

The last three equations show that the recursive form of the full objective incorporates both the corrective factors for the start-gap and the end-gap, and that each solution for eqn.(4.39) takes into account any intermediate observations $\overline{\mathbf{Y}}(t_1, t_2, w)$ as well, ensuring that the full objective in eqn.(4.37) is maximized.

A note on computing the "in-between" observations $\ln \mathcal{P}(\overline{\mathbf{Y}}(t', t, w))$. In practice, the observations that are not influenced by any arrival might come from a wide variety of distributional models. However, if these observations are not i.i.d., then computing $\ln \mathcal{P}(\overline{\mathbf{Y}}(t', t, w))$ can be non-trivial: the overall expression is not additive in each time step in (t', t). In this case, an appropriate model must take the correlations between all intervening observed variables between t and t'. This can present a challenging modeling problem. The solution to this problem might be attempted with models of continuous and/or highly sampled stochastic models such as a Gaussian process.

Consider the difficulty of providing an observational model for these data that are uninfluenced by the arrival chain. Since the segments can be of arbitrary length, the distributional model must accomodate high dimensions and also varying dimensions, i.e.: it may have to be modeled as its own point process. If a distribution is chosen that does not properly reflect the observed data, then the term(s) of $\ln \mathcal{P}(\overline{\mathbf{Y}}(t', t, w))$ may accrue large negative values during the decoding sequence. As a result, the decoder will tend to "find" events that are as close together as possible in order to avoid the associated loss in the objective.

In addition, if these observations are not truly independent of those influenced by the arrival chain, then errors in the total log-likelihood expression can accrue due to the hard segmentation between the two sets of variables. This is particularly problematic for harmonic representations using scale as a parameter to the feature set, e.g. wavelets with a larger scale capture many more time steps than those at a smaller scale, making the segmentation between observed and unobserved variables all the more indistinct.

If the distributions of these observations differs significantly from their conditional distribution given a mark type, then often this term does not contribute meaningfully to the objective. This will often be the case when the mark points designate some interesting pattern in the signal data while the waiting times are characterized by periods of silence and/or baseline noise. In these cases, this term can be dropped from the objective to arrive at an approximation to the true solution to the maximum log-likelihood.

Having stated this, the log-likelihood of these in-between observations can sometimes be neglected in practice as is detailed in the following note.

4.3.4 Proof of Convergence

Given the optimization program:

$$\Phi^* \;=\; \max_{t,a} \Phi[t, a] + \Upsilon^{(e)}[t, a] \tag{4.46}$$

with the recursive definition:

$$\Phi[t,b] \;=\; \max\{\, \max_{t'<t,a} \Phi[t',a] + \Psi[t',a]\,,\; \Upsilon^{(s)}[t,b] \,\} \tag{4.47}$$

Let Q be the chain of arrivals whose times and labels are those produced by dynamic programming using the values given by $\Phi[t,b]$ as memoization for the scores of the subproblems, and the terms $\Psi[t',a,t,b]$, $\Upsilon^{(s)}[t,b]$, and $\Upsilon^{(e)}[t,a]$ as score increments for the problem substructure.
Claim: Q is a maximum solution to (4.46).
Proof: Assume that $\exists\, Q'$ with strictly greater Φ^* than Q with $Q \neq Q'$, so that Q is sub-optimal. This means that there exists some initial interval in the recording ending with some arrival in Q that would have been ignored in the search that solves the subproblem for some longer chain in (4.47) or for the entire chain including the end-gap in (4.46). Since all such suboptimal chains are ignored during the search, Q could not have been found by dynamic programming. This results in a contradiction.

4.4 Application

We've now described a probability model, its learning procedures and a method for estimating the best chain of latent points from a sequence of observable data. This model can now be applied to heart sounds by supposing the existence of two heartbeat types $E = \{\mathrm{S1}, \mathrm{S2}\}$ which alternate deterministically and whose times must be estimated in new recordings. The training procedures described in section 4.2.2 will not be necessary because the ordering of the heartbeat labels is always alternating and is strictly deterministic. Also, we will see in the following section that the training procedure described by section 4.2.1 for the waiting time model does not suffice when the distribution of a specific patient does not match that of the overall population or of the training data that represents the population.

Chapter 5

Modeling waiting times using signal filtering

In the treatment of parameter estimation in the point process model outlined in the previous chapter, the parameters of the conditional probability densities $\mathcal{P}(\mathbf{T}|\mathbf{X})$ are trained from the sample-weighted likelihood over the entire training set. This approach produces robust estimators of the distribution's parameters *if all recordings in the population obey the same distribution for the waiting times.* This is an unlikely occurrence in biological signal analysis; individual subjects may exhibit considerable differences in the timing and "rhythm" of the arrivals (heartbeats and neuronal firings).

For example, individual heart sound recordings may exhibit a waiting time distribution that is substantially more precise than that of the overall population. The cardiac period of a healthy adult can be easily double or triple that of an infant. The use of a pooled model for the waiting times between the beats would produce poor results when estimating the arrival sequence of heartbeats. The intra-subject variance in the cardiac period is likely to be quite low due to the heart's natural pacemaker, while inter-subject variance found in the greater clinical population will be much higher. Consequently, individual parameters for each subject must be in order for inference to be accurate. This scenario presents challenges for estimation of the arrival sequence.

When both the parameters and the arrival chain are treated as unknowns, both must be estimated simultaneously. In order to find maximum likelihood estimates, this leads to an optimization that is intractable: the best sequence of arrivals must maximize probability under the timing model, and the parameters of the timing model are dependent on the arrival times themselves - a circular dependency. This produces a difficult optimization program:

$$\max_{\theta,\mathbf{X},\mathbf{T}} \mathcal{P}(\mathbf{X}, \mathbf{T} \mid \mathbf{Y}, \theta) \tag{5.1}$$

The main offering of this section is an approximate inference method that is designed to work around this difficulty by estimating parameters of the waiting time density functions. Note that this is more difficult for an unlabeled test recording than a labeled recording because the arrival times are not known and are co-dependent on these waiting time parameters. This is done by creating a separate feature representation of the recorded signal, hereafter called an *intensity map*. Using this intensity map representation the probability density for the waiting times can be approximated reliably without knowing the actual arrival times for a given recording.

5.1 Repeating signals and alternating labels

In heart sound identification, the sequence of arrivals $\mathcal{P}(X_q \,|\, X_{q-1})$ is entirely deterministic although the waiting times between these arrivals are stochastic. Recall that the waiting time $Z_q = T_{q+1} - T_q$ is often easier to model than the times T_q themselves. In this case, the task is to model the waiting times $\mathcal{P}(Z_q \,|\, X_{q-1}, X_q)$ for each pair of labels X_{q-1}, X_q and to do this before estimating the times of the labels themselves.

If each arrival produces observations that are similar to those of neighbouring arrivals of the same label, one can design filters that respond to patterns of "output" (here, the Y_t variables) separated by fixed intervals in time. The current approach uses a filter designed to produce a response proportional to the intensity of repetition at two fixed length intervals. A filter bank can then be constructed so that there is one filter for each combination of interval lengths in some bounded set. This filter bank can be used to implement an operator which takes the time-dependent feature representation to a new time-invariant representation that contains information about the interval lengths found in the recording.

A single filter in the filter bank is determined by a selection of interval width parameters $\langle \alpha, \beta \rangle$.

$$\text{let } \mathbb{I}_{\{x\}} = \begin{cases} 1 \text{ if } x = 0 \\ 0 \text{ otherwise} \end{cases} \tag{5.2}$$

$$h_{\alpha\beta}(\tau) = \mathbb{I}_{\{\tau\}} + \mathbb{I}_{\{\tau-\alpha\}} + \mathbb{I}_{\{\tau-\beta\}} + \mathbb{I}_{\{\tau-\alpha-\beta\}} \tag{5.3}$$

Using an arbitrary time-dependent feature $f[t]$, (for example the waveform signal itself) the response can be written:

$$H[\alpha, \beta] = \frac{1}{L-\alpha-\beta} \sum_{t=1}^{L-\alpha-\beta} \prod_{\tau} f[t+\tau]^{h_{\alpha\beta}(\tau)} \tag{5.4}$$

$$= \frac{1}{L-\alpha-\beta} \sum_{t=1}^{L-\alpha-\beta} f[t] f[t+\alpha] f[t+\beta] f[t+\alpha+\beta] \tag{5.5}$$

For an alternating label sequence of only two label types, there are three intervals of interest: the waiting time between an arrival of type A followed by one of type B, the waiting time between an arrival of type B followed by an arrival of type A, and the time between two arrivals of the same mark type which are punctuated by one of the opposing type. These are depicted in figure 5.1.

