

Probabilistic Modeling of Kidney Dynamics for Renal Failure Prediction

by

Boon Teik Ooi

S.B., Massachusetts Institute of Technology (2012)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
August 7, 2013

Certified by
Prof. Peter Szolovits
Professor of Computer Science and Engineering
Thesis Supervisor

Certified by
Dr. William J. Long
Principal Research Scientist
Thesis Supervisor

Accepted by
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Probabilistic Modeling of Kidney Dynamics for Renal Failure Prediction

by

Boon Teik Ooi

Submitted to the Department of Electrical Engineering and Computer Science
on August 7, 2013, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The large quantity of clinical data collected from the Intensive Care Unit (ICU) has made clinical investigation by a data-driven approach more effective. In this thesis, we developed probabilistic models for modeling variable kinetics and temporal dynamics of states. We applied the models to the prediction of acute kidney injury (AKI), but the models are applicable to other medical conditions as well.

It is known that serum creatinine follows first-order clearance kinetics. We developed a stochastic kinetic model for first-order clearance and used it to model creatinine kinetics. Some properties implied by the model that are verifiable with the available data are consistent with the empirical results. Those properties are mean-reversion, variation with linear standard deviation, and convergence of variance to a finite value.

Based on the stochastic kinetic model, creatinine can be treated as a lognormal random variable with state-dependent parameters. We model the temporal dynamics of kidney states and creatinine using a Hidden Markov Model. Observations of creatinine are assumed to be random variables, with baseline creatinine as mean. Each individual baseline is itself a random variable sampled from a population distribution. Baseline for each patient can be estimated by combining the population distribution and all creatinine observations of the patient using techniques similar to Bayesian inference. Prediction of acute kidney injury with this generative model gives an AUC of 0.8259 and 0.8497 for female and male population respectively.

Thesis Supervisor: Prof. Peter Szolovits
Title: Professor of Computer Science and Engineering

Thesis Supervisor: Dr. William J. Long
Title: Principal Research Scientist

Acknowledgments

The time when I was looking for a thesis project, at the last minute, in Prof. Szolovits's office, is still fresh in my mind. I would like to thank him for offering me a project in the group, even though I may not have had the necessary background. Pete has been a good mentor to me and discussions with him gave useful insights into the project. Several times when we encountered issues that he is not familiar with, he took the trouble to schedule meetings with people who might be able to help with those questions. I am also grateful for the RA-ship support provided by Pete, as that allowed me to spend more time in research.

I would like to thank Dr. Bill Long, another advisor of mine, for his guidance in this project. He has been very patient in explaining to me about medical concepts and the MIMIC II database whenever I have questions. Completion of this thesis would not have been possible without his help.

My Research Assistantship was funded by Siemens Corporation through a project titled "Predictive Analytics and Dashboard for Population Management," grant U54 LM008748 from the National Library of Medicine, titled "Informatics for Integrating Biology & the Bedside," and grant R01 EB001659 from the National Institute of Biomedical Imaging and Bioengineering. I am grateful for their support.

The past 5 years at MIT have been an invaluable experience to me. In this place, I have had the opportunity to explore my interests, take interesting classes, acquire useful skills, and improve my problem solving ability. While I did not improve on my social skills, I made many good friends here.

Finally, I would like to thank my family for their love and support. None of this is possible without them.

Contents

1	Introduction	15
1.1	Overview	15
1.2	Thesis Organization	17
2	Related Work	19
2.1	Diagnostic Criteria	19
2.2	Baseline Estimation	22
2.3	Creatinine Kinetics	24
2.4	Empirical Bayes Method	26
3	Dataset Preparation	29
3.1	MIMIC II Database	30
3.2	Relevant Variables	30
3.2.1	Demographic Variables	31
3.2.2	Chart Variables	31
3.2.3	Lab Variables	31
3.2.4	IO Variables	32
3.2.5	Ground Truth	32
3.3	Selection Criteria	33
3.4	Issues with Dataset	33
3.4.1	Discretization	34
3.4.2	Control Groups	36

3.4.3	Class Imbalance	37
3.4.4	Unequal Intervals	37
4	Modeling Variable Kinetics	41
4.1	First-order Clearance Kinetics	42
4.2	Stochastic Kinetic Model	42
4.2.1	Initial Value	43
4.2.2	Equilibrium Value	44
4.2.3	Variable Properties	46
4.3	State Abstraction	47
4.4	State Transition	47
5	Temporal Dynamics of States	49
5.1	Generative Model for State Dynamics	50
5.1.1	Compound Sampling Observation	50
5.1.2	Hidden Markov Model	51
5.1.3	Variational Inference	52
5.2	Learning Algorithm	54
5.2.1	EM Algorithm	54
5.2.2	Estimation of Population Parameters	55
5.3	Prediction	56
5.3.1	Baseline Estimation	56
5.3.2	Inference of States	57
5.4	Discussion	58
6	Results	61
6.1	Urine Output	62
6.2	Creatinine Kinetics	63
6.2.1	Heuristic Interpretation	64
6.2.2	Mean-reverting Drift	65

6.2.3	Linear Diffusion	66
6.2.4	Convergence of Variance	66
6.2.5	Generation and Clearance	67
6.3	Kidney State Dynamics	69
6.3.1	Lognormal Model for Creatinine	70
6.3.2	Goodness of Fit	71
6.3.3	Stages of Acute Kidney Injury	72
6.3.4	Classification Results	74
7	Conclusion	77
7.1	Summary	77
7.2	Future Work	78
A	Linear Stochastic Differential Equation	81
A.1	The General Solution	81
A.2	Solution to SDE with Constant Coefficients	83

List of Figures

2-1	The three levels of renal dysfunction (Risk, Injury, Failure) and two clinical outcomes (Loss, ESRD) of the RIFLE classification system. The shape of the figure that becomes increasingly narrower indicates the increasing specificity across the stages of RIFLE.	21
2-2	The three stages of the AKIN staging system. The increase in creatinine must occur within 48 hours.	22
3-1	Histogram of log-creatinine with evenly spaced bins	35
3-2	Histogram of log-creatinine with bins that respect the discretization	36
3-3	Histogram of creatinine measurements intervals for normal patients	38
3-4	Histogram of creatinine measurements intervals for patients with renal failure	39
6-1	Mean of the change in creatinine for different values of creatinine	65
6-2	Standard deviation of the change in creatinine for different values of creatinine	66
6-3	Standard deviation of the change in creatinine for different durations after measurement	68
6-4	Histogram of creatinine values of normal patients and the probability density functions of the lognormal distribution. Left: female, Right: male	71
6-5	Probability plot of creatinine values of normal patients versus the lognormal distribution. Left: female, Right: male.	72

List of Tables

2.1	Definitions of acute kidney injury that have been used in several published studies.	20
3.1	Common attributes of clinical events with the descriptions.	31
3.2	ICD9 codes for Acute Kidney Injury.	32
6.1	The number of patients, hospital admissions, and creatinine samples for each gender.	69
6.2	Results of Pearson's chi-squared test for goodness of fit	72
6.3	The creatinine criteria in RIFLE and AKIN classification system. For AKIN, the increase in creatinine must occur within 48 hours. Cr, serum creatinine. .	73
6.4	Population mean of creatinine for each state	73
6.5	Typical values of creatinine for each state	73
6.6	Sensitivity, specificity and the area under ROC of the model.	74

Chapter 1

Introduction

Rapid advancement of technology is transforming modern healthcare into a more complex and data-intensive environment. This is especially true in the Intensive Care Unit (ICU), which provides continuous and comprehensive care for critically ill patients. Because these patients are already suffering from severe conditions, clinical decision making in the ICU is very challenging. Any misjudgment can easily cost a life. Therefore, there has been increasing interest in automated patient monitoring and decision support systems that can assist physicians in decision making.

The large amount of data generated from constant monitoring of intensive care patients is an invaluable resource for that purpose. The rich collection of clinical data, which includes clinical measurements, lab tests, and medications, can be used for building machine learning models that can improve the efficiency and timeliness of clinical decision making. Our collaborators have taken the initiative to collect and disseminate such data for research purposes [43]. This project utilized such data for the study of acute kidney injury (AKI).

1.1 Overview

In the past, investigations of diseases have been mostly based on animal models, which are thought to be unavoidable without better alternatives. Many people, however, remain

skeptical about the applicability of such animal models to numerous human diseases.

The advent of clinical databases that contain clinical data collected from intensive care patients has made the data-driven approach an attractive option for the investigation of human diseases. Many old models can be verified and new questions can be answered using the data. The major limitation is that the data are collected during the delivery of care, without specific research purposes. Thus, not all modeling tasks are feasible with these data. Nevertheless, this is a big step towards more use of data-driven models in medicine.

An example of a disease that is usually studied by utilizing animal models is acute kidney injury. There are three basic types of animal models that are used in the studies of AKI, namely ischemia, toxin and sepsis models, and several subtypes [19, 48, 50]. However, no single model is universally applicable as each model has its pros and cons. In fact, none of the existing models gives a reproducible model of AKI on intensive care patients. Despite these limitations, the animal models were carefully considered during the design of the definitions of AKI [5]. It is agreed that better models are definitely needed, but animal models remain fundamental to improving our understanding of acute kidney injury.

Acute kidney injury is a complex disorder that is common in the intensive care setting. It was reported that AKI affects 1% to 31% of intensive care patients. Studies have shown that AKI is a key risk factor for mortality, with mortality rate ranging from 20% to 82%, depending on the population and the criteria used to define AKI [12]. Despite being an important clinical issue, a universally accepted definition of AKI did not exist until the invention of the RIFLE classification system in 2004. Before that, various definitions had been used in the literature, making it difficult to make comparisons across studies.

The goal of this project is to improve the prediction of renal failure by using a data-driven approach. We hope that our model can complement the RIFLE criteria by providing a probabilistic view to the diagnosis of renal failure. Serum creatinine is undeniably one of the most important indicators of renal health. In order to have a better understanding of the nature of creatinine, we model creatinine kinetics as a stochastic process. Based on that, a generative model for the temporal dynamics of kidney states is developed and is used for

AKI prediction. We then discuss some statistical properties of the model and its connection to the RIFLE criteria.

1.2 Thesis Organization

This thesis is organized into the following chapters.

Chapter 2 provides an overview of the existing work on diagnosis of acute kidney injury and statistical analysis techniques.

Chapter 3 introduces the clinical database, from which the dataset used in this project is obtained. Some problems with the dataset that might affect statistical learning are also described.

Chapter 4 describes the stochastic kinetic model for modeling the first-order clearance kinetics of clinical variables. Then, some theoretical properties implied by the model are discussed.

Chapter 5 develops the generative model for the temporal dynamics of states and clinical observations. The algorithms for parameters estimation and state prediction with the model are also described.

Chapter 6 examines the modeling of creatinine kinetics and the progression of acute kidney injury using the models developed in this thesis.

Chapter 7 concludes the thesis with a summary of contributions and potential directions for future research.

Chapter 2

Related Work

This chapter provides an overview of the work that have been done on the diagnosis of acute kidney injury and statistical learning that are relevant to this project. We start by providing some background on the advancement in the definitions of AKI over the past decade. Due to the importance of creatinine in diagnosis of AKI, there have been some studies on AKI through modeling of creatinine kinetics. We survey some of these works, as our stochastic kinetic model can be considered as an extension to their models. Estimation of baseline creatinine is fundamental to diagnosis, as the stages of AKI are defined by the increase in creatinine relative to the baseline. We describe some contemporary methods for estimating baseline creatinine. Finally, we introduce an idea in statistics that is relevant to creatinine observations and estimation of individual baselines.

2.1 Diagnostic Criteria

Before the invention of the RIFLE classification system in 2004, there was no consensus in the definition of acute kidney injury. More than 30 definitions of AKI have been used in the literature and we list some of these definitions in Table 2.1 [30].

All definitions use increases in serum creatinine as a criterion for acute kidney injury. However, the magnitude of the required increase varies between definitions. Another notable

Author	Definition
Taylor, et al.	0.3 mg/dl increase in serum creatinine
Soloman, et al.	0.5 mg/dl increase in serum creatinine within 48 hours
Hou, et al.	0.5 mg/dl increase in serum creatinine if baseline below 1.9 mg/dl, 1.0 mg/dl increase in serum creatinine if baseline between 2.0 - 4.9 mg/dl, 1.5 mg/dl increase in serum creatinine if baseline above 5.0 mg/dl
Levy, et al.	25% increase in serum creatinine to at least 2.0 mg/dl within 48 hours
Parfrey, et al.	50% increase in serum creatinine to at least 1.4 mg/dl
Cochran, et al.	more than 0.3 mg/dl and 20% increase in serum creatinine
Hirschberg, et al.	serum creatinine above 3.0 mg/dl with baseline below 1.8 mg/dl, or “acute decrease” in creatinine clearance to below 25 ml/min

Table 2.1: Definitions of acute kidney injury that have been used in several published studies.

distinction between the definitions is whether they are based on value increase or percentage increase. The criteria by Hou, et al. can be considered as percentage increases. Another distinction is the duration constraint on the creatinine increase, where some definitions imposed a 48-hour time constraint to the increase in the criteria.

In search for a standard definition of AKI, the Acute Dialysis Quality Initiative (ADQI) group conducted a conference to gather the experts in the field. Sufficient consensus was achieved for most of the proposed questions, resulting in the RIFLE classification system for acute kidney injury. The RIFLE criteria classify acute kidney injury into three levels of renal dysfunction and two clinical outcomes. The three levels of renal dysfunction are *Risk* of renal dysfunction, *Injury* to the kidney, and *Failure* of kidney function. Each level has separate criteria for creatinine and urine output (UO). The two clinical outcomes are *Loss* of kidney function and End-stage renal disease (*ESRD*).

While the RIFLE criteria are termed the definition of AKI, they are used more like a standardized guideline for diagnosis. It was mentioned in the report that patients classified into the Risk stage might not actually have renal failure because the Risk stage is designed to have high sensitivity. Specificity of the RIFLE classification system increases from the Risk stage to the ESRD stage. The RIFLE criteria for the three stages and two outcomes are given in Figure 2-1. For the creatinine criteria, the increases are relative to the baseline.

Of course, the RIFLE criteria are not perfect. To overcome the limitations of RIFLE,

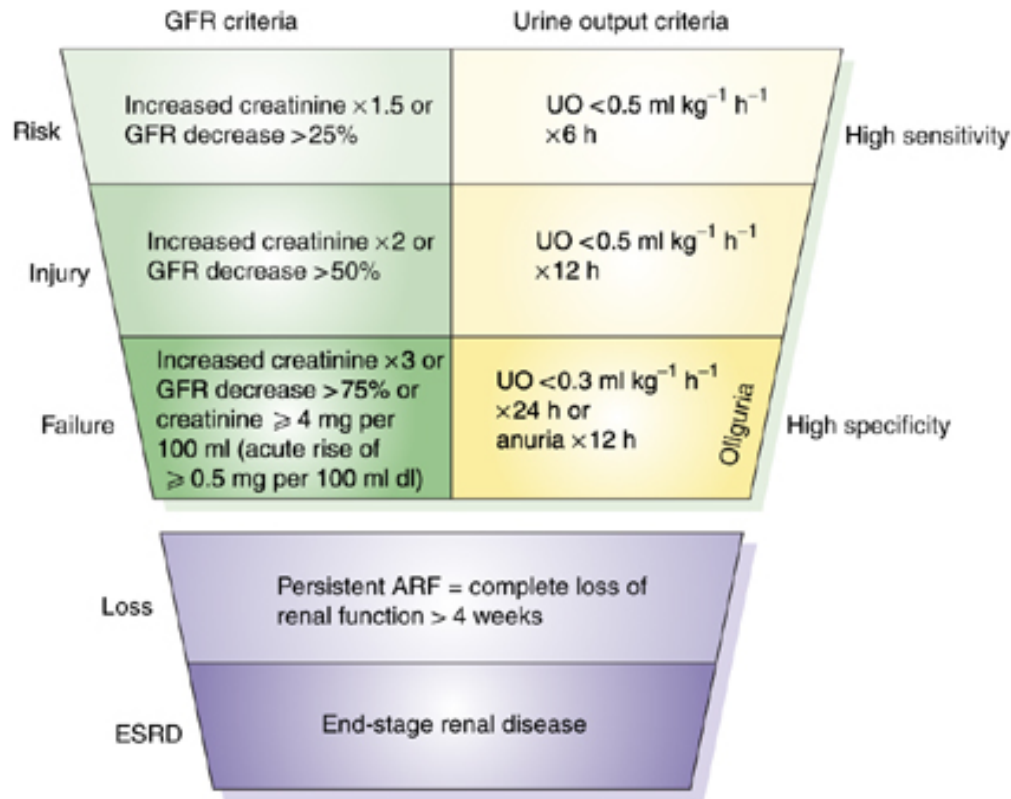


Figure 2-1: The three levels of renal dysfunction (Risk, Injury, Failure) and two clinical outcomes (Loss, ESRD) of the RIFLE classification system. The shape of the figure that becomes increasingly narrower indicates the increasing specificity across the stages of RIFLE.

the Acute Kidney Injury Network (AKIN) introduced the AKIN staging system as a refinement to the RIFLE criteria [31]. The two outcomes were removed from the staging system and remain outcomes. There has been increasing evidence that small increments in serum creatinine are associated with adverse outcomes that manifest in increased risk of mortality [8, 39, 26]. To account for that, AKIN modified the criteria of the Risk stage to include an absolute increase of 0.3 mg/dl from baseline creatinine, in addition to the 1.5 \times percentage increase. Lastly, the criteria are based on changes that occur within 48 hours. New criteria for the three levels of severity of AKI are given in Figure 2-2.