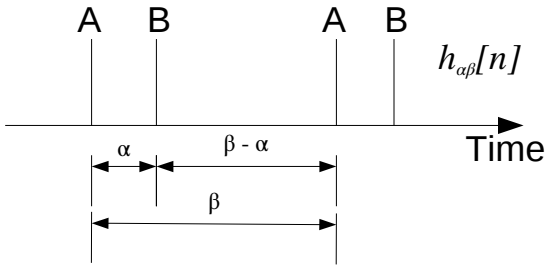


Figure 5.1: Parameterization of a quartic filter. The intervals of interest can be captured by two parameters.

In practice, this quartic filter can be further improved upon by making use of a time-dependent *feature vector* rather than a scalar feature. This feature vector represents observed signal data. For example, let $y_{[t,\cdot]}$ be a vector of time-frequency features corresponding to time t, where $y_{[t,k]}$ refers to the kth scalar feature. Then eqn. (5.4) can be rewritten:

$$H'[\alpha,\beta] \;=\; \frac{1}{L-\alpha-\beta}\sum_{k=1}^{F}\sum_{t=1}^{L-\alpha-\beta}\nu[t,k]\,\nu[t+\beta,k]\,\nu[t+\alpha,k]\,\nu[t+\alpha+\beta,k] \tag{5.6}$$

5.2 Conversion of filter output to probability density

The filtration operator developed in the previous section can be used to provide an estimate of the density function given in eqn.(4.6). The probability model is extended to include the parameters of the waiting time distribution as separate variables:

$$\mathcal{P}(\mathbf{Y},\mathbf{X},\mathbf{T},\boldsymbol{\theta}_{\mathbf{T}}) \tag{5.7}$$

Here, $\boldsymbol{\theta}_{\mathbf{T}}$ represents a set of parameters that are specific to a particular patient or to a given recording. This model admits a factorization similar to that given in eqn.(4.4):

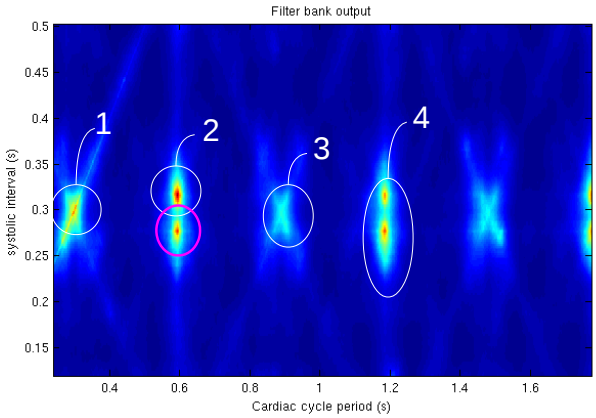
$$\mathcal{P}(\mathbf{T},\boldsymbol{\theta}_{\mathbf{T}}|\mathbf{X}) \;=\; \mathcal{P}(\mathbf{T}|\boldsymbol{\theta}_{\mathbf{T}},\mathbf{X})\,\mathcal{P}(\boldsymbol{\theta}_{\mathbf{T}}) \tag{5.8}$$

By taking the parameters of this model into account as random variables, it is possible to associate timing properties of the arrival sequence with a specific recording or subject. The central idea is to scale and normalize the filter bank responses $f[\alpha, \beta]$ so that their values integrate to 1 using a suitable measure, i.e. $\iint f[\alpha, \beta]\mu(d\alpha, d\beta) = 1$ allowing them to serve as a probability density function.

A bounded set of parameter values are discretized ("gridded") to create the domain for a function approximator whose values will be stored in a large array ("intensity map"). In the case of heart sound data, the cardiac period and the systolic interval are used to form the grid. The third quantity, the diastolic interval, can be computed by a selection of the two former parameters. This is depicted in figure 5.2.

One of the obstacles met by this filteration approach is that a repeating input pattern can produce a strong response for many combinations of input that are dissimilar to the "true" pattern. When a single referent goes by a true identity and several false identities, we refer to the false identities as *aliases*. This is a different use of the word than that used to describe the effect of undersampling a signal (i.e.: sampling at less than twice the rate of its highest non-zero frequency band).

We use this term "aliases" here to describe strong responses to a particular input pattern at false values of the parameters. Usually, this occurs because the input pattern assumes that no beats fall in between the intervals defined by α and β. However, when a filter is configured with values of these parameters at integer multiples of the true parameters, these filters will often respond strongly. This produces "echo" patterns in the intensity map. These echoes tend to occur at constant multiples of the true parameter values.



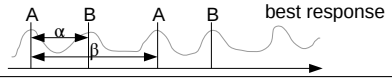
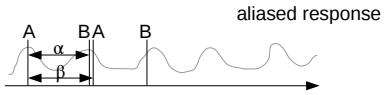
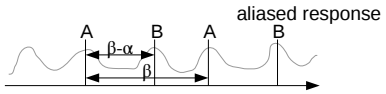
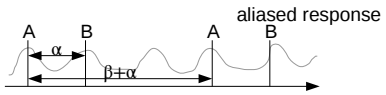
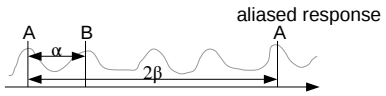
Label	Description	Illustration (input signal in gray)
pink	This is the optimal alignment of filter parameters for the example waveform in grey.	
1	This effect occurs when the systolic interval is close to equal the diastolic interval.	
2	This effect occurs when the true parameters $\langle \alpha, \beta \rangle$ produce a response at $\langle \beta - \alpha, \beta \rangle$. The cardiac period is accurate, but the systolic and diastolic intervals are reversed. This artifact is always directly above the optimal parameter value and tends to vertically mirror the response about the line $2\alpha = \beta$.	
3	This effect occurs when the systolic interval parameter is correct, but the cardiac period is off by $\beta' = \beta \pm \alpha$. The first pair of impulse responses align cleanly with the two beats of a cardiac cycle, but the second pair are mismatched. This artifact is more intense when the systolic interval and diastolic interval are close to equal, or $2\alpha \approx \beta$.	
4	This effect occurs when the filter is configured so that the cardiac period is exactly twice the optimal parameter value, or $\beta' = 2\beta$. In this case, the filter is aligned with two complete heartbeats that are separated from each other by an entire period instead of being truly sequential.	

Figure 5.2: This example shows that a few symmetries in the rhythm produce aliases in the intensity map. The artifacts are labeled and the causes described in the table. The true interval values fall approximately in the region of the pink circle in the intensity map. Further artifacts can be seen toward the right of the intensity plot as instances of these examples for larger integer multiples.

5.3 Anti-aliasing of the filtered output

5.3.1 Method 1: Lateral inhibition

In order to produce an accurate distribution of the waiting times from the intensity map depicted in figure 5.2, the "alias" artifacts must be removed or suppressed from the intensity map before it is normalized. The main idea is to treat each response point in the intensity map as though it were that of the optimal parameter values, and then to subtract this response away from any other point in the map that can be interpreted as an alias of these parameters. Thus, the larger values in the map tend to "inhibit" the smaller values, and the pattern of inhibition is determined by the aliasing patterns described heretofore.

The removal of the aliasing artifacts in f can be expressed as:

$$f'[\alpha, \beta] \;=\; \max(0, f[\alpha, \beta] - g[\alpha, \beta]) \tag{5.9}$$

Here g is called the *alias map*. It is computed from the values of f in accordance with valid aliasing patterns such as those in figure 5.2. These are described algorithmically below. Once g has been computed, it can be subtracted from f and the result thresholded at zero to produce a new non-negative intensity map f'.

Let $\mathcal{D}$ be the set of grid points of $\langle \alpha, \beta \rangle$ which forms the domain of the intensity map. Then the anti-aliasing procedure is as follows:

The parameter C specifies an arbitrary convolution kernel that is used to smooth the alias map in the penultimate step. This smoothing step is essential because the indices used in the innermost loop are formed by integer multiples; the output values of g in this step are thus strided at $2\times, 3\times, ...$ and so on. The parameter H is meant to limit the number of multiples in the cardiac period domain. In our experiments, the value used was $H = 3$. The constant λ is a value large enough to "zero out" the values of F for which the systolic interval is much larger than the diastolic interval (by some threshold ϵ). This constant should be equal to or greater than the peak magnitude of the convolution kernel in order to ensure that the resulting map is zero over this range of parameter values.

An illustration of the lateral inhibition method is given below in figure 5.3. The result of subtracting the aliased points from the intensity map can be visualized in figure 5.3.