Despite their limitations, introduction of the RIFLE and AKIN criteria has been a big step towards a more standardized definition, allowing meaningful comparisons across studies. Several studies have been published aiming to validate the two criteria from their applications

Medscape® www.medscape.com		
Stage	Serum creatinine criteria	Urine output criteria
1	Increase of $\geq 26.4 \mu\text{mol/l}$ (0.3 mg/dl) OR to 150–200% of baseline (1.5–2.0-fold)	$< 0.5 \text{ ml/kg/h}$ for $> 6 \text{ h}$
2	Increase to $> 200\text{--}300\%$ of baseline ($> 2\text{--}3\text{-fold}$)	$< 0.5 \text{ ml/kg/h}$ for $> 12 \text{ h}$
3 ^a	Increase to $> 300\%$ of baseline ($> 3\text{-fold}$; or serum creatinine $\geq 354 \mu\text{mol/l}$ [4.0 mg/dl] with an acute rise of at least $44 \mu\text{mol/l}$ [0.5 mg/dl])	$< 0.3 \text{ ml/kg/h}$ for 24 h OR anuria for 12 h
Only one criterion (creatinine or urine output) needs to be fulfilled to qualify for a stage. ^a Patients who receive renal replacement therapy are considered to have met the criteria for Stage 3, irrespective of the stage that they are in at the time of commencement of renal replacement therapy. Permission obtained from BioMed Central © Mehta RL et al. (2007) <i>Crit Care</i> 11: R31.		
Source: Nat Clin Pract Nephrol © 2007 Nature Publishing Group		

Figure 2-2: The three stages of the AKIN staging system. The increase in creatinine must occur within 48 hours.

in clinical practice [41]. Most of the studies are based on retrospective analysis, though there are some with prospective analysis. We are not aware of any attempt to validate the criteria with statistical models, and our work is aimed toward that direction.

2.2 Baseline Estimation

Conceptually, AKI signifies a rapid worsening of kidney function from pre-morbid levels. Creatinine criteria of the RIFLE and AKIN classification system are based on increase in creatinine from the baseline value, which reflects the patients pre-morbid kidney function. All changes should be compared to the individual baseline since the renal capability of each individual is different. Therefore, accurate estimation of baseline creatinine has become a fundamental component of the diagnosis of AKI.

However, we again face the problem that there is no standard definition of baseline creatinine. Worse still, creatinine values of patients before hospitalization are usually not available. That makes accurate evaluation of baseline from such pre-hospitalization data totally hopeless for many patients and has spurred the development of various strategies for estimating baseline creatinine without using pre-hospitalization creatinine values.

The ADQI group, who developed the RIFLE criteria, recommend back-calculation of creatinine from the Modification of Diet in Renal Disease (MDRD) formula, by assuming an

estimated GFR of 75 ml/min/1.73m² for every individual [5]. The original purpose of the MDRD formula was to estimate GFR from creatinine value, thus the term back-calculation. Calculation with the formula only relies on demographic information such as gender, ethnicity and age.

Other definitions of baseline creatinine that have been used in the literature are the creatinine at the time of hospital admission, the minimum creatinine value during the hospital stay, estimation using some other formulas, or the lowest value among these [20, 47].

The viability of determining baseline creatinine by back-calculation with the MDRD formula is assessed in a study by Pickering, et al. [38]. They conducted a retrospective study on patients with known baseline creatinine. The patients were classified according to the RIFLE criteria using the following baseline estimates:

- C_{75} , back-calculation with MDRD assuming a GFR of 75 ml/min/1.73m²
- C_{100} , back-calculation with MDRD assuming a GFR of 100 ml/min/1.73m²
- C_{ln} , average of 1000 random values from a lognormal distribution fitted to the baselines of all patients.
- C_{low} , the lowest creatinine value in the first week in the ICU

According to their results, C_{75} and C_{100} greatly overestimated AKI; C_{low} overestimated AKI according to AKIN but correctly classified AKI according to RIFLE; C_{ln} correctly classified AKI under both criteria. Among the baseline estimates considered, C_{ln} has the best overall performance.

The authors explained that C_{ln} performed better than C_{75} and C_{100} because back-calculation relies only on age and race, but not the actual renal function. On the other hand, distribution of C_{ln} is based on the aggregate renal function of the population and therefore, can better estimate the baseline of individuals that belong to the population.

The main difference between C_{ln} and C_{low} is that C_{ln} estimates the average baseline of the population, whereas C_{low} estimates the individual baseline. However, if the renal functions of the patients are independent, we would expect C_{ln} to outperform C_{low} . This result can

be attributed to what is known as Stein’s paradox in estimation theory. Stein’s paradox will be discussed in Section 2.4, as it is related to the empirical Bayes method.

2.3 Creatinine Kinetics

Because serum creatinine is an important indicator of renal health, there have been several studies on creatinine kinetics in the context of acute kidney injury.

Moran, et al. investigated the course of acute kidney injury through a creatinine kinetic model in 1985, before the invention of the RIFLE criteria [32]. They developed a model of creatinine kinetics that assumes a constant generation rate and first-order clearance rate, and used that model to predict the relationship between creatinine clearance and creatinine concentration in patients with postischemic acute renal failure. Several patterns of changes in the creatinine clearance of the patients were identified, including abrupt step decrement, ramp decrement, and exponential decrement. Their models were shown to have good predictive performance on the course and prognosis of acute renal failure in individual patients.

In a study by Waikar, et al., the authors investigated the connection between creatinine kinetics and severity of AKI [46]. They considered both a single-compartment model and a two-compartment model for modeling creatinine kinetics. The single-compartment model assumes a single compartment where creatinine is uniformly distributed and is the same model as that used in Moral, et al. [32]. The two-compartment model assumes that creatinine is generated in the intracellular compartment and then diffuses by first-order kinetics into the extracellular compartment, where clearance by the kidney occurs. Although the two-compartment model better represents the metabolism of creatinine, the trajectories of changes in creatinine predicted by their two models are almost identical. Therefore, the modeling of creatinine kinetics in later chapters assumes a single-compartment model.

The authors simulated creatinine kinetics after AKI in the setting of normal kidney function, and stage 2, 3 and 4 chronic kidney disease (CKD). They showed that 24 hours after a 90% reduction in creatinine clearance, the percentage changes in creatinine are highly dependent on baseline kidney function, whereas the absolute increases are approximately the

same across all levels of baseline kidney function. From another perspective, the time to reach a 50% increase in creatinine ranges from 4 hours for normal baseline to 27 hours for stage 4 CKD, while the time to reach a 0.5 mg/dl increase was virtually identical if the reduction in creatinine clearance is more than 50%. Based on that, they proposed an alternative definition of AKI that incorporates absolute changes in serum creatinine over a 24-48 hour time period.

It occurs to me that the authors have made an implicit assumption, for which no justification is provided. They claimed that AKI definitions that use percentage increase in creatinine are flawed because for any given percentage reduction in creatinine clearance, percentage increase in creatinine is slower for patients with high baseline. Consequently, identical percentage reduction in creatinine clearance results in different classifications of AKI severity depending on baseline kidney function. The implicit assumption is that the same percentage reduction in creatinine clearance represents the same degree of severity increase for all levels of baseline.

I actually hold the opposite opinion, that identical absolute changes in creatinine clearance result in the same severity level. My reasoning for that is based on the assumption that random fluctuations in creatinine clearance are due to additive noise. Consider a creatinine clearance rate k with variation δ . The likelihood of fluctuating to $k + \delta$ and $k - \delta$ are the same and by symmetry, changes in severity should have equal magnitude. If their assumption is true, then the noise in creatinine clearance should be multiplicative. If the noise is indeed additive, changes in creatinine after a short interval should have variance that is linear in the creatinine value. This is consistent with an empirical property shown in Figure 6-2. Nevertheless, the evidence is not strong due to various deficiencies of the data.

In a nutshell, we agree with the RIFLE criteria that the conceptual model of AKI severity should be based on percentage changes in serum creatinine instead of absolute changes. A creatinine kinetic model with noise will be discussed in Chapter 4.

2.4 Empirical Bayes Method

Charles Stein shocked the statistical world in 1956 with his proof that maximum-likelihood estimation for Gaussian models, used for more than a century, was inadmissible beyond the simple two-dimensional situation. The simplest form of Stein's paradox involves estimation of parameters $\theta_1, \theta_2, \dots, \theta_k$ with $k \geq 3$, and for each parameter, we have one observation

$$x_i \sim \mathcal{N}(\theta_i, 1), \quad i = 1, \dots, k.$$

The maximum-likelihood estimator of $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ is just the vector of observations $\mathbf{x} = (x_1, x_2, \dots, x_k)$. Under squared error loss, frequentist risk of estimator \mathbf{x} is 1 for every value of $\boldsymbol{\theta}$. James and Stein showed that the estimator $\mathbf{y} = (y_1, y_2, \dots, y_k)$ where

$$y_i = \left(1 - \frac{k-2}{S}\right) x_i, \quad S = \|\mathbf{x}\|^2$$

has frequentist risk strictly less than 1 for all values of $\boldsymbol{\theta}$ [16]. The estimator \mathbf{y} is now known as the James-Stein estimator.

Efron and Morris provided a Bayesian derivation of the James-Stein estimator by assuming a prior distribution $\theta_i \sim \mathcal{N}(0, \tau^2)$. The Bayes estimator that minimizes the Bayes risk is the posterior mean

$$\hat{\theta}_i = \left(1 - \frac{1}{\tau^2 + 1}\right) x_i.$$

From the marginal distribution $x_i \sim \mathcal{N}(0, 1 + \tau^2)$, S has chi-square distribution with k degrees of freedom where

$$\mathbb{E} \left[\frac{k-2}{S} \right] = \frac{1}{1 + \tau^2}.$$

Substituting the unknown $1/(1 + \tau^2)$ in the Bayes estimator with an unbiased estimator $(k-2)/S$ gives the James-Stein estimator.

Although we have chosen a prior with mean $\mu_i = 0$, the James-Stein estimator dominates the maximum-likelihood estimator for every choice of μ_i . The James-Stein estimator can be thought as shrinking each observed value x_i towards 0.

In fact, the James-Stein estimator is a special case of the empirical Bayes method. Consider a more general setting with

$$x_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \quad \theta_i \sim \mathcal{N}(\mu, \tau^2), \quad i = 1, \dots, k$$

where σ is known. The empirical Bayes approach finds the parameter estimates that maximize the marginal likelihood

$$x_i \sim \mathcal{N}(\mu, \sigma^2 + \tau^2),$$

which gives $\hat{\mu} = \bar{x}$, $\hat{\tau} = S - \sigma$. The James-Stein estimator for this general case is given by

$$\hat{\theta}_i = \bar{x} + \left(1 - \frac{(k-3)\sigma^2}{S}\right) (x_i - \bar{x})$$

which also has lower frequentist risk than the maximum likelihood estimator. However, we now require $k \geq 3$, as we estimated one additional parameter μ from the data.

James-Stein estimator can be applied to the estimation of baseline creatinine, where x_i is the log of a creatinine observation and θ_i is the log baseline. While individual baselines are independent, pooling information from the population can improve the estimates.

We use the empirical Bayes idea to model the population and individual baseline distributions. However, estimation of individual baselines is not as straightforward in our case because we have multiple states, and the state that each observation corresponds to is unknown. In order to estimate the parameters of each state, we need to model the dependency structure of observations and the corresponding hidden states. Our method for parameter estimation with hidden states will be discussed in Section 5.1.

Chapter 3

Dataset Preparation

The dataset used for machine learning and modeling in this thesis is obtained from the MIMIC II database. The MIMIC II database contains detailed clinical records, including lab results, bedside monitoring waveforms and electronic documentation taken during the delivery of care in the Intensive Care Unit (ICU). In other words, the data are not collected with any specific research purpose in mind. While the rich collection of clinical data is invaluable for data analytics research, there are several implications of that that makes machine learning challenging. Special attention is required for handling problems like selection bias, missing data, non-uniform sampling, etc. Like any data sources that required manual input by human, the MIMIC II database is subjected to recording error. Data cleaning can be helpful in mitigating the susceptibility to this kind of error.

This chapter gives an overview of the MIMIC II database and the various types of data that are available in the database. The clinical variables to consider in the dataset preparation and their relevance to renal failure are discussed. Finally, we describe some issues with the dataset that may cause problems in statistical modeling. These issues need to be taken into account in the design of learning algorithms, as different algorithms are affected differently by those issues. For instance, class imbalance can be detrimental to standard classification algorithm while its effect on unsupervised learning is small.

3.1 MIMIC II Database

The Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC II) database was created to facilitate the development and evaluation of ICU decision-support systems [10, 43]. After years of data collection, the MIMIC II database now contains physiological information on over 30,000 patients who have been admitted to the ICU of a Boston teaching hospital. Indeed, one of the goals of the MIMIC project is to encourage research in the development of patient monitoring technology and there have been successful attempts in constructing models for predicting hazardous episodes in the ICU [22].

The MIMIC II database can be divided into two components, namely the Waveform Database and the Clinical Database. The Waveform Database contains data recorded from bedside monitoring such as Electrocardiograph Monitors (ECG), Arterial Blood Pressure Monitors (ABP) and Respiratory Monitors. The Clinical Database contains complete records of clinical events that occurred during the ICU stay such as administration of medications, lab tests, and clinical measurements, along with the timestamps. Therefore, clinical events for each variable can be viewed as unevenly spaced time series. Besides time series data, the Clinical Database also contains demographic information, ICD9 codes, clinical notes, etc.

This project only uses the Clinical Database, as the waveform data are less useful for our modeling objective. Instead, time series data are used extensively for unsupervised learning. Evaluation of the learning algorithm is based on ground truth extracted from the ICD9 codes and clinical notes.

3.2 Relevant Variables

The Clinical Database is organized as a relational database, where each table corresponds to a different record type. For time series data, each clinical event is stored as a row in the appropriate table. While the attributes vary between tables, some attributes that are common to most tables are listed in Table 3.1.

Attribute	Description
Subject_ID	Unique identifier for each patient
Hadm_ID	Unique identifier for each hospital admission
ICUStay_ID	Unique identifier for each ICU admission
Charttime	Time at which the event occurred
ItemID	Identifier to specify the exact event
Value	Numeric value associated with the event

Table 3.1: Common attributes of clinical events with the descriptions.

3.2.1 Demographic Variables

Demographic information of patients are stored in the **D_Patients** table. It contains information such as gender, date of birth, date of death (if applicable) and the **Hadm_ID** of the patients' admissions. Gender and age, which can be calculated from the date of birth, are the two main demographic variables that are relevant to the prediction of AKI. The generation rate of creatinine depends on the muscle mass, which is highly correlated with age and gender. The creatinine clearance rate of the kidney is also affected by age and gender.