Algorithm 2 $[\ f'\ ] =$ `lateral_inhibition`$(f;\, C, H, k)$

$g \leftarrow \emptyset^{\mathrm{Domain}(\alpha,\beta)}$
for all $\langle \alpha, \beta \rangle \in \mathcal{D}$ **do**
  if $2\alpha > \beta + \epsilon$ **then**
    $g[\alpha, \beta] \leftarrow \lambda f[\alpha, \beta]$
    continue
  end if
  for all $\langle p, q \rangle \in [0 : H] \times [0 : 1]$ **do**
    $c \leftarrow q\alpha + p\beta$
    for all $\langle m, n \rangle \in [0 : 1] \times [-1 : 1]$ **do**
      $s \leftarrow n\alpha + m\beta$
      if $\langle c, s \rangle \in \mathcal{D}$ **then**
        $g[c, s] \leftarrow \left(g[c, s]^k + f[\alpha, \beta]^k\right)^{\frac{1}{k}}$
      end if
    end for
  end for
end for
$g \leftarrow g \bigotimes C$
$g \leftarrow g^2$
$f' \leftarrow f - g$
for all $\langle s, c \rangle \in \mathcal{D}$ **do**
  if $c > 2s$ **then**
    $f'[c, s] = 0$
  end if
end for

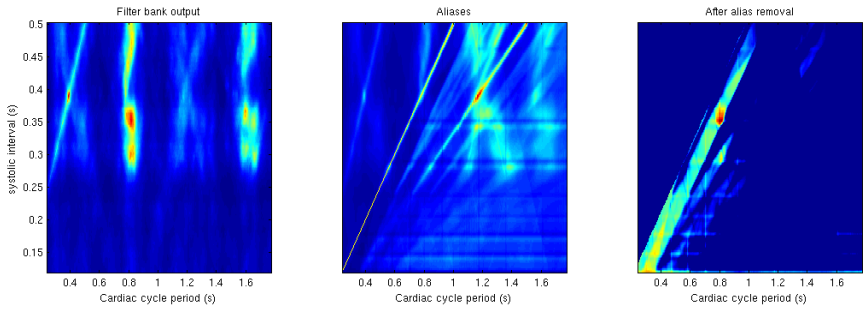


Figure 5.3: The lateral inhibition method. The cardiac period width is on the abscissa, while the systolic interval width is on the ordinate axis. In the left plot is the original intensity map f produced by the filter bank. The middle plot is the map g which contains the image of the aliased points. The right plot shows the difference map $f - g$.

One of the drawbacks with this method is that it tends not to remove enough of the "mass" of the intensity map near the line $2s = c$, where the systolic and diastolic intervals are roughly half that of the cardiac period and are roughly equal with each other. This is because, in close proximity to this line, there are few other points in the domain which can cause an alias effect to occur in this region. As a result, the alias map g tends to exhibit a darkened region slightly below the line $2s = c$. The subsequent result after subtracting the alias map from the result is increased probability mass near this line, meaning that the estimator tends to favor heartbeat sequences in which the systolic interval is about equal to the diastolic interval.

Although there are likely to be further steps that one could take to improve the performance of this method, it is believed to be less robust overall. This is because of the need to set several parameter settings and the likelihood that further gains in performance due to programmatic tweaks are likely to overfit in an algorithmic or structural sense. Since no aspect of this method relies on learning, it is difficult to guarantee that an increase in performance will generalize outside of the training set.

5.3.2 Method 2: Least maximal peak

One of the drawbacks of the lateral inhibition method is that if the signal contains too much noise or high-frequency repetition, these can produce high "false" values in the filter bank response which in turn mitigate the "true" responses.

The least maximal peak method uses the signal normalization method from section 3 to determine the positions of maxima in the smoothed filter bank output, f. Once these have been localized, the maximal point with the

smallest parameter values is chosen as the best and a distribution is created by zeroing out all the responses that are sufficiently distant from this best set of parameter values. Here, "sufficiently distant" means using a pair of operating constants that determine the size of the region containing the final total probability mass. These constants are given in appendix I.

This has the effect of producing a highly localized distribution that typically has a very small number of peaks concentrated in a tight region. Empirically, this method has proven to be the most robust at identifying the proper timing of heart sounds (see section 6). A listing of this algorithm is given below. Here, the **maximal**(f) function selects all points (c, s) for which $f(c, s)$

Algorithm 3 $[\ f'\ ] =$ `least_maximal_peak`$(f, k_\sigma, d_c, d_s, \epsilon)$

$\bar{f} \leftarrow$ `mean`(f)
$f' \leftarrow$ `relative-scaling`$(f, \bar{f}, k_\sigma)$
$E =$ **maximal**(f')
$c^*, s^* = \arg\min_E 2s + c$
for all $c, s \in$ **domain**(f') **do**
  if $|c - c^*| > d_c$ **and** $|s - s^*| > d_s$ **then**
    $f'[c, s] \leftarrow 0$
  end if
end for

attains a maximal value and f has a second-order difference value (i.e. not on the borders of the map).

The drawback to this method is that it is difficult to provide a strong analytical justification for its use. However, the motivation for the method is quite simple: the intensity maps produced by filtration invariably contain several maximal points. On inspection of several intensity maps, it has been found that the true beat parameters usually have the smallest projection on the line $c = 2s$.

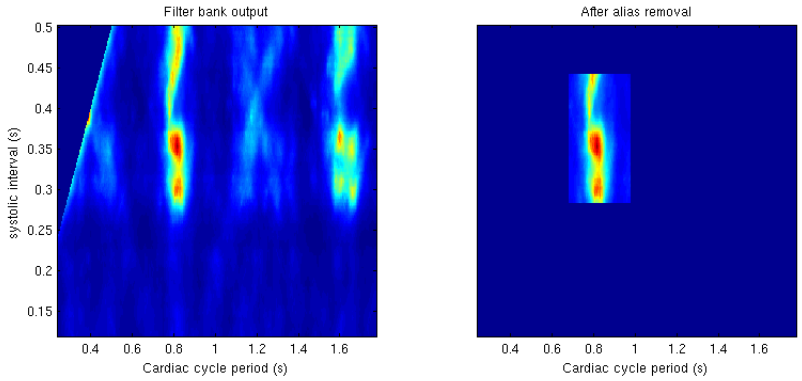


Figure 5.4: The least maximal peak method. On the left is the original input, on the right is the density of f'. Although this method introduces a strong discontinuity in the shape of a box around the peak of the distribution, the discontinuity tends to manifest itself in regions of low probability mass.

Aside from being easy to implement, the main strength of this method is that produces relatively reliable distributions for use in estimation of the arrival sequence. When the intensity map is created, the aliased responses tend to accumulate at parameter values larger than the true parameter values, with little or no aliasing artifact in the lower regions of the 2-dimensional domain. This explains why choosing the lowest of the maximal values is a good heuristic - it picks the peak that is least likely to have been contributed to by the aliasing process.

5.4 Combining maps from several recordings

It is possible to obtain a more reliable distribution by taking the product of intensity maps created by recordings taken at different chest positions on the same subject. In practice, this technique seems to negate some of the noise associated with a single chest position. See section 6.1.1 for more details.

Chapter 6

Experiments

As mentioned earlier, most research attempts which require the identification of S1 and S2 have relied heavily on the ECG signal to register the cardiac cycle. Without the ECG signal, there are relatively few methods which classify S1 from S2 based on the sound alone. In the context of those methods that have made the attempt, none look exclusively at the identification problem itself. So, it is difficult to compare our algorithm with any current state of the art.

To compare a set of four variations of the methodology presented here, a set of 35 human heart sounds were collected and annotated by a human non-expert with the two labels *S1* and *S2*. These labels stand for the noises produced by the heart valves at the start of ventricular systole and diastole, respectively. See [23, 2] for coverage on the mechanics of heart sound production. The goal of the study is to determine whether heart sound identification is possible using the available data, to determine what the major obstacles are, and to compare a few solutions to the problem that rely on probabilistic modeling.

A sample of 43 patients' heart sounds were recorded for 20 seconds using a 3M Littman stethoscope at the left sternal region (position 2L) using the Bell end of the scope. These recordings were exported and stored in waveform at 4kHz without resampling. Although the Nyquist frequency is effectively 2kHz for this sampling rate, the time-frequency features are computed at 500Hz. In addition, the frequency bands of these features are limited to between 10Hz-100Hz. Of the 43 recordings, 35 were manually labeled using an unspecified wavelet transform as a guide for detecting the precise location of the beats; the remaining 8 were either too noisy to be labeled by sight or not properly recorded. Experimenter selectivity thus added significant bias to the result and is one reason why the results should not be interpreted as clinically generalizable. Rather, this selective method was used in order to obtain a "reasonable" data set with which to learn from. The unlabeled recordings were simply too difficult for a non-expert to judge, and so the resulting classifier cannot be trusted to identify heartbeats for new recordings that similarly exhibit high

noise or variation in the heartbeat rhythm. The labelings were not validated by a clinician.