3.2.2 Chart Variables

The **ChartEvents** table contains clinical measurements taken from patients while receiving care in the ICU. Chart variables include heart rate, respiration rate, blood pressure, blood gases, etc. The intervals between measurements vary depending on the measured quantities, and range from minutes to days. Patients in a more severe condition might have higher measurement frequency. We considered including blood pressure as a feature because it is an indicator of decreased blood flow to the kidney. However, the correlation between blood pressure and AKI is very weak and it introduces a lot of noise. Therefore, it was eliminated from the feature set.

3.2.3 Lab Variables

Values from lab tests are recorded in the **LabEvents** table, which includes variables like serum creatinine and BUN (blood urea nitrogen). Unlike chart variables, data for lab variables

outside the ICU are also available in the database. Hence, we have many more samples for lab variables than for chart variables. This project started by using both BUN and creatinine for AKI prediction. It is known that BUN is not specific enough for the diagnosis of AKI compared to creatinine, because it is also reacts to dehydration and heart failure. BUN was initially included, as we thought it might give additional information about pre-renal disease. However, BUN does not seem to carry more signal beyond that of creatinine and it actually introduces noise in the prediction. Therefore, BUN was eventually excluded from the final feature set.

3.2.4 IO Variables

The `TotalBalEvents` table recorded cumulative urine output of each patient for every 24-hour interval in the ICU. More detailed breakdown of urine output data is also available in the `IOEvents` table. Of course, the values in `TotalBalEvents` are more stable and robust, due to the smoothing over a 24-hour period. Urine output is one the criteria in the RIFLE classification system. However, we decided not to use urine output for prediction due to some problems with this variable in our data set. A more detailed discussion on this is in Chapter 6.

3.2.5 Ground Truth

The `ICD9` table contains the ICD9 codes for disease information of patients for each hospital admission. A hospital admission is associated with acute kidney injury if the ICD9 codes for that admission include any of the following.

445.81	580-580.99
583-584.99	586-586.99
590.1	593.89-593.99
646.21-646.22	669.32
866-866.99	

Table 3.2: ICD9 codes for Acute Kidney Injury.

3.3 Selection Criteria

As mentioned, the database is not free from errors. There are many incidents of missing data that may hinder analysis on the data. In our case, missing ICD9 codes is a huge problem because the ground truth based on clinical notes alone is not reliable.

Furthermore, not all patients may have the necessary data for our learning objective. The available data for each patient are primarily based on the patient's physiological condition and the physicians' opinion. For example, creatinine values for a patient would not be available if the physician does not think that the creatinine value would offer any additional insight into the patient's condition.

In the preparation of the dataset, we need to exclude patients that are not eligible according to some criteria. In particular, we include an admission in the final dataset if all of the following criteria are satisfied.

- ICD9 codes are not missing
- patient was between 20 and 75 years old
- has at least one creatinine observation
- patient not receiving dialysis treatment
- patient have not not had kidney transplant

We limit the dataset to adult patients with ages between 20 and 75 years old to avoid the extreme variation in kidney function due to the effects of age. In addition, that assumption makes the age distribution within our dataset more uniform, so that the average age effect on baseline creatinine is negligible compared to gender, AKI severity and noise.

3.4 Issues with Dataset

This section describes some problems with the dataset of that may hinder the performance of various learning algorithms. We need to keep that in mind when learning from the data, as

different algorithms have different degrees of vulnerability to each problem. It is important to make sure that none of the modeling assumptions are violated.

3.4.1 Discretization

All clinical measurements are limited by the precision of the measurement apparatus. As a result, measurements of variables that are continuous in nature can be regarded as discretizations of the real values. In the database, values of BUN are integers whereas values of creatinine are multiples of 0.1, meaning that values of two variables are rounded to the nearest 1.0 and 0.1 respectively.

While the effect of discretization may be negligible most of the time, it could be problematic when taking non-linear transformation of the values such as logarithms and exponentials. The set of possible values that becomes unevenly spaced after transformation is problematic for plotting histograms. For instance, the possible values for creatinine are $\{0.0, 0.1, 0.2, \dots\}$ and taking the logarithm results in $\{-\infty, -2.303, -1.609, \dots\}$. Plotting the histograms of log-creatinine require unevenly-spaced bins that respect the discretization in order to get a decent visualization of the distribution. Figure 3-1 and Figure 3-2 illustrate the difference between simple evenly spaced bins and bins that respect the discretization.

For that reason, goodness of fit tests that require value transformation may be unsuitable for testing distributions of clinical variables. The test result would not be accurate as the original distribution has been distorted by the discretization, limiting the choice of statistical tests that can be performed on the data. For testing of distributions, a statistical test that can be used in our case is the Pearson's chi-squared test. The test establishes whether or not an observed frequency distribution differs from a theoretical distribution. For example, suppose we would like to test the hypothesis that creatinine has lognormal distribution. The observed frequency distribution is just the normalized counts of all possible values $\{0, 0.1, 0.2, \dots\}$. For each possible value x , the theoretical distribution is the probability of the precision range $p(x - 0.05 < X \leq x + 0.05)$ where X is the lognormal distribution to compare against. Testing on creatinine data gives a p -value of 1.0. By contrast, testing the hypothesis that

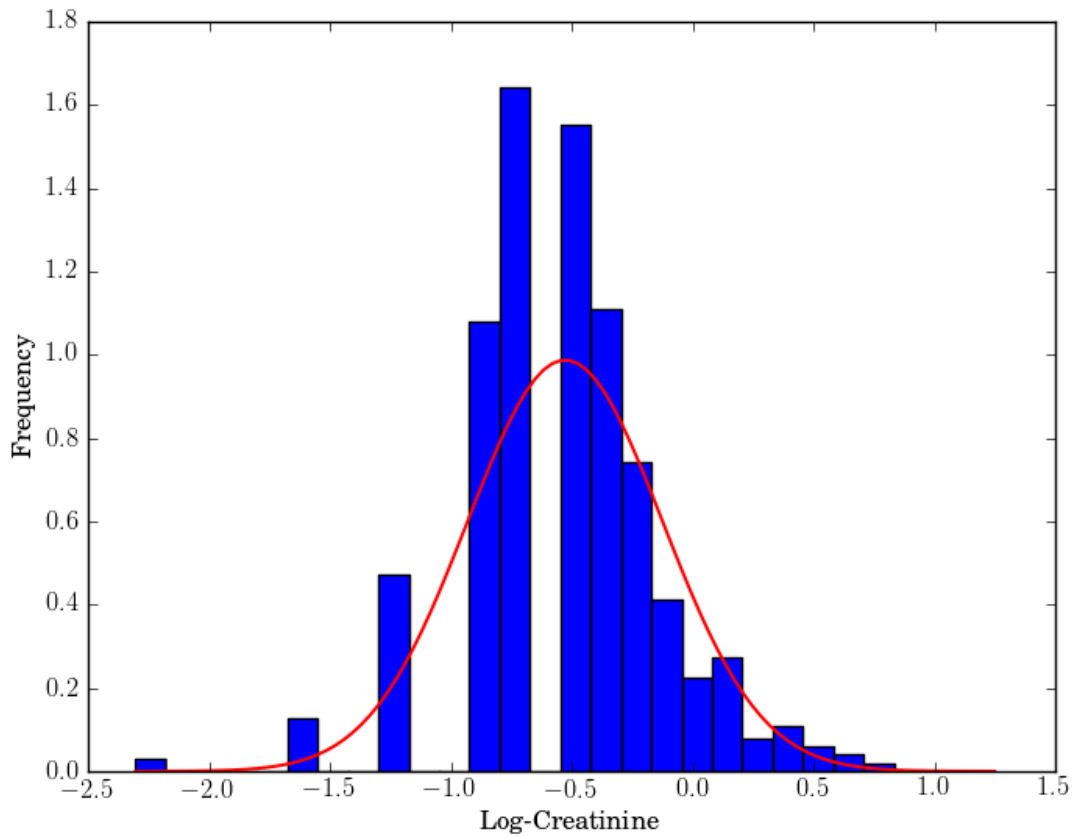


Figure 3-1: Histogram of log-creatinine with evenly spaced bins

log-creatinine has normal distribution gives p -value that is close to zero. Normality testing with other more powerful tests like the Anderson-Darling test, the Shapiro-Wilk test, or the Kolmogorov-Smirnov test also results in near zero p -value.

Parameter estimation from discretized samples should take into account the uncertainty within the precision range. One way of doing that is by using the EM Algorithm, where the undiscretized values are modeled as hidden variables [13]. Depending on the degree of discretization, ignoring that can result in poor estimates, especially if the value distribution within the precision range is highly asymmetric.

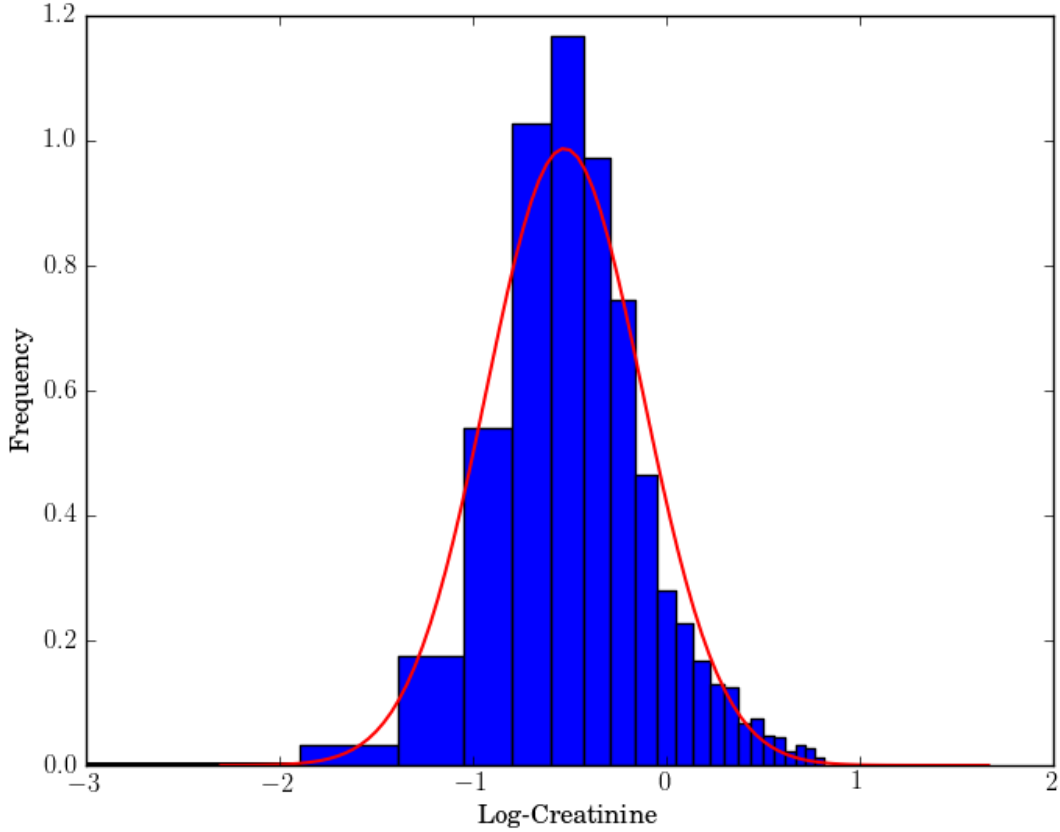


Figure 3-2: Histogram of log-creatinine with bins that respect the discretization

3.4.2 Control Groups

As the MIMIC II database only contains information on patients that are admitted to the ICU, we do not have data for other hospitalized patients who are less severely ill, not to mention healthy individuals. The lack of control groups is obviously not favorable for inference, but fortunately, this is not an issue for AKI prediction. In fact, only a small fraction of patients in the database have AKI and the rest can act as the control group. Besides, data for creatinine measurements before being admitted to the ICU are also available in the database. That is, before the overall physiological states becomes too severe. Therefore, we have sufficient negative samples for modeling normal kidney states.

3.4.3 Class Imbalance

As the number of patients with renal failure is only about a quarter of all patients, we have to worry about the imbalanced class distribution instead. Standard classification algorithms that minimize training error are vulnerable to the class imbalance problem, resulting in poor generalization performance. Several techniques have been proposed to deal with class imbalance, such as resampling and cost-sensitive learning, using different evaluation metrics [23]. Nevertheless, the proposed solutions have other issues and their effectiveness depend on the degree of imbalance, size training set, classifier involved, etc.

Our final model uses unsupervised learning, which is less affected by the imbalanced class distribution. For performance evaluation, we consider multiple evaluation metrics including the confusion matrix, area under ROC curve (AUC), sensitivity and specificity.

3.4.4 Unequal Intervals

The intervals between measurements of clinical variables are not constant, but vary depending on the patient's condition. Patients in critical condition usually have the relevant variables measured more often compared to patients in stable condition, so that timely treatment can be delivered in case the condition deteriorates.

Figure 3-3 and Figure 3-4 illustrate the distribution of creatinine measurement intervals for normal patients and renal failure patients respectively. For both groups of patients, most measurements are separated by a 24-hour interval. Observe that the histogram for renal failure patients has its frequency concentrated more on left. It has local maxima at 6 hour and 12 hour intervals, and the frequency of the two intervals are considerably higher than that in the histogram for normal patients. All these are to be expected.

The distribution of measurement intervals that are condition-dependent is a source of sampling bias when modeling with the intervals. Besides, certain intervals simply do not have enough samples for statistical analysis. That makes certain modeling task infeasible. Moreover, many time series analysis techniques that assume evenly-spaced samples will not be applicable. Hence, modeling of temporal dynamics is difficult with the available data.