Two separate feature representations were applied in the study. The first was a simple signal energy representation, where the signal energy $|f[t]|^2$ is smoothed using convolution with a Gaussian kernel with $\sigma = 0.15$ seconds. This representation gives a good indication of the overall volume of the signal at a level of detail that is as coarse as possible while still providing enough resolution that two adjoining beats are not blurred together. The idea here is that one finds a small number of maximal points when the signal energy is blurred and that these maxima tend to encompass the sound from a whole individual heartbeat. It is relatively easy to run the Viterbi algorithm by computing likelihood for a small number of candidate points that occur at these maxima compared to computing the likelihood for the entire dense sequence. Furthermore, it was observed that these maxima tend to occur at or near the training labels.

The second representation was a set of time-frequency features computed using the pseudo Wigner-Ville distribution (PWVD) as found in [29]. The PWVD was configured with a set of 32 frequency bands chosen from 10Hz to 100Hz uniformly spaced on the natural logarithmic scale, and with a time resolution of 500Hz. The `relative_scaling` function (see section 3) was used to control for large spikes in amplitude: the parameters of the two-dimensional Gaussian kernel passed to this function included a standard deviation of $\sigma_{\text{time}} = 0.050$ in the time domain and $\sigma_{\text{freq}} = 16$ bands in the frequency domain.

There are two parts to the study. The first part is a demonstration of subject specificity in heart sounds. These illustrations show the inherent variability of heart sounds by presenting the statistics and distributions of the labeled training data. This part is intended to motivate the estimation procedures described in previous chapters. In the second part, a number of predictors are constructed using the methods proposed, and are then compared demonstrating the effectiveness of each method.

6.1 Waiting time distributions

In order to motivate the estimation of waiting time distributions for the individual, this section provides a comparison of distributions learned for the waiting times of individual training subjects versus those learned for the entire pooled sample. These distributions were selected on the basis of their being distributions over a positive random variable. They differ in terms of parameterization complexity, but also because they offer a wide range of flexibility to skewness and heavier-tailed data.

6.1.1 Configuration

To capture the waiting times conditioned on the heart sound type, $\mathcal{P}(Z_q|X_q, X_{q-1})$, the following distributions were fit to data.

Name	Parameters	Density / Likelihood
Log-Normal	μ, σ	$\frac{1}{\sqrt{2\pi}\lvert\sigma\rvert} \exp\{-\frac{(\ln x-\mu)^2}{2\sigma}\}$
Log-Laplacian	μ, σ	$\frac{1}{2\lvert\sigma\rvert} \exp\{-\frac{\lvert \ln x-\mu\rvert}{\sigma}\}$
Asymmetric Log-Laplacian [28, 46]	μ, m, σ	$\frac{2\exp\{\frac{xm}{\sigma}\}}{(2\pi)^{\frac{d}{2}}\sigma^{\frac{1}{2}}} \frac{x^{\nu}\sigma^{\frac{\nu}{2}}}{2+m^{\nu}\sigma^{\frac{\nu}{2}}} K_{\nu}\left(\sqrt{(2+\frac{m^2}{\sigma^{-1}})(\frac{y^2}{\sigma})}\right)$
Asymmetric Log-Laplacian [53]	μ, p, σ	$\frac{p(1-p)}{\sigma} \exp\{-\frac{x-\mu}{\sigma}(p - I(x \leq \mu))\}$
Gamma	α, β	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\{-\frac{x}{\beta}\}$

Table 6.1: Candidate distributions for the waiting times.

Each recording $s \in S_{\text{train}}$ contains some number of heartbeat arrivals of each type: $N_{s,c}$. In order to negate the effects of bias caused by recordings with larger numbers of beats, a desired "sample size per recording" is chosen $\hat{N}$. For every recording and for each class, the $N_{s,c}$ samples are replicated $\hat{N}$ **div** $N_{s,c}$ times, followed by drawing $\hat{N}$ **mod** $N_{s,c}$ samples uniformly at random with replacement. This procedure attempts to minimize the bias induced by random sampling in order to provide an equal number of recordings for each sample. Thus there are $(|S_{\text{train}}| \cdot \hat{N})$ data points altogether. Parameters for each of the waiting time class conditionals are then fit for each of the distributions given above, and the *normalized negative log-likelihood* is computed:

$$\mathcal{I}(\mathcal{P}, \theta; \omega) = -\frac{1}{D} \sum_{i} \log_2 \left(\mathcal{P}(x_i(\omega)|\theta)\right) \tag{6.1}$$

This statistic is used as a rough goodness-of-fit measure of the distribution, with higher values indicating more samples landing in the tails of a distribution. It can also be seen as an estimate of how robust the distribution is to outliers in the training data set.

In every case, the systolic interval and cardiac period are modeled independently, though in practice it is known that these two are correlated. If a heart sound label corresponds to the column of the data matrix and each individual datum corresponds to the row, then the rows within a column cannot be compared between columns because of the way the bootstrapping procedure resamples from existing data in order to obtain $\hat{N}$ samples for each class and subject. Thus, each column must be modeled separately as a consequence of the bootstrap method.

The estimator used to find the maximum likelihood parameters of the Kotz asymmetric log-Laplacian is given in [46]. This estimator can produce a non-positive-definite estimate of Σ (here: σ) for some data sets. The usual remedy for this deficiency is to project the intermediate matrix onto the positive definite cone. However, for univariate data this amounts to forcing σ to equal some small positive constant, which was set to 0.01.

It should be noted that PWVD features were used to compose the intensity maps $H[\alpha, \beta]$ using the dot-product filter given by eqn. (5.6), and this was true even for predictors that did not use the PWVD features for their observational model. In the results section below, the density estimation method relying on intensity maps $H[\alpha, \beta]$ are abbreviated as the "H-map" method in the tables and figures.

In order to gauge the effectiveness of the intensity map approach described in section 5, an intensity map was created for each one of the subjects over the parameter domain of the systolic intervals and the diastolic intervals. The parameters of the intensity map used are given in appendix 1. In addition to the use of the intensity map to describe the distribution for a single recording, a second set of distributions were created using the pooled intensity maps from all four chest recordings corresponding to a single subject. The "pooled" intensity map amounts to the product of the individual maps formed over each of the recordings. This technique has been observed to provide a more accurate distribution for any single recording in the set of four because it is effective at cancelling out the bias associated with a single chest position. Different chest positions are known to differentially propagate noise from the heart, whereas the heart rate and overall rhythm is assumed to remain relatively constant over all four recording sessions.

6.1.2 Results

The interval lengths for the systolic interval and the cardiac period were taken for the pooled sample and for three individuals. These data are plotted in histograms in figure 6.1.

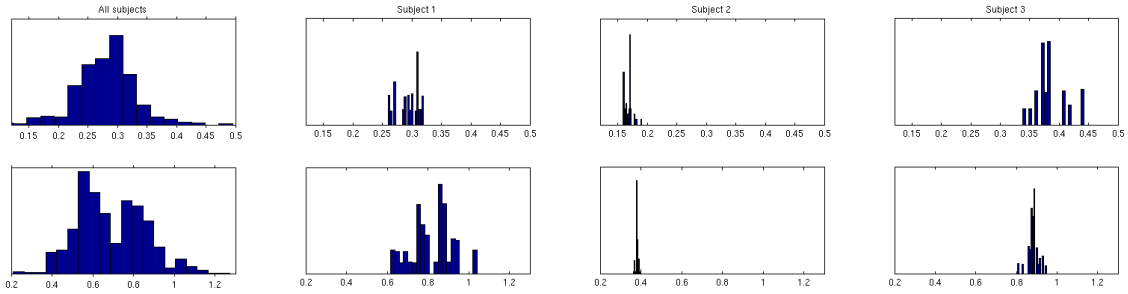


Figure 6.1: Histograms of waiting times for the pooled sample and for three individuals with 20 bins. The systolic interval is along the top row with the cardiac period along the bottom row. Note that a distinctive feature is the presence of multiple modes in the distributions of the subjects.

For a total sample size of $N = 10500$ heart sounds taken from $|S_{\text{train}}| = 35$ labeled recordings, the statistics of the pooled sample were generated. For three randomly selected training subjects, the above distributions were also fitted to $\hat{N} = 300$ data points for each subject and each class.

Densities	Pooled		Subject 1		Subject 2		Subject 3	
Log-normal	-0.33	0.47	-1.90	-0.22	-2.71	-3.04	-1.86	-2.06
Log-Laplacian	-0.41	0.62	-1.75	-0.13	-2.69	-3.00	-1.90	-2.19
Kotz Log-AL [28, 46]	0.70	1.01	-0.90	0.48	-1.32	-1.30	-0.95	-0.85
Yu Log-AL [53]	-0.48	0.57	-2.06	-0.27	-2.82	-3.19	-2.01	-2.26
Gamma	-2.22	-0.92	-3.67	-1.20	-5.28	-5.28	-3.23	-3.07
Single Recording H-map	n/a		-3.19	-0.47	4.61	4.67	-3.13	-2.73
Whole Session H-map	n/a		-3.30	-0.58	4.61	4.67	-3.05	-3.06

Figure 6.2: Normalized negative log-likelihood of selected density functions. Each major column contains a pair of minor columns giving the normalized negative log-likelihood for the systolic interval distribution and the cardiac period distribution. The rows show these scores for each model. Lower scores indicate that the distribution was a better fit for the pool or sample of data, and for the interval given. Here, the distributions created by signal filtering

The values reported in figure 6.2 were stable under the bootstrap sampler. Although the scores were observed to vary by small amounts ($< 1\%$), it was not necessary to use error bars to report the results. The pooled scores were only recorded for the parameteric distributions as the intensity maps are designed to be used specifically with either one recording or one subject's set of recordings. The bars in figure 6.3 are plotted on an arbitrary scale of scores that is intended to show the overall differences between the distributions and does not imply a baseline.

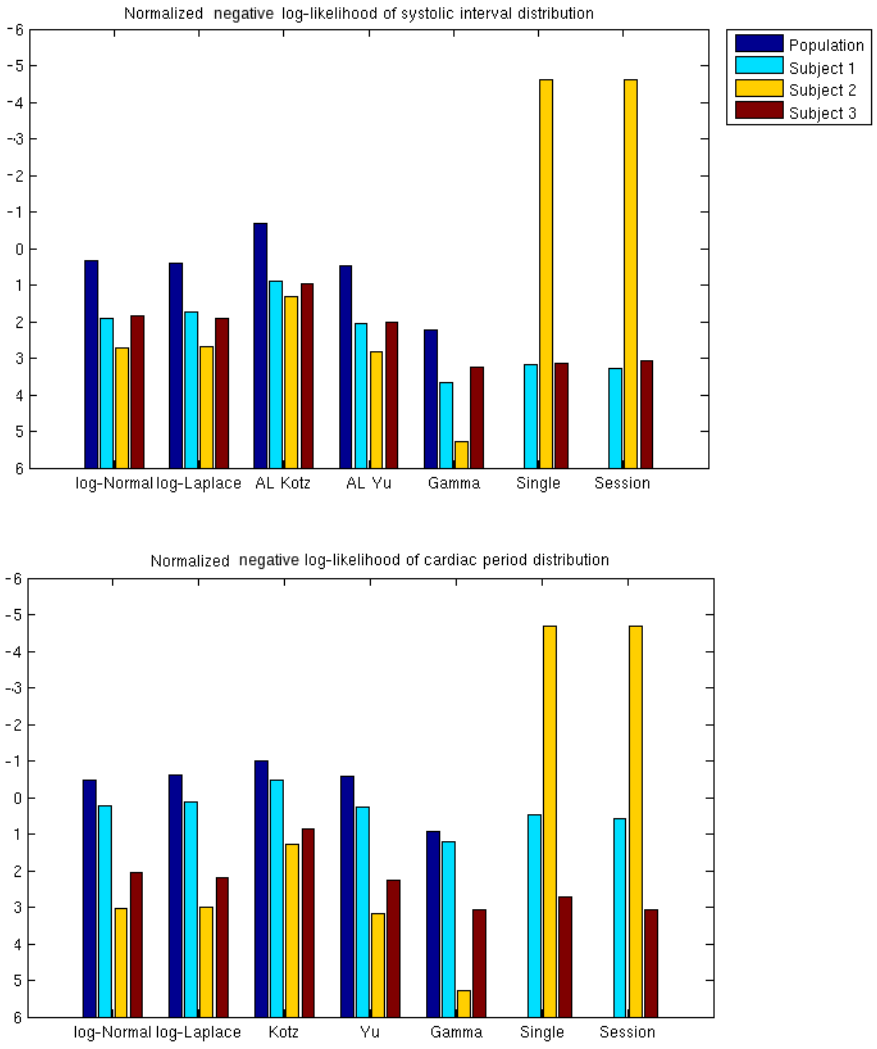


Figure 6.3: Normalized negative log-likelihood of selected density functions. Lower values in the graph indicate fewer outliers and a better fit. The best distribution fit given the training labels is the Gamma distribution. The two (nonparameteric) methods on the rightmost side of the graph obtain an average log-likelihood that is comparable with the parameteric methods on their left, although with a high variance.

6.1.3 Discussion

As far as modeling the systolic interval goes, it is clear that the Gamma distribution provided the best fit overall, though the intensity map approaches were competitive. There was more variance in the log-likelihood scores of the cardiac period model, though the Gamma was also a good fit in this category. One distinguishing feature appeared to be performance on the multimodal distributions seen in the subject histograms. It may be that this will prove to be an important feature in future efforts to model heart sounds; there are

good reasons to believe that the timing of the heartbeat may follow a bimodal distribution. For example, breathing and physical activity can be treated as bimodal processes. These activities can affect the dynamics of valvular flow by offsetting the shutting of valves and by causing splits in the heart sound as a consequence [23, 2].

One criticism of the experimental design is that in comparing distributions in terms of their likelihood, the complexity of the model (specifically, the number of its free parameters) is left out. Broadly speaking, models with greater parameteric complexity have a greater range of flexibility and thus thend to attain higher likelihood scores than models of less parameteric complexity. A better measure of the model fit might have been the Akaike Information Criterion (AIC) which contains a penalty term that directly punishes model complexity in terms of the number of its parameters.

Even in consideration of this fact, one would expect that the models with higher parameteric complexity in this study would have had more freedom to fit the data, resulting in lower negative log-likelihood scores for those models. However, this isn't what happened. The best-fitting model, the Gamma, had only two parameters compared to the three parameters of the asymmetric Log-Laplace distributions. This actually lends even more support to the Gamma as the appropriate distribution for heart sound modeling.

In comparing the intensity map methods against their counterparts, it seems that these methods are competitive with the Gamma distribution at modeling the systolic interval, and slightly worse than average when it comes to modeling the cardiac period. It is notable that the intensity map methods did poorly at modeling subject 2's data because of a specific characteristic discussed below in section 6.2.5. Here, the intensity map estimators identify the wrong distribution entirely and rely on the uniform density mixture component almost entirely. Since this component accounts for only 4% of the probability mass, the overall negative log-likelihood for this one recording is very low.

The log-likelihood values tended to be lower when fitted to subject data than for the overall population. This does not mean that the sufficient statistics of the subject are more informative than that of the overall pooled distribution, even though casual observation of the histograms may appear to support this claim. Rather, the lower scores are a result of low variance in the waiting times of a given subject, resulting in data that were easier to fit using unimodal density functions.

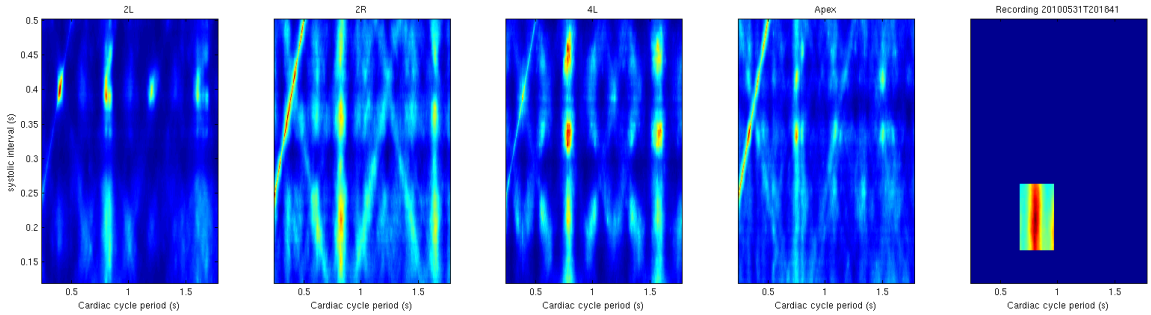


Figure 6.4: Intensity maps created for four different recordings of each of the different chest listening positions combined into a single map on the right which identifies a much stronger peak than is present in any one of the separate maps.

The difference between goodness of fit and distribution accuracy can be seen in comparing the intensity map distributions produced by a single recording versus those produced by combining the maps from all four chest listening positions. Although both classes of distributions produce roughly equivalent scores, the combination of all four listening positions seems to identify the "true" peaks of the distribution more readily (see figure 6.4).