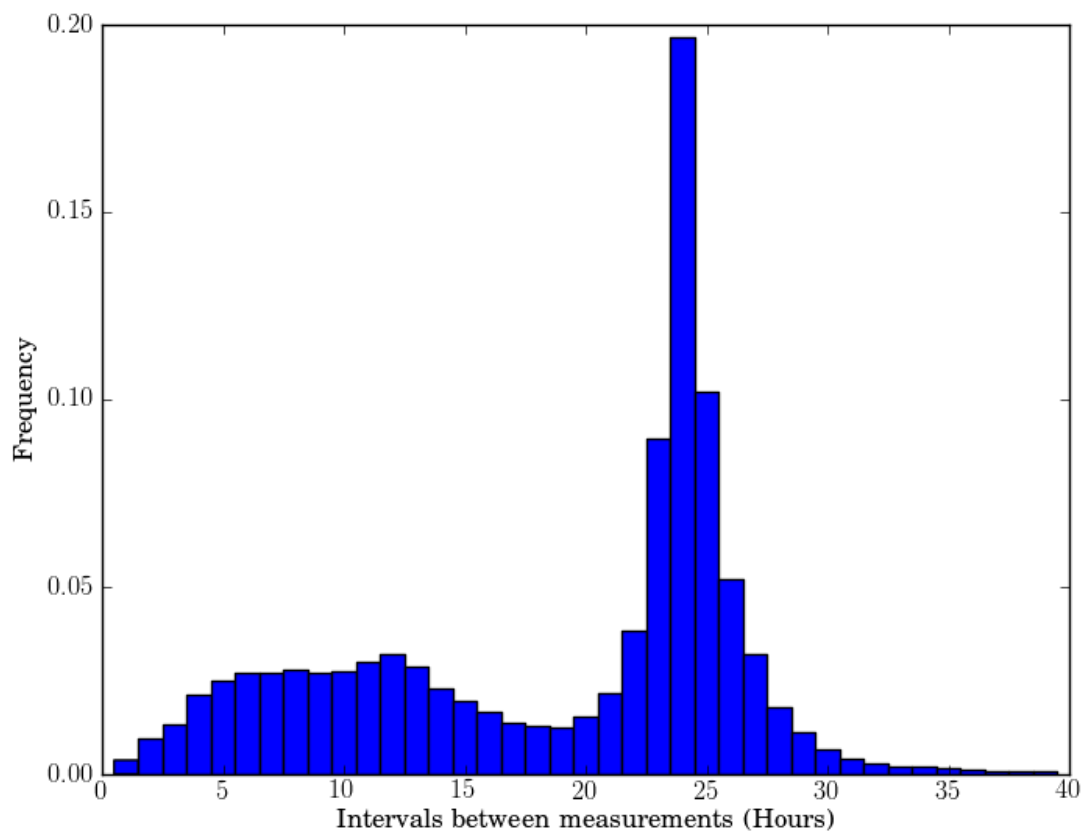


Figure 3-3: Histogram of creatinine measurements intervals for normal patients

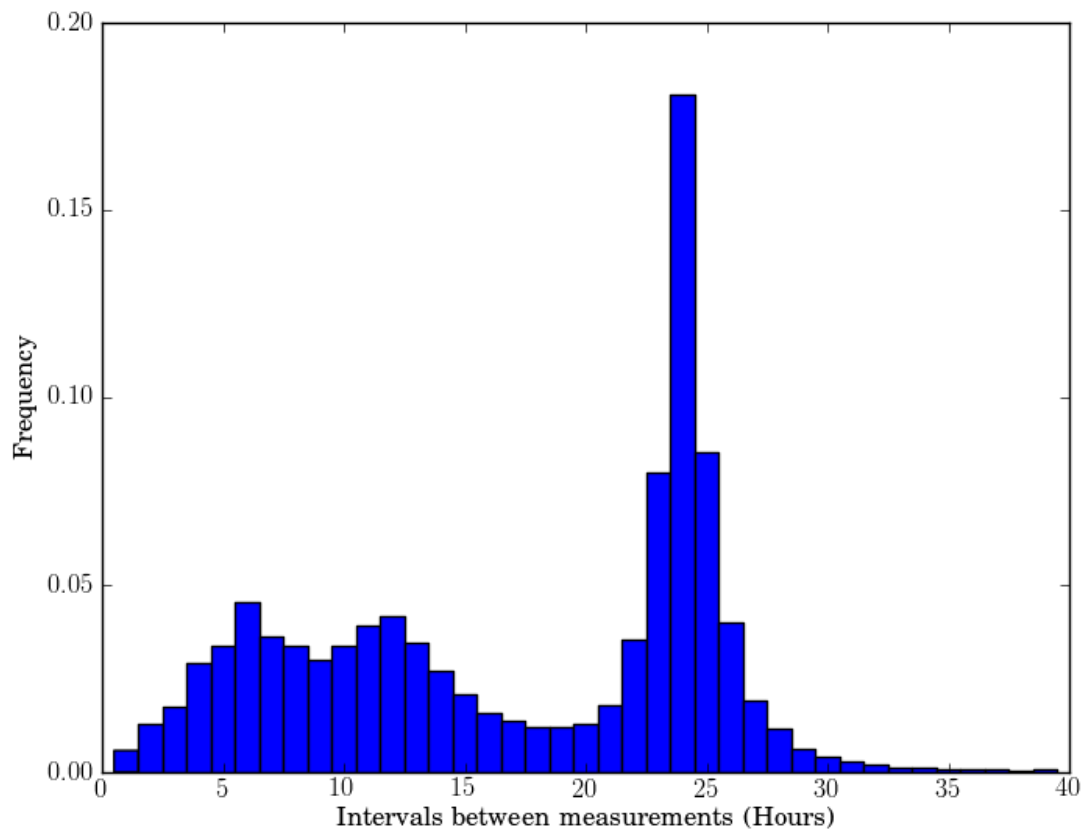


Figure 3-4: Histogram of creatinine measurements intervals for patients with renal failure

Chapter 4

Modeling Variable Kinetics

As mentioned earlier, measurement frequencies vary among clinical variables, and an important factor that affects the measurement frequency is rate at which each variable varies. After each measurement, our uncertainty in the variable necessarily increases with the elapsed time until the next measurement is made. Quantification of the random variation of the variable with time can be useful in determining the optimal measurement intervals, for timely detection of clinical deterioration.

Time evolution of substances in biological system is governed by their kinetics and the type of kinetics depends primarily on the substance type. Waste substances that are removed through the kidneys are typically assumed to follow first-order clearance kinetics, where the generation rate of the substance is constant and the removal rate is proportional to the concentration of the substance.

This chapter investigates the kinetic model for first-order clearance, as many clinical variables are associated with waste substances such as BUN and creatinine. Understanding the kinetics allows us to deduce properties of the variable that can be utilized in the design of learning algorithms. For variable with first-order clearance kinetics, the existence of a normal range, and consequently possibly multiple abnormal ranges, suggests that modeling the variable as being in one of several states is fruitful. This state abstraction assumes that for each possible state, the variable has a fixed distribution with parameters that only

depends on the current state.

4.1 First-order Clearance Kinetics

Many clinical variables are associated with the concentration of substances that follow first-order clearance kinetics, where the generation rate is constant and the clearance rate is proportional to the current concentration. Let x be such a variable and v be the total volume in which x is distributed. The mass-balance equation of x after a short interval Δt satisfies

$$\Delta(xv) = (vg - vkx)\Delta t \quad (4.1)$$

where g and k are the generation rate and clearance rate per unit volume. We can assume that there is no change in v . Taking the limit of Δt , we arrive at the following differential equation for the first-order clearance kinetics:

$$\frac{dx}{dt} = g - kx. \quad (4.2)$$

Solving the differential equation gives

$$x_t = x_0 e^{-kt} + \frac{g}{k}(1 - e^{-kt}). \quad (4.3)$$

At any given time, the variable is a weighted combination of the initial value x_0 and the equilibrium value g/k . Observe that the weight of x_0 decreases exponentially with time and x_t approaches the equilibrium value as $t \rightarrow \infty$.

4.2 Stochastic Kinetic Model

One limitation of the model above is that it does not take into account randomness in the physiological constants. In order to model the randomness, we add Gaussian white noise to

the constant g and k . The mass-balance equation after augmenting with noise is

$$\Delta x = (g\Delta t + \epsilon Z_{\Delta t}^g) + x(-k\Delta t + \sigma Z_{\Delta t}^k), \quad (4.4)$$

where $Z_{\Delta t}^g$ and $Z_{\Delta t}^k$ are independent normal random variables with mean zero and variance Δt . We use time-dependent variance to model the fact that our uncertainty increases with time. When the physiological condition is stable, we should have $\epsilon^2 \ll g$ and $\sigma^2 \ll k$, so the variance per unit time is small compared to the physiological constants.

In order to simplify the analysis of the model, we let $\epsilon = 0$, and focus only on the noise in the clearance rate. Moreover, that also ensures that $x \geq 0$ at all time, since x is the concentration of a substance and cannot be negative. Taking the limit, we arrive at the stochastic differential equation for the first-order clearance kinetics

$$dX_t = (g - kX_t) dt + (\sigma X_t) dW_t \quad (4.5)$$

where W_t is a Brownian motion. Complete derivation of the solution to the stochastic differential equation can be found in Appendix A.

Therefore, we can model the clinical variable as the stochastic process

$$X_t = X_0 \exp(-\alpha t + \sigma W_t) + g \int_0^t \exp(\alpha(s - t) - \sigma(W_s - W_t)) ds \quad (4.6)$$

where $\alpha = k + \frac{1}{2}\sigma^2$. The stochastic process X_t gives the distribution of the clinical variable after time t , conditioned on the initial value X_0 . At any given time, X_t is the sum of two random variables and we will look into each of the components.

4.2.1 Initial Value

Let

$$A_t = X_0 e^{-\alpha t} e^{\sigma W_t} \quad (4.7)$$

be the initial value component of X_t . Recall that $W_t \sim \mathcal{N}(0, t)$ and therefore, A_t is a lognormal random variable with parameter $(\ln X_0 - \alpha t, \sigma^2 t)$. We can think of A_t as a random variable centered at $X_0 e^{-kt}$ with multiplicative noise. The mean and variance of A_t are given by

$$\mathbb{E}[A_t] = X_0 e^{-kt} \quad (4.8)$$

$$\text{Var}[A_t] = X_0^2 e^{-2kt} (e^{\sigma^2 t} - 1). \quad (4.9)$$

Since $\sigma^2 \ll k$, the variance decreases exponentially to zero as $t \rightarrow \infty$.

4.2.2 Equilibrium Value

Carmona et al. showed in [7] that the variables

$$e^{-\alpha t + \sigma W_t} \int_0^t e^{\alpha s - \sigma W_s} ds \quad \text{and} \quad \int_0^t e^{-\alpha s + \sigma W_s} ds$$

have the same distribution. Thus, we let

$$B_t = g \int_0^t e^{\alpha(s-t) - \sigma(W_s - W_t)} ds \quad (4.10)$$

$$= g \int_0^t \exp(-\alpha s + \sigma W_s) ds \quad (4.11)$$

be the equilibrium value component of X_t .

The variable B_t is the time integral of a geometric Brownian motion, which has lognormal distribution for any fixed s . Yor gave the explicit density function of B_t and the formula for computing the moments in [29]. The first two moments of B_t are as follow.

$$\mathbb{E}[B_t] = \frac{g}{k} (1 - e^{-kt}) \quad (4.12)$$

$$\mathbb{E}[B_t^2] = \frac{2g^2}{k(2k - \sigma^2)(k - \sigma^2)} (ke^{-(2k - \sigma^2)t} - (2k - \sigma^2)e^{-kt} + (k - \sigma^2)) \quad (4.13)$$

Notice that as $t \rightarrow \infty$, the variance of B_t converges to the stationary value

$$\begin{aligned} V &= \mathbb{E}[B_\infty^2] - (\mathbb{E}[B_\infty])^2 \\ &= \frac{\sigma^2}{2k - \sigma^2} \left(\frac{g}{k} \right)^2. \end{aligned} \quad (4.14)$$

The expression $\text{Var}[B_t]$ is more complicated, but since $\sigma^2 \ll k$, we make the approximation

$$\mathbb{E}[B_t^2] \approx \frac{2g^2}{k(2k - \sigma^2)(k - \sigma^2)} (k(1 - e^{-kt})^2 - \sigma^2(1 - e^{-kt})). \quad (4.15)$$

The variance formula can then be simplified to

$$\begin{aligned} \text{Var}[B_t] &\approx V(1 - e^{-kt}) \left(\frac{3k - \sigma^2}{k - \sigma^2} (1 - e^{-kt}) - \frac{2k}{k - \sigma^2} \right) \\ &= V(1 - e^{-kt}) \left(1 - \frac{3k - \sigma^2}{k - \sigma^2} e^{-kt} \right). \end{aligned} \quad (4.16)$$

From the formula, we can see that the standard deviation of B_t increases from 0 to V at a rate slower than $(1 - e^{-kt})$.

The distribution B_t is commonly used in finance for pricing Asian options. As the exact density function is too sophisticated for practical purpose, B_t is often approximated using a lognormal distribution. That has been shown to work well in practice. Nevertheless, theoretical justification for the lognormal approximation does not appear until the paper by Dufresne in 2004 [15]. In particular, he showed that B_t converges in distribution to the lognormal distribution as $\sigma \rightarrow 0_+$. This is a useful result as the assumption that σ is small usually holds in biological systems.

4.2.3 Variable Properties

As X_t is the sum of two random variables for any fixed t , by linearity of expectation,

$$\begin{aligned}\mathbb{E}[X_t] &= \mathbb{E}[A_t] + \mathbb{E}[B_t] \\ &= X_0 e^{-kt} + \frac{g}{k}(1 - e^{-kt}).\end{aligned}\tag{4.17}$$

which coincides with the solution to the deterministic differential equation. The expected value converges to the equilibrium value $\frac{g}{k}$.

Recall that the weight of A_t decreases exponentially with time and that $\text{Var}[A_\infty] = 0$. As $t \rightarrow \infty$, we have $X_t \approx B_t$ and the variance converges to

$$\text{Var}[X_\infty] = \frac{\sigma^2}{2k - \sigma^2} \left(\frac{g}{k}\right)^2.\tag{4.18}$$

The stochastic process has finite variance in the long run, and our uncertainty in X_t does not grow unbounded with time. Therefore, clinical variables that follow first-order clearance kinetics should have a normal range. As long as the physiological constants doesn't change, our uncertainty in the variables is confined by the normal ranges.

Observe that the stochastic differential equation for the first-order clearance kinetics can be written as

$$dX_t = k(\mu - X_t) dt + \sigma X_t dW_t\tag{4.19}$$

where $\mu = \frac{g}{k}$ is the equilibrium value. In this form, we can see that the drift coefficient is proportional to $(\mu - X_t)$, so X_t always drifts towards μ with magnitude proportional to its deviation from μ . Indeed, the solution shows that X_t approaches the equilibrium value μ in the long run with finite variance. Therefore, clinical variables that follow first-order clearance kinetics should be mean-reverting.

4.3 State Abstraction

The mean-reverting property and the convergence of variance suggest that variable X_t has a normal range around the equilibrium value. The existence of the normal range motivates the abstraction that observations of clinical variables with first-order clearance kinetics are independent and identically distributed random variables given the states. We assume that X_t has lognormal distribution with the equilibrium value as its mean.

The states are discrete representations of physiological conditions relevant to X_t . The major difference between the states is in the physiological constants. Therefore, the distributions of X_t for different states have different parameters.

The distribution types, however, are the same for all states since they are only affected by the kinetics, and not the constants. Hence, variables that follow first-order clearance kinetics have lognormal distributions with parameters that are determined by the underlying state. It is possible that the changes in the constants also change the kinetics. Some possible instances of that include negation of the sign of k , violation of $\sigma^2 < 2k$, or $k \rightarrow 0$.

As an example, let X_t be the serum creatinine and the states be different stages of renal failure. All states have the same g , as the generation rate does not change with the progression of renal failure. The main difference between the states is the clearance rate k , as that is the main indicator of renal health. As a result, states with higher severity have higher equilibrium value due to the decrease in k .

4.4 State Transition

The previous model assumes a stable condition where g and k are constants. The stochastic kinetic model can also be applied to studying the progression of clinical deterioration, where the coefficient g and k are time-dependent.

In the context of acute renal failure, the worsening of kidney function can be modeled by a clearance rate that decreases with time $k(t)$. The stochastic differential equation is unlikely to have nice analytical solution. Nevertheless, it is still possible to simulate the trajectory by

approximating with discrete time steps Δt . That results in a sequence of random variables, where the expected values forms a trajectory that approximates the solution to the deterministic differential equation. Variance of the sequence of variables gives the uncertainty at those points along the trajectory.

Chapter 5

Temporal Dynamics of States

The underlying state of diseases are usually reflected through some clinical variables. In diagnosis of the disease, we often have to infer the states based on the observed values of those clinical variables. Changes in states can be detected from the deviations of the variables from the baseline, but it is not easy to decide whether the deviation is significant enough to conclude a state change. Our decision should also take into account the odds of state transitions.

To make things more complicated, the baseline values are different for each individual. While it is possible to estimate the baseline from past measurements, not every patient has enough data to produce reliable estimates. In the estimation of individual baseline, a model that utilizes the past measurements when the data are available, and “borrows strength from the ensemble” otherwise, would be very useful in this scenario. In addition, baseline distribution of the population can help to improve the estimates of individual baselines, as demonstrated by Stein’s paradox.

This chapter describes the generative model for the state dynamics and observations of the relevant clinical variables, followed by the algorithm for learning the model parameters from data. The trained model can then be used to infer the states given the observation sequence of the variables.

5.1 Generative Model for State Dynamics

This section develops the model for the temporal dynamics of a clinical variable X and the associated state. We assume that the state abstraction holds, so the observations of X are independent and identically distributed random variables given the states. Note that X can be multivariate and the components need not be independent.

5.1.1 Compound Sampling Observation

Consider the following compound sampling model for the distribution of X given the state. For each patient, we model X as a random variable sampled from a distribution with individual parameters that reflects the patient's state. The individual parameters of the patients are also random variables, sampled from the population distribution.

Let x_k be the observed value for patient k , who is in state $s_k = i$. We have

$$x_k \sim f(x_k \mid \theta_{k,i})$$

where $\theta_{k,i}$ is the individual parameter of patient k in state i . From the state abstraction, the individual distribution should be the same for all states, though the parameters are different. For each state i , the individual parameter $\theta_{k,i}$ is a random variable sampled from the population distribution

$$\theta_{k,i} \sim g(\theta_{k,i} \mid \eta_i)$$

where η_i is the population parameter for state i . Note that η_i is not a random variable and this notation is just to make the parameter explicit.