Judging from the variability in the data alone, the histograms show that the systolic interval exhibits a fairly stable and repetitious behaviour. This behaviour is even consistent across patients, with the peak occurring consistently at 300ms. This is likely due to the heart's natural pacemaker ability, which controls the width of the systolic interval with high precision using a small network of nerve fibers which drive the cardiac rhythm and reach peak activity during ventricular systole. The distribution of the systolic interval is therefore the most useful piece of information to be gleaned from the timing of the heart when attempting to predict the arrival times of heartbeats.

6.2 Estimation of point-labeled heart sounds

6.2.1 Methods

In order to assess performance on the identification task, a "continuous" version of the precision-recall score was developed for comparing the predicted labels of the heart sound sequence with the training labels. This metric is intended to be analogous to the precision/recall scores derived from the confusion matrix of a binary classification task. Recall that each heart sound arrival is centered on a window of width $2w$ in the signal in which the arrival influences the observed feature variables. Each predicted label and training label are associated with such a window, and the total width of the overlap between these two classes of labels can be divided into either the sum of the prediction

label window widths or the sum of the training label window widths. These two ratios provide the analogues to precision and recall, respectively.

Formally, the *continuous precision* and *continuous recall* are defined in the following terms. Denote by Q_{train} and $\hat{Q}$ the set of labeled heart sound arrivals and a set of predicted arrivals, respectively.

$$W = \sum_{q'}^{\hat{Q}} \sum_{q}^{Q_{\text{train}}} \max\{\, 0,\, 2w - |t_{q'} - t_q| \,\} \tag{6.2}$$

$$U = \sum_{q'}^{\hat{Q}} w + \min\{\, w,\; t_{q'},\; L - t_{q'} \,\} \tag{6.3}$$

$$V = \sum_{q}^{Q_{\text{train}}} w + \min\{\, w,\; t_q,\; L - t_q \,\} \tag{6.4}$$

The scores are computed as:

$$P = \frac{W}{U} \tag{6.5}$$

$$R = \frac{W}{V} \tag{6.6}$$

The continuous precision can be interpreted as the percentage of correct overlap between the predicted labeled windows and the true labeled windows out of the sum of all the predicted windows. This interpretation is analogous to the precision of a classification task. The motivation for this score is that a correct labeling is not binary in the sense of a classification task. Rather, a predicted label must be compared to a true label by means of an overlap between the window of observations centered on each one.

For a given predictor and a labeled training set, the continuous precision/recall scores of a predictor form a point on the unit square. A predictor generates a two dimensional scatterplot consisting of the points of all the recordings. The scatterplot can be useful for identifying the predictor's robustness and for identifying cases or groups of cases that the predictor has difficulty with. One can use it to determine whether a predictor's accuracy is specialized toward generating more labels (emphasizing recall) or generating fewer labels (emphasizing precision). There is a clear difference between a classical PR-curve diagram and the scatterplot generated by these "continuous PR" scores: in the classical version, there is a continuous parameter being varied which generates the curve. Each point on this curve is intended to represent a unique classifier which generated the point's precision and recall. In this continuous PR score, a small finite set of classifiers are being compared. Ironically, the classical PR score for discrete labeling tasks produces

a continuous curve and our continuous PR score generates a discrete set of scores.

6.2.2 Experimental Configuration

As mentioned in the introduction of the thesis, few existing methods rely exclusively on the phonocardiogram signal to identify S1 and S2. The only method that focuses exclusively on the identification of S1 and S2 is a neural network method [22] for which the topological properties of the network (its sizes and connectivity) are not known. Furthermore, the misclassification rates cited by this study indicate that the network

Four configurations of heart sound identification predictors were created and are described below.

Predictor 1. Abbreviation: PS The distribution of both the waiting times and the observed data are pooled across the entire sample S_{train} with no subject specificity built into the model. The representation chosen for the observed data is the signal energy representation. Both distributions are modeled with the univariate asymmetric Laplacian described in [53]. Likelihoods are calculated for every time step of the sequence.

Predictor 2. Abbreviation: PP This predictor uses the same pooled distribution to model the waiting times. The observation model consists of the PWVD features captured by the multivariate asymmetric Laplace distribution described in [28] and whose parameters are also shared across the entire pooled sample. Likelihoods are also calculated for every time step of the sequence.

Predictor 3. Abbreviation: SD Here, a subject-based model is used to capture the waiting times between the heartbeats. The method outlined in section 5 is used to approximate the waiting time distribution for the specific patient before the estimates of the arrival chain have been computed. The intensity maps for all four chest positions of a given subject are combined in order to produce the distribution. The observational model uses the signal energy representation with a univariate asymmetric Laplace distribution described in [53]. Likelihoods are calculated for every time step of the sequence.

Predictor 4. Abbreviation: SP This predictor is fundamentally the same as SD, except that likelihood is now calculated only at the maximal points of the signal energy representation. That is, only the time steps that belong to the set $\{\ t\ :\ s'[t-1] \leq s'[t]$ and $s'[t+1] \leq s'[t]\ \}$ are considered, where $s'[t]$ is the convolution of the signal energy by a Gaussian kernel[1] with scale parameter $\sigma = 0.100s$. This greatly speeds up the computation but at the risk of loss in accuracy since it is not guaranteed that the maximal points in this

[1]This choice of kernel is discussed in section 6.

representation will contain the training labels. Formally, this is like including a binary feature $y_{t,0}$ which is only true at the maximal points in the smoothed signal energy and conditioning all probability mass on the truth of this feature.

6.2.3 Predictions

The relevant comparisons to be made between these configurations are as follows.

PS vs. PP: Between these configurations, the effectiveness of the PWVD representation is tested against the smoothed signal energy representation. Although the PWVD has been recommended for use in heart sound analysis before [29], it is difficult to anticipate whether the spectral signature of a given heartbeat type will be consistent enough across the entire population to be useful for labeling the beats. It is very likely that large intersubject variation will exist in the time-frequency representation of a heartbeat. If it does, it is non-trivial to capture this highly multivariate density while simultaneously estimating the positions of the beats themselves. However, the time-frequency representation might still be useful if it can be modeled by a pooled sample model, and that is what this comparison attempts to determine. *Prediction: no significant difference between PS and PP.*

PS vs. SD: These two configurations both use a similar observation model, but the PS model uses a shared parameteric density to describe the waiting times where the SD model uses a subject-specific density based upon the intensity map technique from section 5. This comparison illustrates the benefit of using the intensity map technique to capture a distribution that is unique to the individual subject. Since this technique is one of the main contributions of this thesis, the comparison is an important one. *Prediction: SD should offer better significantly better precision than PS.*

SD vs. SP: This comparison tests whether there is any drop in performance when likelihood is computed only for the sparse set of maximal points in the smoothed signal energy rather than for the entire sequence. The former technique offers a dramatic increase in computational speed since the decode phase of the dynamic programming algorithm is a major computational bottleneck for the entire predictor. When using only the maximal points as candidates for arrivals, the number of points under consideration drops by a factor of 100. If the prediction performance of the algorithm is unaffected by this change, then it is worthwhile to estimate the best chain in this way. *Prediction: no significant difference between SD and SP.*

6.2.4 Results

The plots in figure 6.5 shows a scatterplot containing the overall precision/recall scores. It is not clear if the PS or PP predictors are performing better than chance on the vast majority of the recordings, and may even be performing worse. In lieu of error bars, shaded ellipses have been drawn in the neighbourhood of the prediction scores for a given predictor, these have been calculated using the means μ^{pr} and covariance Σ^{pr} of each predictor's precision and recall. Small eigenvalues of the covariance matrix are scaled upwards using the following (given in MATLAB syntax):

$$[V, D] = \text{eig}(\Sigma) \tag{6.7}$$

$$m = \max \mathbf{diag}(D) \tag{6.8}$$

$$\Sigma' = \min(\Sigma, 0.15m) \tag{6.9}$$

Each ellipse is drawn to cover approximately 68% of the associated prediction scores for a predictor.

Due to the nature of the heart sounds alternating in sequence, it was a frequent occurrence that the predictor would get the beats mismatched with one another, e.g. predicting S2 for true S1. For this reason, the predictors all obtained continuous precision/recall scores at zero for some number of the recordings in the training data. This tended to give a skewed picture of the results since it sometimes was that the predictor was arguably identifying the primary rhythm but mismatching the beats' label types.

Because this occurrence tended to deflate a predictor's score, these "zero-score" predictions were removed and the data was re-plotted on a log scale in order to focus on differences between the predictors in the low range of the accuracy scale. This plot should be viewed with some skepticism as it omits the numerically worst output from each classifier.