We choose $f(x_k \mid \theta_{k,i})$ and $g(\theta_{k,i} \mid \eta_i)$ such that the marginal distribution

$$m(x_k \mid \eta_i) = \int_{\theta_{k,i}} f(x_k \mid \theta_{k,i}) g(\theta_{k,i} \mid \eta_i) d\theta_{k,i} \quad (5.1)$$

is tractable for parameter re-estimation in the Baum-Welch algorithm [1]. A sufficient condition for that is that $f(x_k \mid \theta_{k,i})$ is a natural exponential family distribution with quadratic

variance function (NEF-QVF) and $g(\theta_{k,i} \mid \eta_i)$ is the conjugate prior. NEF-QVF is a subset of the exponential family that includes normal, Poisson, gamma, binomial and negative binomial distributions. Some interesting properties NEF-QVF are discussed by Morris in [33].

While η_i captures the population characteristics of X , the choice of population is up to the user. For instance, learning η_i from each gender separately allows us to capture the gender effect on X , assuming that we have enough data for both genders. This gives us flexibility in choosing the control variables based on our interest and the available data.

Although semantically different, the compound sampling model is very similar to Bayesian inference. However, we do not assume any prior knowledge and the inference is completely data-driven.

5.1.2 Hidden Markov Model

Let $\mathbf{s}_k = \{s_{k,1}, \dots, s_{k,T_k}\}$ be the state sequence of patient k and $\mathbf{x}_k = \{x_{k,1}, \dots, x_{k,T_k}\}$ be the corresponding observations. As the states are unknown, we model the temporal structure of the states and observations as a Hidden Markov Model (HMM) with M states. This model will be used for learning the population parameters and, eventually, for inference of the hidden states.

The hidden state sequence is modeled as a Markov chain with transition probability $p(s_{k,t} \mid s_{k,t-1})$. Given the state $s_{k,t}$, the corresponding observation $x_{k,t}$ is distributed according to the compound sampling model with

$$p(x_{k,t} \mid s_{k,t} = i, \theta_{k,i}) = f(x_{k,t} \mid \theta_{k,i}), \quad p(\theta_{k,i}) = g(\theta_{k,i} \mid \eta_i).$$

Let $z_{k,t}^i = \mathbb{1}\{s_{k,t} = i\}$ be the indicator variables for states. The complete likelihood for patient k can be written as

$$p(\mathbf{s}_k, \mathbf{x}_k, \boldsymbol{\theta}_k \mid \boldsymbol{\eta}) = p(s_{k,1}) \left\{ \prod_{t=2}^{T_k} p(s_{k,t} \mid s_{k,t-1}) \right\} \left\{ \prod_{i=1}^M g(\theta_{k,i} \mid \eta_i) \prod_{t=1}^{T_k} f(x_{k,t} \mid \theta_{k,i})^{z_{k,t}^i} \right\} \quad (5.2)$$

Ideally, we would like to find $\boldsymbol{\eta} = \{\eta_i\}$ that maximizes the observed likelihood

$$p(\mathbf{x}_k \mid \boldsymbol{\eta}) = \sum_{\mathbf{s}_k} \int_{\boldsymbol{\theta}_k} p(\mathbf{s}_k, \mathbf{x}_k, \boldsymbol{\theta}_k \mid \boldsymbol{\eta}) d\boldsymbol{\theta}_k. \quad (5.3)$$

However, as marginalization of \mathbf{s}_k is intractable, we resort to the EM algorithm with \mathbf{s}_k as latent variable. We cannot leave $\boldsymbol{\theta}_k$ as a latent variable and we will describe the problem with that later. Marginalizing $\boldsymbol{\theta}_k$ gives the observation distribution

$$p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i) = \int_{\theta_{k,i}} \prod_{t=1}^{T_k} f(x_{k,t} \mid \theta_{k,i})^{z_{k,t}^i} g(\theta_{k,i} \mid \eta_i) d\theta_{k,i}. \quad (5.4)$$

Marginalization of the common parameters coupled all observations of the same states together. Thus, the observations are no longer independent given the states. This makes parameter estimation with the EM algorithm intractable.

5.1.3 Variational Inference

In order to circumvent the tractability issue after marginalization, we approximate the observation distribution with the closest tractable distribution. In particular, we approximate the joint distribution as factorizable into

$$p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i) \approx \prod_{t=1}^{T_k} p(x_{k,t} \mid s_{k,t}, \eta_i) \quad (5.5)$$

where $p(x_{k,t} \mid s_{k,t}, \eta_i)$ are to be determined. Let

$$v(\mathbf{x}_k \mid \eta_i) = \prod_{t=1}^{T_k} p(x_{k,t} \mid s_{k,t}, \eta_i) \quad (5.6)$$

$$= \prod_{t=1}^{T_k} v_t(x_{k,t} \mid \eta_i)^{z_{k,t}^i}. \quad (5.7)$$

By variational inference, we choose $v(\mathbf{x}_k)$ that minimizes the Kullback-Leibler divergence:

$$v(\mathbf{x}_k \mid \eta_i) = \arg \min_v D_{KL}(v(\mathbf{x}_k \mid \eta_i) \parallel p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i)) \quad (5.8)$$

Observe that $v_t(x_{k,t} \mid \eta_i)$ can be any distribution if $z_{k,t}^i = 0$. Without loss of generality, we assume that $z_{k,t}^i = 1$ for all t . The minimization gives

$$\ln v_t(x_{k,t} \mid \eta_i) \propto \mathbb{E}_{-x_{k,t}}[\ln p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i)]$$

where $\mathbb{E}_{-x_{k,t}}[\ln p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i)]$ denotes the expectation over $\ln p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i)$ with respect to all the variables except for $x_{k,t}$. This can be simplified to

$$v_t(x_{k,t} \mid \eta_i) \propto \exp \left(\int_{\theta_{k,i}} \ln(f(x_{k,t} \mid \theta_{k,i})g(\theta_{k,i} \mid \eta_i)) d\theta_{k,i} \right) \quad (5.9)$$

$$= \exp \left(\int_{\theta_{k,i}} \ln p(\theta_{k,i} \mid x_{k,t}) + \ln m(x_{k,t} \mid \eta_i) d\theta_{k,i} \right) \quad (5.10)$$

$$= m(x_{k,t} \mid \eta_i) \quad (5.11)$$

where $m(x_{k,t} \mid \eta_i)$ is the marginal distribution of a single observation. As a result, the variational solution gives the approximation

$$p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i) \approx \prod_{t=1}^{T_k} m(x_{k,t} \mid \eta_i)^{z_{k,t}^i} \quad (5.12)$$

The marginal likelihood for each patient is then

$$p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta}) = p(s_{k,1}) \prod_{t=2}^{T_k} p(s_{k,t} \mid s_{k,t-1}) \prod_{i=1}^M p(\mathbf{x}_k \mid \mathbf{s}_k, \eta_i) \quad (5.13)$$

$$\approx p(s_{k,1}) \left\{ \prod_{t=2}^{T_k} p(s_{k,t} \mid s_{k,t-1}) \right\} \left\{ \prod_{i=1}^M \prod_{t=1}^{T_k} m(x_{k,t} \mid \eta_i)^{z_{k,t}^i} \right\}. \quad (5.14)$$

5.2 Learning Algorithm

Combining the previous discussions, we now have a tractable model for the temporal dynamics of states and observations. This section describes the estimation of population parameters using the EM algorithm.

5.2.1 EM Algorithm

The expectation-maximization (EM) algorithm is an iterative procedure for finding the maximum likelihood estimates of parameters in models with latent variables. More details about the EM algorithm can be found in Dempster et al. [13].

In our case, the latent variables are the hidden states \mathbf{s}_k . The log-likelihood for each patient can be written as

$$\begin{aligned}\ell(\boldsymbol{\eta}; \mathbf{y}_k) &= \ln \sum_{\mathbf{s}_k} p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta}) \\ &= \ln \sum_{\mathbf{s}_k} q(\mathbf{s}_k \mid \mathbf{x}_k) \frac{p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta})}{q(\mathbf{s}_k \mid \mathbf{x}_k)} \\ &\geq \sum_{\mathbf{s}_k} q(\mathbf{s}_k \mid \mathbf{x}_k) \ln \frac{p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta})}{q(\mathbf{s}_k \mid \mathbf{x}_k)} \\ &= \mathbb{E}_q \left[\ln \frac{p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta})}{q(\mathbf{s}_k \mid \mathbf{x}_k)} \right] \\ &= \tilde{\ell}(q, \boldsymbol{\eta})\end{aligned}$$

where the inequality follows from Jensen's inequality. The algorithm initializes the parameter estimates to some values $\boldsymbol{\eta}^{(1)}$, then repeats the following steps for $t = 1, 2, \dots$ until convergence.

E-step Find

$$q^{(t+1)} = \arg \max_q \tilde{\ell}(q, \boldsymbol{\eta}^{(t)})$$

M-step Update the parameter estimates to

$$\boldsymbol{\eta}^{(t+1)} = \arg \max_{\boldsymbol{\eta}} \tilde{\ell}(q^{(t+1)}, \boldsymbol{\eta})$$

The sequence of parameter estimates $\boldsymbol{\eta}^{(t)}$ increase the lower bound of the log-likelihood in each iteration until the local maximum is reached.

5.2.2 Estimation of Population Parameters

The optimal $q^{(t+1)}$ in the E-step can actually be solved explicitly to give

$$q^{(t+1)} = p(\mathbf{s}_k \mid \mathbf{x}_k, \boldsymbol{\eta}^{(t)}).$$

According to the Rao-Blackwell theorem, the optimal parameter in M-step will be some function of the sufficient statistics. Therefore, the E-step can be reduced to computing the posterior expectation over the sufficient statistics.

Here is the final algorithm for learning the population parameters of the model.

Algorithm 1 Algorithm for learning the population parameter

Initialization: set $\boldsymbol{\eta}^{(t)} = \boldsymbol{\eta}^{(1)}$ and $t = 1$.
while $t < T_{\max}$ **and** $|\boldsymbol{\eta}^{(t)} - \boldsymbol{\eta}^{(t-1)}| > \epsilon$ **do**
 Evaluate $p(x_{k,t} \mid s_{k,t} = i) = m(x_{k,t} \mid \eta_i^{(t-1)})$ for each state.
 Compute $p(s_{k,t} = i \mid \mathbf{x}_k)$ using the forward-backward algorithm.
 Collect the sufficient statistics $\boldsymbol{\gamma}_i$ for each state.
 Compute the optimal parameter estimates from $\boldsymbol{\gamma}_i$.
 Set $t = t + 1$ and update the estimates $\boldsymbol{\eta}^{(t)}$.
end while

Suppose we are interested in estimating the model parameters for a population that consists of patient $1, 2, \dots, K$. We can train the model on these patients' data to maximize the log-likelihood of the population

$$\ell(\boldsymbol{\eta}) = \sum_{k=1}^K \ln p(\mathbf{s}_k, \mathbf{x}_k \mid \boldsymbol{\eta}) \quad (5.15)$$

using Algorithm 1. That gives the estimates for the population parameters, initial state distribution, and the transition probability of the Hidden Markov Model. Now that we have the full model, the model can be used for prediction.

5.3 Prediction

This section applies the model to the inference of individual distributions and state sequences. Inference of individual distributions is a general case of baseline estimation, since the distributions contain baseline information for each states. Decoding of the most-likely state sequence use the individual distributions for variable observations in the Viterbi algorithm.

With the estimates of the population parameter $\hat{\eta}_i$, the posterior distribution of the individual parameter $\theta_{k,i}$ given the observations \mathbf{x}_k is

$$p(\theta_{k,i} \mid \mathbf{x}_k, \hat{\eta}_i) \propto \prod_{t=1}^{T^k} f(x_{k,t} \mid \theta_{k,i})^{z_{k,t}^i} g(\theta_{k,i} \mid \hat{\eta}_i). \quad (5.16)$$

For evaluating the posterior distribution, the variables $z_{k,t}^i$ are substituted by the estimator $\mathbb{E}[z_{k,t}^i \mid \mathbf{x}_k, \hat{\eta}_i]$, which can be computed using the forward-backward algorithm.

With the posterior distribution of the individual parameter, the predictive distribution of future observations given the state is then

$$p(x_k \mid s_k = i, \mathbf{x}_k, \hat{\eta}_i) = \int_{\theta_{k,i}} f(x_k \mid \theta_{k,i}) p(\theta_{k,i} \mid \mathbf{x}_k, \hat{\eta}_i) d\theta_{k,i}. \quad (5.17)$$

The predictive distribution is tractable, since $p(\theta_{k,i} \mid \mathbf{x}_k, \eta_i)$ will have the same form as the population distribution $g(\theta_{k,i} \mid \eta_i)$.

5.3.1 Baseline Estimation

Deviations of clinical variables from the baseline are used as diagnostic criteria for some diseases. For instance, the stages of acute kidney injury are actually defined in terms of the

increase in creatinine from the baseline. Despite its definitive role in diagnosis, a universal definition of baseline creatinine has yet to emerge. Several definitions have been used in various studies, such as the creatinine at the time of hospital admission, the minimum creatinine value during the hospital stay, the creatinine estimated from the Modification of Diet in Renal Disease (MDRD) formula or the lowest value among these [20].

We consider a probabilistic interpretation of the baseline for variable X_k . Let i be the normal state. Our belief in the baseline for patient k after we observed \mathbf{x}_k is fully characterized by $p(x_k \mid s_k = i, \mathbf{x}_k, \eta_i)$. The posterior distribution naturally combines the population baseline with past observations. It is also consistent with our hope that as more past observations become available, the influence of the population baseline will also decrease.

Therefore, the posterior distribution itself is a good description of the baseline. Another description that is more concise is the expected value of x_k under this distribution. If the distribution is lognormal, the median might be another good representation of the baseline [27].

5.3.2 Inference of States

After training the model on the dataset, the model can be used for state prediction. Given a sequence of observations for a patient, the most likely state sequence can be inferred using the Viterbi algorithm [40]. If history of past observations is available, the predictive distribution $p(x_k \mid s_k = i, \mathbf{x}_k, \hat{\eta}_i)$ is used as the observation distribution of x_k . Otherwise, we just use the marginal distribution $m(x_k \mid \hat{\eta}_i)$.

However, we need to be careful in the computation of the predictive distribution from previous observations. Observations that are too old should be discarded, as the individual parameters might have changed due to aging.

5.4 Discussion

We mentioned that the individual parameters cannot be treated as latent variables. Zhang et al. [52] suggested an evidence framework for learning Bayesian HMM with an exponential family distribution for the observations. They also use the EM algorithm to optimize the likelihood, but θ_k is treated as latent variables rather than being marginalized.

The problem with their approach is that the algorithm estimates $\theta_{k,i}$ for every i and k in the E-step. In the M-step, the population parameter η_i is estimated from sufficient statistics that depend only on $\theta_{k,i}$, and not \mathbf{x} . Since the estimation of $\theta_{k,i}$ is entirely based on \mathbf{x}_k , that gives horrible estimates if patient k is never in state i . Moreover, the estimation of η_i gives equal weight to each $\theta_{k,i}$. So the estimator would be highly biased if a substantial fraction of patients are never in state i .

A necessary condition for this model to work well is that for each state, the number of patients that are never in that state is small. Consequently, we would have to marginalize the individual parameters as that condition is far from the truth in our dataset.

We have developed a generative model for inference of kidney states given observations of the relevant clinical variables. Baseline estimation based on available observations is a common clinical problem with practical importance. The two main challenges in baseline estimation are the lack of past observations for certain patients and the unknown state sequence. We tackled the problem of insufficient data by utilizing the idea of “borrowing strength from the ensemble” in statistics. In that way, patients with limited observations data can borrow baseline information of the population. Besides, our method is an unsupervised learning and the true state sequence is not required as an input. Another popular unsupervised learning technique that can be used is clustering, which clusters the observations into different states. Our approach has the advantage of being able to exploit the temporal structure of the data. Clustering does not take the temporal dependency between the observations and just assign each observation to the most likely state. In addition, our model is generative while the clustering model is not.

The concept of individual distribution is more general than the variable baseline. Ac-

According to our model, the baseline can be assumed to be the posterior mean of the variable, which can be easily calculated. The model also tells us the uncertainty associated with the baseline estimates based on the number of observations we have.

Chapter 6

Results

This chapter discusses the application of the models developed in previous chapters to the diagnosis of acute kidney injury (AKI). Renal health is mainly reflected through serum creatinine and urine output. Therefore, changes in the two variables have been used to define the stages of acute kidney injury in the RIFLE classification system.