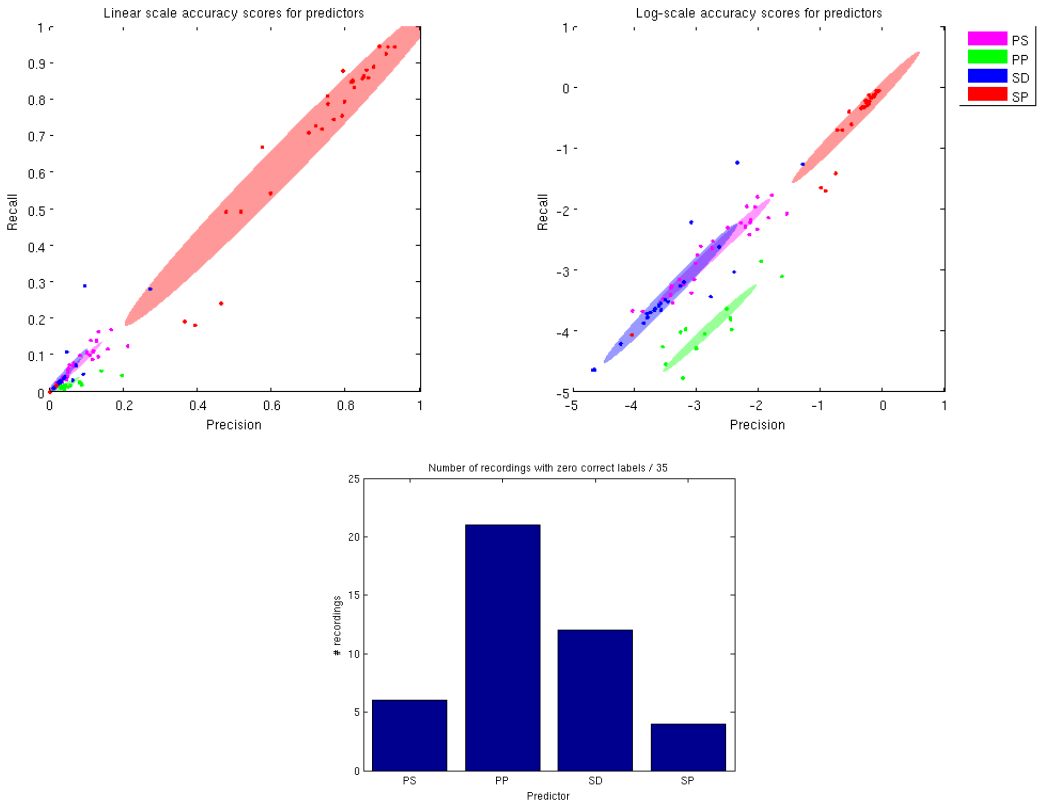


Figure 6.5: Continuous precision and recall results for four predictor configurations. Above, individual dots show training sample recordings, with ellipses drawn to indicate overall accuracy and robustness. A number of data points occur at the origin for each predictor. The log-scale plot on the right omits data points in which a predictor did not successfully predict any heart sounds. On the bottom, the number of recordings that achieved zero precision/recall score are given for each predictor based on a training sample of size $D = 35$. Results indicate that the SP combination of using the maximal points of the likelihood, subject-specific modeling, and a simple energy representation provided the best overall performance.

6.2.5 Discussion

The immediate finding is that the SP predictor dominated the prediction scores by a substantial margin. This is largely supported by the continuous precision/recall scatterplot showing the performances of each predictor on training data, but also by the low number of "zero-score" labelings which may have mismatched the labels. The second finding is that conditioning every predicted beat on a maximal point in the signal energy representation was the single most effective way to identify heart sounds in the training data. This is likely due to the fact that such time steps stand out prominently when examining the sequence visually, and make for obvious candidates for the precise arrival time of a given beat.

It is challenging to model a segment of PWVD features with a parameteric multivariate distribution. This is thought to be the main reason for the PP predictor's poor performance. As the complexity of the observed variables increases, it becomes more difficult to learn an accurate generative model.

The implications of a complex input such as the PWVD is that the observation model tends to produce log-likelihood values that are far too small. As a result, the decoding algorithm tends to avoid placing events altogether and opts to put them as far apart as possible in order to avoid paying these large penalties. This accrues "less small" log-likelihood values in the model of the arrival times and results in an arrival chain that is spaced very far apart, missing the true beats entirely. The visible effect of this is that overall fewer label predictions are emitted, and so precision is somewhat increased (see the green region of figure 6.5) while recall is diminished. This interpretation is also supported by the large number of zero scores by the PP predictor which attempts to model the PWVD features. Apart from this artifact, most of the prediction scores tended to gather near the line where precision and recall were nearly equal.

Another finding was that in comparing PS with SD, the use of the intensity map approach to density estimation did not significantly improve performance and may have even hurt performance for these configurations. This suggests that there may still be room for the pooled density estimate to be successful if coupled with other techniques such as that used by SP to only compute likelihood for peak values in the smoothed signal energy.

The success of the density estimation method of section 5 does come with some clear drawbacks. The method is constrained to a domain of heart sound parameters that may not contain the entire population's heart sounds. This might be remedied by enlarging the domain of the intensity map, though it would best be done with more data or clinical input to inform the selection of the domain. The bounds of this map are given in appendix 1.

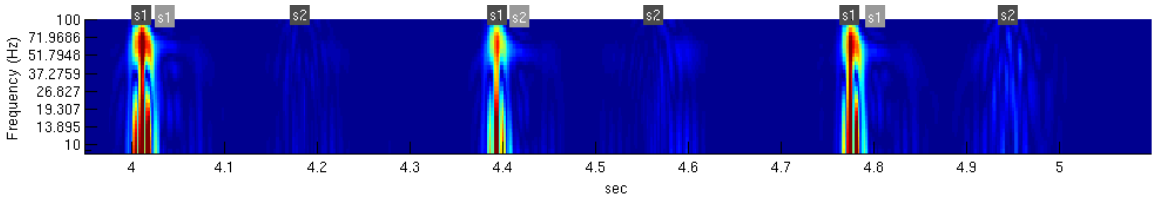


Figure 6.6: The output of the SP predictor. A strong dominant S1 beat coupled with a muted S2 beat, together beating at over 150bp/min. Dark shaded labels denote training labels while lighter shaded labels denote predictions. In this situation, the SP predictor models only the S1 beats when identifying the relevant waiting time distributions, and may not even compute likelihood for the faint regions where the true S2 occurs.

In figure 6.6, the predictions of the SP predictor and the training labels are overlaid on top of a plot of the PWVD features. In this case, the algorithm has identified the wrong distribution for the waiting times of the arrivals due to its use of the quartic filter used to build the intensity map. When the subject's heart sounds are extremely polarized, consisting of one very loud beat, one very soft beat, and a very high heart rate, the softer beat does not produce a noticeable response to the four indicator functions in the quartic filter. Rather, when the filter aligns only with the dominant beat, this creates a large response in the intensity map that is subsequently selected as the peak of the density. In addition, if the likelihood is only computed at local maxima in the smoothed signal energy representation, the softer beat can be skipped entirely. The presence of a soft S2 beat thus presents an ambiguity in which the cardiac period of length c is easily confused for one of length $\sim 2c$.

Chapter 7

Conclusion

The learning achievements attained during the completion of this work included several items not mentioned earlier in the document. These included the use of discriminative models to model heart sound identification, exponential family models and kernels to capture high-dimensional data. These earlier research paths did have an influence on the preliminary stages of the solution presented here, though they were not documented by this thesis for reasons of brevity.

Future work in this area must rely more heavily on statistical analysis of data in order to quantify the behaviour of the distributions involved. It was found that the distributional analysis presented in the first part of the experiments section was indirectly the most valuable part of designing the predictor since it was able to provide decisive evidence for the appropriateness of the distribution(s) used. Had this analysis taken place earlier in the project schedule, it might have facilitated the design of a better classifier.

The work presented here should be viewed as very preliminary in the area of automatic auscultation. It was found that the most difficult element of this task was that of providing appropriate probability distributions for a clinical sample that exhibits high variance and produces a large amount of observable data. A correct distributional model and feature representation will help provide tractable inference for heart sound analysis and will eventually lead to sound machine-assisted diagnoses.

Appendix I: Constants and operational parameters

Input representation
PWVD features : $M = 32$
PWVD feature rate: $s_p = 500Hz$
PWVD frequency bands: $f_{\min} = 10, f_{\max} = 100$
Sampling rate: $s = 4kHz$
Sample recording length: $L = 20s$
Heartbeat width: $2w = 100ms$

Waiting time distribution
Span of systolic interval used to select peak of distribution: $250ms$
Span of cardiac period used to select peak of distribution: $350ms$
Waiting time distribution uniform density mixture component: $\alpha = 0.04$
Waiting time parameter range: ($s \equiv$ systolic interval, $c \equiv$ cardiac period)

$$\mathbb{P} \;=\; \{\; s, c \;:\; s \in [120ms, 502ms], c \in [240ms, 1774ms], 1.8s < c < 4s \;\}$$

Viterbi decoder
Maximum search depth for systolic interval: $750ms$
Maximum search depth for diastolic interval: $1500ms$

Bibliography

[1] R. Barbieri, E.C. Matten, A.R.A. Alabi, and E.N. Brown. A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability. *American Journal of Physiology-Heart and Circulatory Physiology*, 288(1):H424–H435, 2005.