Due to various problems with urine output, we are only using serum creatinine for our modeling. The reasons for not using urine output for renal failure prediction is explained in the next section.

As discussed earlier, creatinine kinetics can be modeled as a first-order clearance model and this model has been applied in a number of studies [32, 46]. We modeled creatinine kinetics using the stochastic kinetic model developed in Chapter 4 and consider some of its properties. We managed to find empirical evidences that can verify some statistical properties predicted by the model.

The model for state dynamics presented in Chapter 5 is used to model the stages of renal dysfunction in RIFLE. The expected value of the creatinine distribution for each state is computed and compared to the RIFLE criteria. Finally, performance of the model in prediction is assessed based on the ground truth.

6.1 Urine Output

Although urine output is one of the staging criteria of the RIFLE system, we decided not to use urine output for our predictive modeling. The reasons for that are given below.

Some members of the Acute Kidney Injury Network (AKIN) felt that urine output is not specific enough for the designation of AKI. It is known that the urine volume could also be influenced by the factors other than renal health, such as hydration state, use of diuretics, and presence of obstruction. Thus any decrease in urine may not be attributed to renal dysfunction because there are several other possibilities. Therefore, the use of urine output in diagnosis requires careful consideration of the clinical context.

The urine criterion for the Risk stage in RIFLE is urine volume less than 0.5 ml/kg per hour for 6 hours. Erdbruegger, et al. noted in [17] that for a 70 kg male adult, this represents a urine volume of 210 ml in 6 hours which, if maintained, would be 840 ml/day. By limiting fluid intake, many healthy individuals could meet this criterion. Thus, the authors discourage diagnosis of AKI based solely upon urine output.

It is worth noting that the urine output criteria are not in balance with the corresponding creatinine criteria and are too sensitive [41]. In other words, Risk patients defined by creatinine criteria may be more severely ill compared to those defined by urine output criteria. Serum creatinine has also been shown to be a better predictor of mortality than urine output [41].

Diagnosis using urine output has low clinical utility because accurate measurement of urine output is difficult in non-intensive care unit settings. Moreover, in order to get a robust estimate of the average urine volume, urine collection has to be done more than once over the span of several hours. This is more troublesome compared to the measurements of creatinine.

The urine volume can have large variation even for the same individual, as it is also affected by fluctuations in total fluid consumed and the frequency of urination. That causes the variable to have low signal to noise ratio. By looking at the histogram, the empirical distribution of urine output for the population is not smooth either. That makes it hard to

model the variable with any known probability distribution.

Because measurement of urine output is difficult outside the ICU, the MIMIC II database only contains the urine output of the patients during their ICU stay. On the other hand, creatinine values for the whole hospital admission, or even outside the hospital, are included in the database. So there are more data for creatinine than for urine output that are available for machine learning.

More importantly, the available ground truths are for each hospital admission. If the patients had renal failure only when outside the ICU, the available urine output data would not reflect the ground truth. It is well recognized that uncomplicated AKI can usually be managed outside the ICU [30]. Thus, AKI patients in the ICU are likely to have some other problems as well. Learning solely from the data collected during an ICU stay might introduce the issue of selection bias.

We compared several supervised learning models to identify acute kidney injury using each of the following feature sets: urine output only, creatinine only, urine output and creatinine. The true labels were extracted from ICD9 codes and clinical notes. It turned out that the model constructed using only creatinine gave the best performance, which strengthens our belief that urine output has low predictive power due to the various problems mentioned above.

For the ease of modeling, we rely on serum creatinine alone for the prediction of renal failure.

6.2 Creatinine Kinetics

We modeled creatinine kinetics using the stochastic first-order clearance model, and looked at the changes in creatinine over a short time scale where the kinetic effect is visible. Our goal is to validate the applicability of the model to creatinine. Empirical properties of the short-term changes in creatinine are compared to the properties predicted by the stochastic kinetic model.

Parameter estimation for the stochastic kinetic model is difficult due to the uneven distri-

bution of measurement intervals. Most of the creatinine measurements are within 24 hours from the last measurement. Measurements with shorter intervals are usually associated with sick patients, which causes sample selection bias. Nevertheless, if we had better data, estimation of the clearance rate k would be very useful in detecting renal failure.

To investigate the reasonableness of the stochastic kinetic model, we examine how it fits data for adult patients with age between 20 and 75 years old, without renal failure. We will return to including patients with renal failure when applying the model for state dynamics in Section 6.3.

6.2.1 Heuristic Interpretation

Let X_t be the value of creatinine, which is assumed to follow first-order clearance kinetics. Recall that X_t satisfies the stochastic differential equation

$$dX_t = (g - kX_t) dt + (\epsilon + \sigma X_t) dW_t.$$

As we are interested in the short-term changes in X_t , we consider the following heuristic interpretation of the stochastic differential equation. For a short time interval Δt , we have

$$\Delta X_t \approx (g - kX_t)\Delta t + (\epsilon + \sigma X_t) Z_{\Delta t} \tag{6.1}$$

where $Z_{\Delta t} \sim \mathcal{N}(0, \Delta t)$. Hence, ΔX_t can be approximated by a normal random variable with mean $(g - kX_t)\Delta t$ and standard deviation $(\epsilon + \sigma X_t)\sqrt{\Delta t}$.

We then compare the empirical drift and diffusion of ΔX_t to this interpretation. We barely have enough samples that have measurement intervals of 6 hours, not to mention samples with shorter measurement intervals. Therefore, we focus on model validation by comparing to the empirical trend rather than parameter estimation. It is also unclear whether this is a good approximation with Δt being 6 hours.

6.2.2 Mean-reverting Drift

The expected value of ΔX_t is proportional to the drift coefficient $(g - kX_t)$. Figure 6-1 shows the plot of the sample mean of ΔX_t of normal patients against X_t with $\Delta t = 6$ hours. We can see the the sample mean decreases linearly with X_t , which is consistent with the stochastic model.

The population average for μ corresponds to the point with zero mean drift, which is slightly below 0.6 mg/dl. The average drift is positive for $X_t < 0.6$ and negative otherwise. As the normal range of creatinine is 0.5 - 1.2 mg/dl, the distribution of X_t within this range has positive skew.

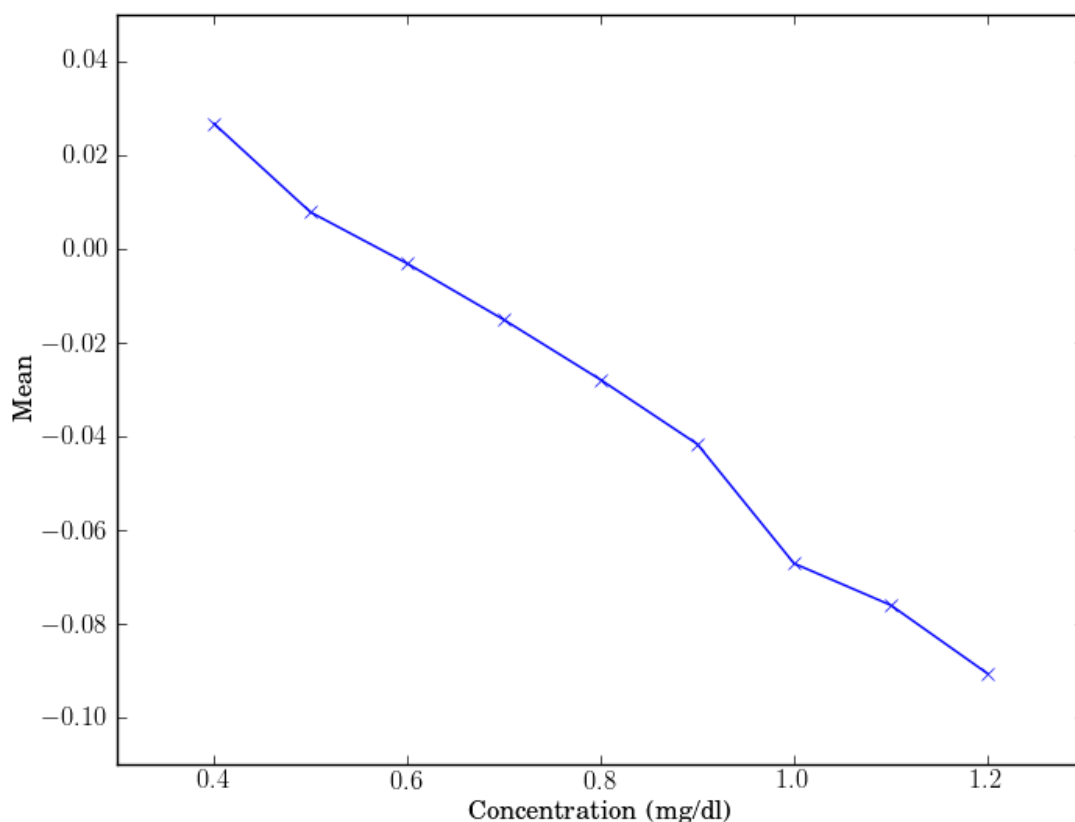


Figure 6-1: Mean of the change in creatinine for different values of creatinine

6.2.3 Linear Diffusion

The standard deviation of ΔX_t is proportional to the diffusion coefficient $\epsilon + \sigma X_t$. We plotted the standard deviation of ΔX_t samples of normal patients against X_t with Δt of 6 hours in Figure 6-2. The sample standard deviation of ΔX_t increases linearly with X_t , which agrees with the model. The multiplicative noise makes the lognormal distribution a good model for the X_t .

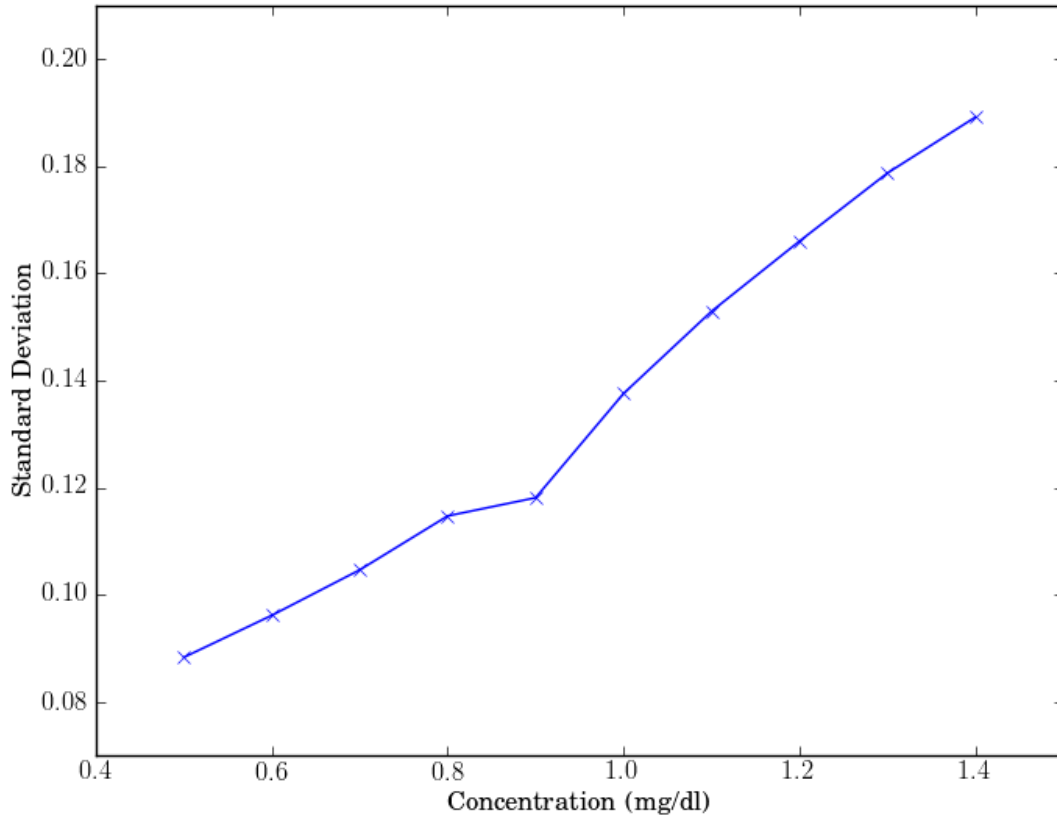


Figure 6-2: Standard deviation of the change in creatinine for different values of creatinine

6.2.4 Convergence of Variance

Figure 6-3 shows the plot of the standard deviation of ΔX_t of normal patients against measurement intervals. Due to the limited data, we plotted samples for every three hours

so the variance at $t = 2, 5, 8, \dots$ are estimated from samples with intervals within $t \pm 1$ hour. The standard deviation of X_t and ΔX_t are the same, because they only differ by a constant X_0 . From the plot, we can see that the standard deviation converges to a finite value as predicted by the stochastic kinetic model.

The decay rate of e^{-kt} can actually be roughly approximated from the rate of increase of the standard deviation. From Equation 4.16, the standard deviation increases at an asymptotic rate that is slightly slower than $(1 - e^{-kt})$. We can see that the increase in standard deviation is negligible after 15 hours, implying that e^{-kt} would be very small by then. That might give us some confidence in the independence assumption of state abstraction, as most values are separated by 24-hour intervals.

An important point to note is that the empirical variance for small Δt may be larger than the theoretical value due to the selection bias problem. Patients with relatively higher measurement frequency for creatinine may indicate an unstable renal condition that the physician perceives as at risk of having AKI. Therefore, the variation in creatinine is probably higher compared to normal patients.

The convergence rate of the variance of X_t might be helpful for determining the measurement interval. As the variance saturates after about 15 hours from the measurement, we do not have more information about X_t beyond what we know about the normal range. Hence, daily measurement works fine for normal patient; for patients at risk however, intervals shorter than 12-hour might be more appropriate.

6.2.5 Generation and Clearance

From the kinetic model, creatinine is a mean-reverting process that fluctuates around the mean $\mu = g/k$, and the mean-reverting property agrees with the empirical result as shown in Figure 6-1. Let v be the total body water. The mean can be written as

$$\mu = \frac{gv}{kv} = \frac{g'}{k'}$$

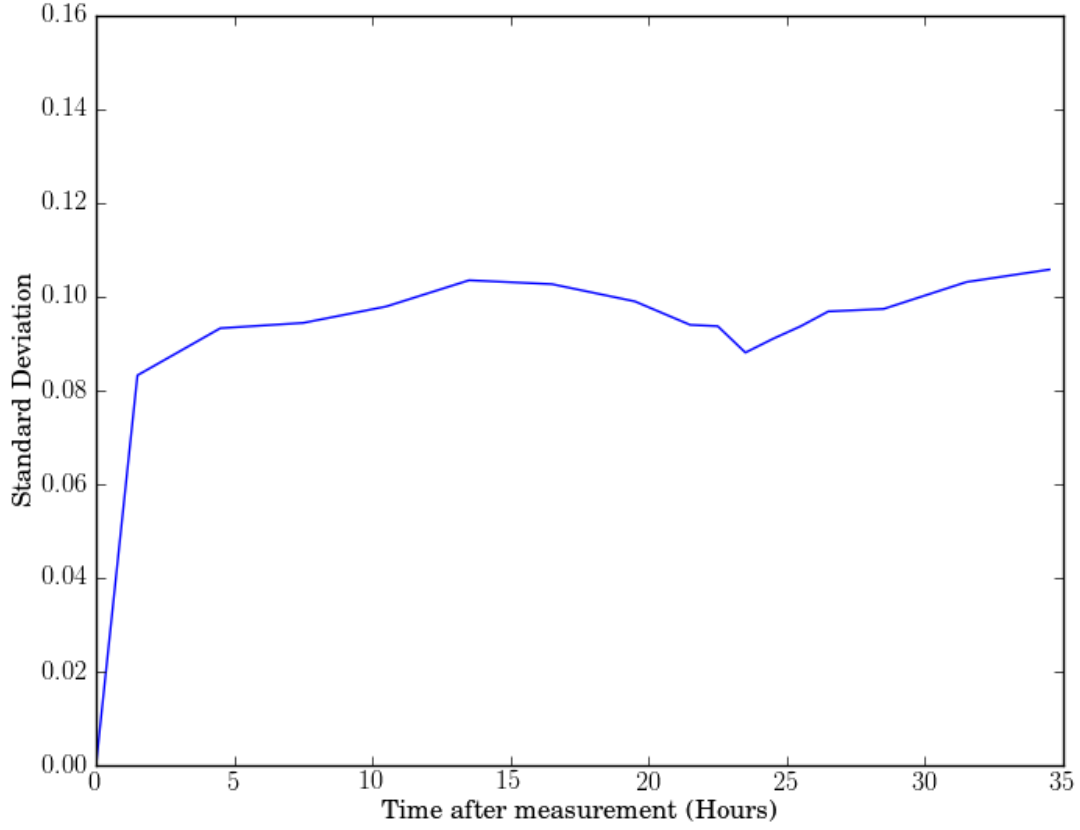


Figure 6-3: Standard deviation of the change in creatinine for different durations after measurement

where g' and k' are the generation and clearance rate of creatinine respectively. Both quantities have been extensively studied by several authors.

The most widely-used formula for estimating creatinine clearance rate is the Cockcroft-Gault formula, proposed by Cockcroft and Gault in [11]. The formula gives

$$k' = \frac{(140 - \text{Age}) \times (\text{Mass in kg}) \times (0.85 \text{ if Female})}{75 \times (\text{Serum Creatinine in mg/dl})} \quad (6.2)$$

where k' has unit ml/min. Using our model, this formula can be interpreted as estimating

$k' = g'/\mu$ with

$$g' = (140 - \text{Age}) \times (\text{Mass in kg}) \times (0.85 \text{ if Female}) \div 75$$

$$\mu = (\text{Serum Creatinine in mg/dl}).$$

The model also suggests that the mean of serum creatinine is proportional to the generation rate g' , which is consistent with the work of Clark, et al. [9] and Shinzato, et al. [44] on creatinine generation rate.

6.3 Kidney State Dynamics

This section applies the generative model for state dynamics of Chapter 5 to model the progression of acute kidney injury. Given the creatinine values as an observation sequence, our goal is to infer the disease state, for classifying the stage of acute kidney injury.

For this experiment, we restrict our dataset to adult patients aged between 20 and 75 years. Each gender is trained separately so that the estimated population parameters capture the characteristics of each gender. The size of the final dataset is summarized in Table 6.1

	Female	Male
Patients	3331	4990
Admissions	3692	5326
Samples	77916	96186

Table 6.1: The number of patients, hospital admissions, and creatinine samples for each gender.

Different hospital admissions of the same patient that are separated by too long are treated as different patients because the individual parameters may have changed. We break the observation sequence of each patient at points where the time interval between consecutive observations exceeds 6 months.

The states are chosen based on the RIFLE and AKIN classification system. In particular, we let the states be Normal, Risk, Injury, and Failure, to correspond to the three levels of

renal dysfunction in both systems.

6.3.1 Lognormal Model for Creatinine

From the stochastic kinetic model, the distribution of creatinine for each state can be modeled using a lognormal distribution. Let X_k be the log of creatinine. The individual distribution and population distribution of X_k are

$$f(x_k \mid \lambda_{k,i}) = \mathcal{N}(\lambda_{k,i}, \sigma_i^2), \quad g(\lambda_{k,i} \mid \mu_i) = \mathcal{N}(\mu_i, \tau_i^2).$$

The variances of individual distributions are assumed to be the same for all patients. Marginalizing the individual parameter $\lambda_{k,i}$ gives the marginal distribution

$$m(x_k \mid \mu_i) = \mathcal{N}(\mu_i, \tau_i^2 + \sigma_i^2).$$

Using our model for state dynamics with 4 states, the parameters μ_i and $\tau_i^2 + \sigma_i^2$ for each state can be estimated from the data. Subtracting a consistent estimate of σ_i^2 from the marginal variance gives the estimates of τ_i^2 .

The posterior distribution $p(\lambda_{k,i} \mid \mathbf{x}_k, \mu_i)$ is a normal distribution with parameter

$$\mathbb{E}[\lambda_{k,i} \mid \mathbf{x}_k, \mu_i] = B_k \mu_i + (1 - B_k) \bar{x}_k \tag{6.3}$$

$$\text{Var}[\lambda_{k,i} \mid \mathbf{x}_k, \mu_i] = B_k \tau^2, \tag{6.4}$$

where

$$B_k = \frac{\sigma^2}{\sigma^2 + \tau^2 \sum_t z_{k,t}^i}, \quad \bar{x}_k = \frac{\sum_t z_{k,t}^i y_{k,t}}{\sum_t z_{k,t}^i}.$$

The variables $z_{k,t}^i$ are replaced with the estimator $\mathbb{E}[z_{k,t}^i \mid \mathbf{x}_k, \mu_i]$, which can be computed using the forward-backward algorithm. The estimators $\mathbb{E}[z_{k,t}^i \mid \mathbf{x}_k, \mu_i] = p(s_{k,t} = i \mid \mathbf{x}_k, \mu_i)$ can be thought of as the “soft count” for the discrete samples.

The predictive distribution for observations of X_k is then

$$\begin{aligned} p(x_k \mid s_k = i, \mathbf{x}_k, \mu_i) &= \int_{\lambda_{k,i}} f(x_k \mid \lambda_{k,i}) p(\lambda_{k,i} \mid \mathbf{x}_k, \mu_i) d\lambda_{k,i} \\ &= \mathcal{N}(B_k \mu_i + (1 - B_k) \bar{x}_k, \sigma_i^2 + B_k \tau^2) \end{aligned} \quad (6.5)$$

With the predictive distribution and the parameter estimates, inference of the patients' states can be done by the Viterbi algorithm.

6.3.2 Goodness of Fit

The marginal distribution of x_k can be verified by comparing the empirical distribution of the creatinine of patients in the same state to the lognormal distribution. Figure 6-4 shows the histogram of creatinine of normal patients and the density function of the lognormal distribution with parameters estimated by maximum likelihood.

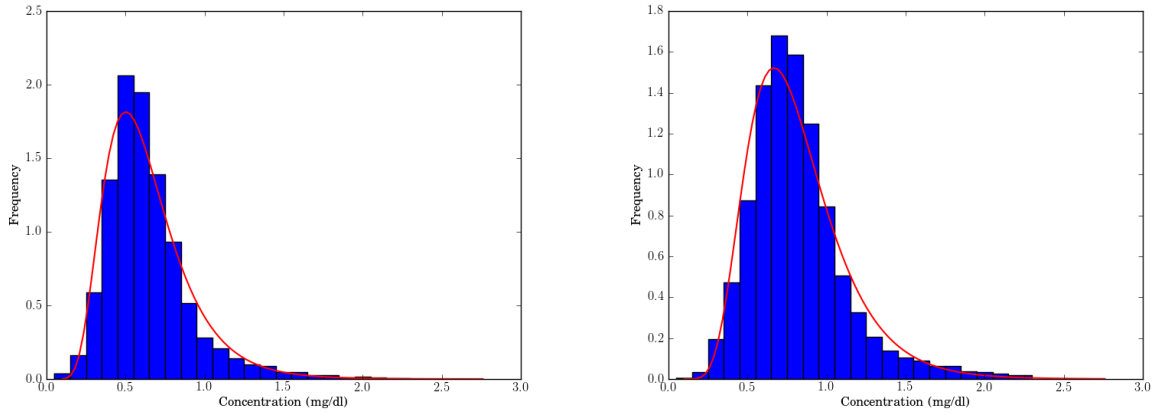


Figure 6-4: Histogram of creatinine values of normal patients and the probability density functions of the lognormal distribution. Left: female, Right: male

The probability plot comparing the empirical distribution of creatinine to the lognormal distribution is shown in Figure 6-5. The data fits the lognormal distribution quite closely except for some outliers at the high end of the range. Those outliers probably belong to other states.

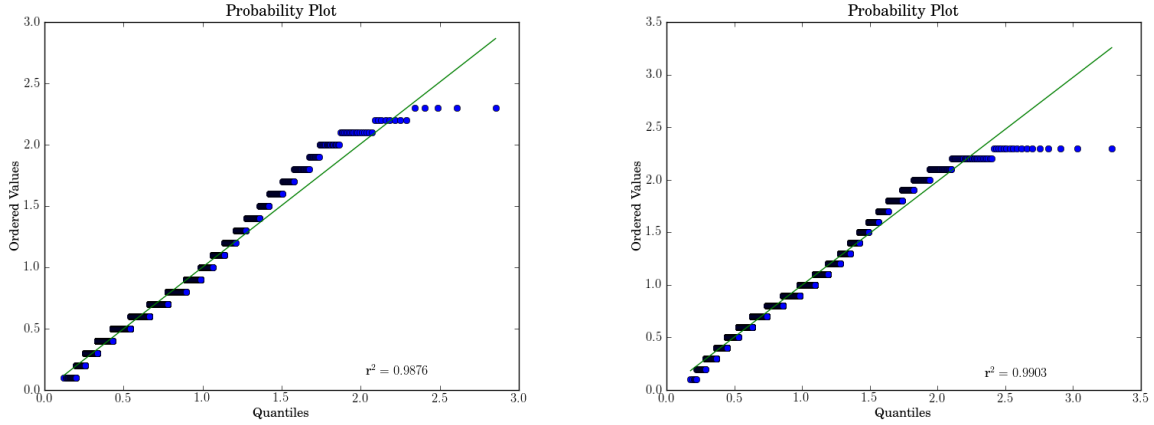


Figure 6-5: Probability plot of creatinine values of normal patients versus the lognormal distribution. Left: female, Right: male.

We tested the goodness of fit with the Pearson’s chi-squared test, which gives the following results. With the high p -value, we accept the hypothesis that the marginal distribution of creatinine is lognormal. In other words, $m(x_k \mid \mu_i)$ is a normal distribution.

	Female	Male
χ^2 statistics	0.0673	0.065
p -value	1	1

Table 6.2: Results of Pearson’s chi-squared test for goodness of fit

6.3.3 Stages of Acute Kidney Injury

Based on the state abstraction, creatinine is assumed to have lognormal distribution with different parameters at different states. We train the model on the patients’ data in order to learn the population parameters of the distributions. The results are then compared to the RIFLE and AKIN scheme. The creatinine criteria for the three stages of renal dysfunction in RIFLE and AKIN are listed in Table 6.3. For comparison with our model,

Table 6.4 shows the population mean of creatinine for each state computed from the lognormal distributions.

The typical values of creatinine for each state are represented as ranges $(m/s, m \times s)$ in

	RIFLE	AKIN
Risk	$\text{Cr} > 1.5 \times \text{baseline}$	$\text{Cr} > 1.5 \times \text{baseline}$, or $\text{Cr} \geq \text{baseline} + 0.3 \text{ mg/dl}$
Injury	$\text{Cr} > 2 \times \text{baseline}$	$\text{Cr} > 2 \times \text{baseline}$
Failure	$\text{Cr} > 3 \times \text{baseline}$, or $\text{Cr} \geq 4 \text{ mg/dl}$ with acute rise of $\geq 0.5 \text{ mg/dl}$	$\text{Cr} > 3 \times \text{baseline}$, or $\text{Cr} \geq 4 \text{ mg/dl}$ with acute rise of $\geq 0.5 \text{ mg/dl}$

Table 6.3: The creatinine criteria in RIFLE and AKIN classification system. For AKIN, the increase in creatinine must occur within 48 hours. Cr, serum creatinine.

	Female	Male
Normal	0.594	0.740
Risk	1.073	1.217
Injury	1.858	2.023
Failure	4.595	4.496

Table 6.4: Population mean of creatinine for each state

Table 6.5, where m and s are the median and geometric standard deviation of the distribution respectively. The range is centered at the median and includes 68.27% of all values. This is a better representation for the multiplicative nature of the lognormal distribution, compared to the (mean \pm standard deviation) representation.

	Female	Male
Normal	(0.405, 0.783)	(0.531, 0.948)
Risk	(0.874, 1.272)	(1.020, 1.413)
Injury	(1.488, 2.228)	(1.665, 2.381)
Failure	(3.137, 6.045)	(3.062, 5.922)

Table 6.5: Typical values of creatinine for each state

Recall that creatinine of normal patients always drifts towards the mean at around 0.6 mg/dl. That is close to the mean of the Normal stage, as expected. The commonly accepted normal range of creatinine is 0.5-1.1 mg/dl for female and 0.6-1.2 mg/dl for male. Lower bounds of the normal ranges are slightly higher than the mean of Normal stage, whereas the upper bounds are just around the mean of Risk stage. This is a reasonable probabilistic interpretation of normal ranges as creatinine value higher than the mean of Risk is likely to be attributed to AKI, even for patients with high baseline.

The 0.3 mg/dl increment in creatinine in the Risk stage is not much larger than the 0.05 mg/dl precision of creatinine measurement, but from the tables, the increase in severity is significant. Observe that the range of typical values that accounts for 68.27% of all values only spans about 0.4 mg/dl. If we consider a female with creatinine of 0.7 mg/dl, an increment of 0.3 mg/dl increases her percentile in the population from 0.75 to 0.96. The patient is still more likely to be in the Normal stage since the increment is less than 50%, but the likelihood of Risk stage has increased substantially.

Ratio of the mean of Risk to the mean of Normal is around 1.7. If we assume that baseline creatinine is the mean of the distribution at the Normal stage, then the ratio for RIFLE is 1.5, which is lower than what we have. Since the Risk stage of RIFLE is known to have high sensitivity, this discrepancy can be attributed to that tradeoff.

There is a difference of about 0.5 mg/dl between the mean of Failure and the value 4 mg/dl used in RIFLE. However, the difference is not large for that level of creatinine, since the noise is multiplicative.

6.3.4 Classification Results

Performance of the model is evaluated by comparing the prediction of the model to the true states of patients extracted from the database for each hospital admission. To get the true states, we look for evidence of acute kidney injury in the patients from the ICD9 codes and clinical notes. For each admission, the observation sequence of creatinine during the admission is given as an input to the trained model, which then return the most likely state sequence. Prediction for that admission is the most severe state in the predicted state sequence. Performance of the model under several metrics is summarized in Table 6.6.

	Female	Male
Sensitivity	0.8271	0.8565
Specificity	0.8702	0.8720
Area under ROC	0.8259	0.8497

Table 6.6: Sensitivity, specificity and the area under ROC of the model.

Our model for state dynamics have demonstrated reasonably well performance on the prediction of acute kidney injury. The statistical properties of the models are also consistent with the universal definitions of AKI. Given the creatinine values, the most likely states of the patients can be efficiently computed using the algorithm discussed previously. In the case when a computer is unavailable, a simple diagnostic test can also be performed based on the mean and typical values provided in the tables. For example, consider a muscular male patient at age of 22 years old and has creatinine value of 1.2 mg/dl. His creatinine is closer to the mean of Risk stage and is outside the typical values of the Normal stage. However, his baseline creatinine should be at high percentile of the population distribution because creatinine generation rate increases with muscle mass. Since his creatinine is not at the higher end of the typical values of Risk state, he is more likely to be in Normal state than the Risk state.

Chapter 7

Conclusion

In this thesis, we have developed probabilistic models for modeling variable kinetics and the temporal dynamics of states. This chapter starts by providing a summary of the thesis, followed by some potential directions for future research.

7.1 Summary

This thesis presents a probabilistic model of the dynamics of kidney function, demonstrates how to estimate parameters of that model from measured creatinine values, and applies the model to diagnosis of acute kidney injury according to a multi-state model. It compares the states automatically discovered by our formalism to the standard RIFLE and AKIN criteria.

Chapter 1 introduced the concepts of kidney injury and the contemporary approach of using data-driven statistical models for clinical analysis and prediction problems, and provided an outline of the rest of the thesis.

In chapter 2, we provided some of the relevant background for this thesis. The chapter begins with a brief summary of the evolution of the diagnostic criteria of acute kidney injury. Then, we discussed the role of creatinine baseline estimation in diagnosis and some existing methods for baseline estimation. We also analyzed some studies on creatinine kinetics in the context of acute kidney injury.

Chapter 3 gave an overview of the MIMIC II database and the various types of data that are available for modeling. Then, some problems with the dataset that may hinder machine learning performance were described.

In chapter 4, we developed the stochastic kinetic model for first-order clearance kinetics, which can be used to model the kinetics of creatinine and other waste substances. Then, we discussed some statistical properties of the model. The chapter concluded with the concept of state abstraction for modeling clinical variables with first-order clearance kinetics.