[2] G. Bojanov. Blood pressure, heart tones, and diagnoses. *Handbook of Cardiac Anatomy, Physiology, and Devices*, pages 243–255, 2009.

[3] D.R. Brillinger. The identification of point process systems. *The Annals of Probability*, pages 909–924, 1975.

[4] E.N. Brown, R. Barbieri, V. Ventura, R.E. Kass, and L.M. Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, 14(2):325–346, 2002.

[5] DS Carter and PM Prenter. Exponential spaces and counting processes. *Probability Theory and Related Fields*, 21(1):1–19, 1972.

[6] F. Chatelain, X. Descombes, and J. Zerubia. Parameter estimation for marked point processes. application to object extraction from remote sensing images. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 221–234. Springer, 2009.

[7] Z. Chen, E.N. Brown, and R. Barbieri. Characterizing nonlinear heartbeat dynamics within a point process framework. *Biomedical Engineering, IEEE Transactions on*, 57(6):1335–1347, 2010.

[8] Z. Chen, P.L. Purdon, E.N. Brown, and R. Barbieri. A differential autoregressive modeling approach within a point process framework for non-stationary heartbeat intervals analysis. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 3567–3570. IEEE, 2010.

[9] ES Chornoboy, LP Schramm, and AF Karr. Maximum likelihood identification of neural point process systems. *Biological cybernetics*, 59(4):265–275, 1988.

[10] L. Citi, EB Klerman, E.N. Brown, and R. Barbieri. Point process heart rate variability assessment during sleep deprivation. In *Computers in Cardiology, 2010*, pages 721–724. IEEE, 2010.

[11] L. Cohen. Generalized phase-space distribution functions. *Journal of Mathematical Physics*, 7:781, 1966.

[12] L. Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941 –981, jul 1989.

[13] D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: General theory and structure*, volume 2. Springer Verlag, 2007.

[14] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. Discrete-time processing of speech signals. Institute of Electrical and Electronics Engineers, 2000.

[15] A. Djebbari and F. Bereksi Reguig. Short-time fourier transform analysis of the phonocardiogram signal. In *Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conference on*, volume 2, pages 844–847. IEEE, 2000.

[16] A. Doucet and C. Andrieu. Iterative algorithms for state estimation of jump markov linear systems. *Signal Processing, IEEE Transactions on*, 49(6):1216–1227, 2001.

[17] A. Doucet and P. Duvaut. Bayesian estimation of state-space models applied to deconvolution of bernoulli–gaussian processes. *Signal Processing*, 57(2):147–161, 1997.

[18] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, 1946.

[19] LG Gamero and R. Watrous. Detection of the first and second heart sound using probabilistic models. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 3, pages 2877–2880. IEEE, 2003.

[20] Robert Grant and Victor A. McKusick. Symposium on cardiovascular sound. *Circulation*, 16:414–436, 1957.

[21] A. Haghighi-Mood and JN Torry. A sub-band energy tracking algorithm for heart sound segmentation. In *Computers in Cardiology 1995*, pages 501–504. IEEE, 1995.

[22] J.E. Hebden and JN Torry. Neural network and conventional classifiers to distinguish between first and second heart sounds. In *Artificial Intelligence Methods for Biomedical Data Processing, IEE Colloquium on*, pages 3–1. IET, 1996.

[23] P.A. Iaizzo. *Handbook of cardiac anatomy, physiology, and devices*. Humana Pr Inc, 2009.

[24] M. Jacobsen. *Point process theory and applications: marked point and piecewise deterministic processes*. Birkhauser, 2006.

[25] FC Jandre and MN Souza. Wavelet analysis of phonocardiograms: differences between normal and abnormal heart sounds. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 4, pages 1642–1644. IEEE, 1997.

[26] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[27] J. Kormylo and J. Mendel. Maximum likelihood detection and estimation of bernoulli-gaussian processes. *Information Theory, IEEE Transactions on*, 28(3):482–488, 1982.

[28] S. Kotz, T.J. Kozubowski, and K. Podgorski. An asymmetric multivariate laplace distribution. Technical report, Department of Statistics and Applied Probability, University of California, 2003.

[29] V. Kudriavtsev, V. Polyshchuk, D.L. Roy, et al. Heart energy signature spectrogram for cardiovascular diagnosis. *Biomedical engineering online*, 6(1):16, 2007.

[30] Brandt A. Last, G. *Marked Point Processes on the Real Line: The Dynamic Approach*. Springer, 1995.

[31] Aubrey Leathem. Splitting of the first and second heart sounds. *The Lancet*, pages 607–613, 1954.

[32] TS Leung, PR White, J. Cook, WB Collis, E. Brown, and AP Salmon. Analysis of the second heart sound for diagnosis of paediatric heart disease. In *Science, Measurement and Technology, IEE Proceedings-*, volume 145, pages 285–290. IET, 1998.

[33] H. Liang, S. Lukkarinen, and I. Hartimo. Heart sound segmentation algorithm based on heart sound envelogram. In *Computers in Cardiology 1997*, pages 105–108. IEEE, 1997.

[34] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[35] K.P. Murphy. Hidden semi-markov models (hsmms). *Unpublished notes.*, 2002.

[36] MS Obaidat. Phonocardiogram signal analysis: techniques and performance comparison. *Journal of medical engineering & technology*, 17(6):221–227, 1993.

[37] MS Obaidat and MM Matalgah. Performance of the short-time fourier transform and wavelet transform to phonocardiogram signal analysis. In *Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing: technological challenges of the 1990's*, pages 856–862. ACM, 1992.

[38] S. Rajan, R. Doraiswami, R. Stevenson, and R. Watrous. Wavelet based bank of correlators approach for phonocardiogram signal classification. In *Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium on*, pages 77–80. IEEE, 1998.

[39] P. Rakovic, E. Sejdic, LJ Stankovic, and J. Jiang. Time-frequency signal processing approaches with applications to heart sound analysis. In *Computers in Cardiology, 2006*, pages 197–200. IEEE, 2006.

[40] BA Reyes, S. Charleston-Villalobos, R. González-Camarena, and T. Aljama-Corrales. Time-frequency representations for second heart sound analysis. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 3616–3619. IEEE, 2008.

[41] E. Shwedyk, R. Balasubramanian, and RN Scott. A nonstationary model for the electromyogram. *Biomedical Engineering, IEEE Transactions on*, (5):417–424, 1977.

[42] Donald Lee Snyder. *Random Point Processes*. John Wiley & Sons, 1975.

[43] B. Tovar-Corona and JN Torry. Time-frequency representation of systolic murmurs using wavelets. In *Computers in Cardiology 1998*, pages 601–604. IEEE, 1998.

[44] J. Tugnait. Adaptive estimation and identification for discrete systems with markov jump parameters. *Automatic Control, IEEE Transactions on*, 27(5):1054–1065, 1982.

[45] Marie-Collette van Lieshout. *Markov Point Processes*. Imperial College Press, 2000.

[46] H. Visk. On the parameter estimation of the asymmetric multivariate laplace distribution. *Communications in StatisticsTheory and Methods*, 38(4):461–470, 2009.

[47] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

[48] W. Wang, Z. Guo, J. Yang, Y. Zhang, L.G. Durand, and M. Loew. Analysis of the first heart sound using the matching pursuit method. *Medical and Biological Engineering and Computing*, 39(6):644–648, 2001.

[49] PR White, WB Collis, and AP Salmon. Analysing heart murmurs using time-frequency methods. In *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*, pages 385–388. IEEE, 1996.

[50] J. Xu, L. Durand, and P. Pibarot. Nonlinear transient chirp signal modeling of the aortic and pulmonary components of the second heart sound. *Biomedical Engineering, IEEE Transactions on*, 47(10):1328–1335, 2000.

[51] J. Xu, L.G. Durand, and P. Pibarot. Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. *Biomedical Engineering, IEEE Transactions on*, 48(3):277–283, 2001.

[52] W. Yanjun, X. Jingping, Z. Yan, W. Jing, W. Bo, and C. Jingzhi. Time-frequency analysis of the second heart sound signals. In *Engineering in Medicine and Biology Society, 1995., IEEE 17th Annual Conference*, volume 1, pages 131–132. IEEE, 1995.

[53] K. Yu and J. Zhang. A three-parameter asymmetric laplace distribution and its extension. *Communications in StatisticsTheory and Methods*, 34(9-10):1867–1879, 2005.

[54] S.Z. Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010.