Chapter 5 described the generative model for the temporal dynamics of states and the relevant clinical variables. To account for the normal range of the variable and the baseline that varies between individuals, we modeled variable distribution by the compound sampling model to distinguish between individual and population distributions. The overall temporal structure was modeled as a Hidden Markov Model with an observation distribution that follows the compound sampling model. Details of approximations to the model to make it tractable and the algorithm for parameter estimation were also included in the chapter. Finally, we discussed the application of the model to state prediction and baseline estimation.

Chapter 6 started with the verification of the stochastic kinetic model with creatinine data. We applied the generative model for state dynamics to the diagnosis of acute kidney injury based on the values of serum creatinine. The state distributions of the model were compared to the RIFLE and AKIN criteria.

7.2 Future Work

This section discusses potential research directions for further investigations, which can be divided into

- applications of the models in other disease domain
- improving the models
- ideas for other related models

The models developed in this thesis can be applied to various clinical problems. The stochastic kinetic model that assumes first-order clearance is applicable to other substances, such as BUN. The model can be easily modified to any linear kinetics, where the general solution is available in Appendix A. The model for state dynamics can be used to model other clinical conditions, as long as the condition is manifested in clinical variables for which the state abstraction hold. For example, observations of BUN can be used to detect conditions like heart failure and dehydration using the model.

A possible improvement that can be made to the model is to relax the approximation assumptions. Recall that we approximated the joint marginal distribution of observations by assuming that the observations are independent. If we are able overcome the tractability issue with weaker assumptions, we will have a more accurate model for the structure of the problem.

Due to the influence of the RIFLE criteria, we modeled acute kidney injury with a discrete number of states, representing different severity levels. An advantage of having discrete states is that it captures clustering effect of different levels of severity in the population. An alternative direction to explore is to use a continuous state representation, in which case state inference can be done by employing techniques such as the Kalman Filter.

In fact, we can go one step further by combining state inference and stochastic kinetic modeling. The strategy would be to infer the creatinine clearance rate from the stochastic process directly. Under this scheme, the clearance rate is no longer a constant, but a continuous function of time. However, this approach is also more susceptible to noise and has many subtle issues to consider. The stochastic differential equation is also unlikely to have an analytical solution.

This concludes our examination of the use of stochastic dynamic modeling to help analyze acute kidney injury. We were able to demonstrate that such a model yields strong diagnostic performance based only on variation in a patient’s serum creatinine, and that the states identified by the model correspond reasonably well to those chosen by human experts to

classify stages of acute kidney injury. We have argued that the methods developed are also more broadly applicable to the dynamics of other clinical measures that exhibit first-order clearance kinetics, and suggested several possible extensions of this work to make better approximations and to use continuous rather than discrete states.

Appendix A

Linear Stochastic Differential Equation

We derive the solution to the general linear stochastic differential equation (SDE). After that, we simplify the solution for the case with constant coefficients that are independent time.

A.1 The General Solution

Consider the general linear stochastic differential equation

$$dX_t = (c(t) + d(t)X_t) dt + (e(t) + f(t)X_t) dW_t \quad (\text{A.1})$$

where c, d, e, f are deterministic function of time, and W_t is a Brownian motion. We first try for a solution of the form $X_t = P_t Q_t$ with

$$\begin{aligned} dP_t &= d(t)P_t dt + f(t)P_t dW_t, & P_0 &= 1 \\ dQ_t &= a(t) dt + b(t) dW_t, & Q_0 &= X_0 \end{aligned}$$

and function a, b to be chosen. By Ito's formula, we have

$$\begin{aligned} dX_t &= P_t dQ_t + Q_t dP_t + fbP_t dt \\ &= dX_t dt + fX_t dW_t + (a + fb)P_t dt + bP_t dW_t. \end{aligned}$$

We choose a, b such that

$$(a + fb)P_t dt + bP_t dW_t = c dt + e dW,$$

which gives

$$a(t) = (c(t) - f(t)e(t))P_t^{-1} \quad (\text{A.2})$$

$$b(t) = e(t)P_t^{-1} \quad (\text{A.3})$$

Notice that

$$P_t = \exp \left(\int_0^t d(s) - \frac{1}{2}f(s)^2 ds + \int_0^t f(s) dW_s \right) \quad (\text{A.4})$$

and $P_t > 0$ almost surely. Hence,

$$Q_t = X_0 + \int_0^t (c(s) - f(s)e(s))P_s^{-1} ds + \int_0^t e(s)P_s^{-1} dW_s \quad (\text{A.5})$$

Let

$$B_t = \int_0^t d(s) - \frac{1}{2}f(s)^2 ds + \int_0^t f(s) dW_s. \quad (\text{A.6})$$

We arrive at the solution to the linear stochastic differential equation:

$$X_t = e^{B_t} \left(X_0 + \int_0^t (c(s) - e(s)f(s))e^{-B_s} ds + \int_0^t e(s)e^{-B_s} dW_s \right) \quad (\text{A.7})$$

A.2 Solution to SDE with Constant Coefficients

This section simplifies the general solution to the case with constant coefficients that are time independent. In particular, we let $c(t) = g$, $d(t) = -k$, $e(t) = \epsilon$, and $f(t) = \sigma$. As a result,

$$B_t = -(k + \frac{1}{2}\sigma^2)t + \sigma W_t \quad (\text{A.8})$$

and the solution becomes

$$X_t = e^{B_t} \left(X_0 + (g - \epsilon\sigma) \int_0^t e^{-B_s} ds + \epsilon \int_0^t e^{-B_s} dW_s \right) \quad (\text{A.9})$$

The last term involves stochastic integral of geometric Brownian motion e^{-B_s} , which can be simplified further using Ito's formula. Let $f(t, W_t) = e^{-B_t}$. By Ito's formula,

$$\begin{aligned} df(t, W_t) &= \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial w^2} \right) dt + \frac{\partial f}{\partial w} dW_t \\ e^{-B_t} - 1 &= e^{-B_t} ((k + \sigma^2) dt - \sigma dW_t) \end{aligned}$$

Rearranging the terms gives us

$$\int_0^t e^{-B_s} dW_s = \frac{1 - e^{-B_t}}{\sigma} + \frac{k + \sigma^2}{\sigma} \int_0^t e^{-B_s} ds \quad (\text{A.10})$$

Substituting A.10 into A.9 gives

$$\begin{aligned} X_t &= e^{B_t} \left(X_0 + \frac{\epsilon}{\sigma} (1 - e^{-B_t}) + \frac{g\sigma + k\epsilon}{\sigma} \int_0^t e^{-B_s} ds \right) \\ &= -\nu + (X_0 + \nu)e^{B_t} + (g + k\nu)e^{B_t} \int_0^t e^{-B_s} ds \end{aligned} \quad (\text{A.11})$$

where $\nu = \frac{\epsilon}{\sigma}$.

Bibliography

- [1] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, Vol. 37, pp. 1554-1563, 1966.
- [2] B. J. C. Baxter and R. Brummelhuis. Functionals of exponential Brownian motion and divided differences. *Journal of Computational and Applied Mathematics*, Vol. 236, Issue 4, pp. 424-433, September 2011.
- [3] M. J. Beal. Variational algorithms for approximate Bayesian inference. PhD thesis, University College London, 2003.
- [4] R. Bellomo, J. A. Kellum and C. Ronco. Defining and classifying acute renal failure: from advocacy to consensus and validation of the RIFLE criteria. *Intensive Care Med*, 33, pp. 409-413, 2007.
- [5] R. Bellomo, C. Ronco, J. A. Kellum, R. L. Mehta, P. Palevsky and the ADQI workgroup. Acute renal failure - definition outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit. Care*, Vol. 8, R204-R210, 2004.
- [6] D. Campbell, C. Fritsche and J. Brandes. A review of urea and creatinine kinetics in predicting CAPD outcome. *Adv. Perit. Dial.*, 8, pp. 79-83, 1992.
- [7] P. Carmona, F. Petit, and M. Yor. On the distribution and asymptotic results for exponential functionals of Lévy processes. In *Exponential Functionals and Principal Values Related to Brownian Motion*, pp. 73-121, 1997.

- [8] G. M. Chertow, E. Burdick, M. Honour, J. V. Bonventre and D. W. Bates. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J. Am. Soc. Nephrol.*, 16(11), pp. 3365-2270, November 2005.
- [9] W. R. Clark, B. A. Mueller, M. A. Kraus and W. L. Macias. Quantification of creatinine kinetic parameters in patients with acute renal failure. *Kidney International*, Vol. 54, pp. 554-560, 1998.
- [10] G. D. Clifford, D. J. Scott and M. Villarroel. User guide and documentation for the MIMIC II database. <http://mimic.mit.edu/documentation.html>, 2011.
- [11] D. W. Cockcroft DW and M. H. Gault. Prediction of creatinine clearance from serum creatinine. *Nephron.*, 16(1), pp. 31-41, 1976.
- [12] D. N. Cruz, Z. Ricci and C. Ronco. Clinical review: RIFLE and AKIN - time for reappraisal. *Crit. Care*, 13:211, 2009.
- [13] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Society (B)*, 39, pp. 1-38, 1977.
- [14] J. L. Doob. The Brownian movement and the stochastic equations. *Annals of Math.*, 43, pp. 351-369, 1942.
- [15] Daniel Dufresne. The log-normal approximation in financial and other computations. *Adv. Appl. Prob.*, 36, pp. 747-773, 2004.
- [16] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association*, Vol. 68, No. 341, March 1973.
- [17] U. Erdbruegger, M. D. Okusa, Etiology and diagnosis of prerenal disease and acute tubular necrosis in acute kidney injury, *UpToDate*, 2013.
- [18] T. Hastie, R. Tibshirani and J. Friedman. The elements of statistical learning: data mining, inference and prediction. *Springer*, 2001.

- [19] S. N. Heyman, S. Rosen, D. Darmon, M. Goldfarb, H. Bitz, A. Shina and M. Brezis. Endotoxin-induced renal failure: II. A role for tubular hypoxic damage. *Exp. Nephrol.*, 8, pp. 275-282, 2000.
- [20] Sérgio Gaião and Dinna N. Cruz. Baseline creatinine to define acute kidney injury: is there any consensus?. *Nephrol. Dial. Transplant.*, 2010.
- [21] Caleb W. Hug. Predicting the risk and trajectory of intensive care patients using survival models. Master's thesis, Massachusetts Institute of Technology, 2006.
- [22] Caleb W. Hug. Detecting hazardous intensive care patient episodes using realtime mortality models. PhD thesis, Massachusetts Institute of Technology, 2009.
- [23] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, Vol. 6, No. 5., pp. 429-449, 2002.
- [24] Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding. *AMIA Annu. Symp. Proc.*, pp. 1276-1283, 2012.
- [25] Kanak Kshetri. Modeling patient states in intensive care patients. Masters thesis, Massachusetts Institute of Technology, 2011.
- [26] A. Lassnigg, D. Schmidlin, M. Mouhieddine, L. M. Bachmann, W. Druml, P. Bauer and M. Hiesmayr. Minimal changes of serum creatinine predict prognosis in patients after cardiothoracic surgery: a prospective cohort study. *J. Am. Soc. Nephrol.*, 15, pp. 1597-1605, 2004.
- [27] Eckhard Limpert, Werner A. Stahel and Markus Abbt. Log-normal distributions across the sciences: keys and clues. *BioScience*, Vol. 51, No. 5, pp. 341-352, 2001.
- [28] David J. C. MacKay. Ensemble Learning for Hidden Markov Models. <http://www.inference.phy.cam.ac.uk/mackay/abstracts/ensemblePaper.html>, 1997.

- [29] H. Matsumoto and M. Yor. Exponential functionals of Brownian motion, I: Probability laws at fixed time. *Probability Surveys*, Vol. 2, pp. 312-347, 2005.
- [30] R. L. Mehta and G. M. Chertow. Acute renal failure definitions and classification: time for change?. *Journal of the American Society of Nephrology*, Vol. 14, No. 8, 2003.
- [31] R. L. Mehta, J. A. Kellum, S. V. Shah, B. A. Molitoris, C. Ronco, D. G. Warnock, A. Levin and the Acute Kidney Injury Network. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. *Crit. Care*, Vol. 11, No. 2, R31, 2007.
- [32] S. M. Moran and B. D. Myers. Course of acute renal failure studied by a model of creatinine kinetics. *Kidney International*, Vol. 27, pp. 928-937, 1985.
- [33] Carl N. Morris. Natural exponential families with quadratic variance functions. *Ann. Statist.*, 10(1), pp. 6580, 1982.
- [34] Carl N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, Vol. 78, No. 381, March 1983.
- [35] Carl N. Morris. Determining the accuracy of Bayesian empirical Bayes estimates in the familiar exponential families. *Statistical Decision Theory and Related Topics IV*, Vol. 1, Proceedings of the 4th Purdue Symposium on Statistical Decision Theory and Related Topics (eds. S. S. Gupta and J. O. Berger), Springer, 1988.
- [36] Kevin P. Murphy. Machine learning: a probabilistic perspective. *The MIT Press*, 2012.
- [37] Saralees Nadarajah. A Review of Results on Sums of Random Variables. *Acta Appl Math*, 103, pp. 131-140, 2008.
- [38] J. W. Pickering and Z. H. Endre. Back-calculating baseline creatinine with MDRD misclassifies acute kidney injury in the Intensive Care Unit. *Clin. J. Am. Soc. Nephrol*, 5, pp. 1165-1173, 2010.

- [39] M. L. Praught and M. G. Shlipak. Are small changes in serum creatinine am important risk factor?. *Curr. Opin. Nephrol. Hypertens.*, 33, pp. 2194-2201, 2005.
- [40] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.
- [41] Z. Ricci, D. Cruz and C. Ronco. The RIFLE criteria and mortality in acute kidney injury: A systematic review. *Kidney International*, 73, pp. 538-546, 2008.
- [42] S. M. Ross. Introduction to Probability Models. *Academic Press*, 9th ed., 2007.
- [43] M. Saeed, M. Villarroel, A. Reisner, T. Heldt, T. H. Kyaw, B. Moody and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39, pp. 952-960, 2011.
- [44] T. Shinzato, S. Nakai, M. Miwa, N. Iwayama, I. Takai, Y. Matsumoto, H. Morita and K. Maeda. New method to calculate creatinine generation rate using pre- and postdialysis creatinine concentrations. *Artificial Organs*, 21(8), pp. 864-872, 1997.
- [45] E. D. Siew, T. A. Ikizler, M. E. Matheny, et al. Estimating baseline kidney function in hospitalized patients with impaired kidney function. *J. Am. Soc. Nephrol.*, 7, pp. 712-719, 2012.
- [46] S. S. Waikar and J. V. Bonventre. Creatinine kinetics and the definition of acute kidney injury. *J. Am. Soc. Nephrol.*, 20, pp. 672-679, 2009.
- [47] Ron Wald. Predicting baseline creatinine in hospitalized patients. *Cli. J. Am. Soc. Nephrol.*, 7, pp. 697-699, 2012.
- [48] L. Wan, R. Bellomo, D. Di Giantomasso and C. Ronco. The pathogenesis of septic acute renal failure. *Curr. Opin. Crit. Care*, 9, pp. 496-502, 2003.
- [49] S. D. Weisbord, H. Chen, R. A. Stone, K. E. Kip, M. J. Fine, M. I. Saul and P. M. Palevsky. Associations of increases in serum creatinine with mortality and length of

- hospital stay after coronary angiography. *J. Am. Soc. Nephrol.*, 17(10), pp. 2871-2877, October 2006.
- [50] K. A. Wichterman, A. E. Baue and I. H. Chaudry. Sepsis and septic shock: a review of laboratory models and a proposal. *J. Surg. Res.*, 29, pp. 189-201, 1980.
- [51] J. Závada, E. Hoste, R. Cartin-Ceba, P. Calzavacca, O. Gajic, G. Clermont, R. Bellomo, J. A. Kellum, and the AKI6 investigators. A comparison of three methods to estimate baseline creatinine for RIFLE classification. *Nephrol. Dial. Transplant*, 25, pp. 3911-3918, January 2010.
- [52] Yu Zhang, Peng Liu, Jen-Tzung Chien and Frank Soong. An evidence framework for Bayesian learning of continuous-density hidden Markov models. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 3857-3860, 2009.