Quantifying the Strength of Evidence in Forensic Fingerprints

Peter George MacEwan Forbes

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2014

Quantifying the Strength of Evidence in Forensic Fingerprints
Submitted for the degree of Doctor of Philosophy in Trinity 2014
Peter George MacEwan Forbes, Somerville College

Statement of Originality

Much of this dissertation is the product of collaborations with others. In particular Steffen Lauritzen has contributed invaluably and guided all aspects of the work.

Large parts of chapters 1, 2, 4 and 5 are the product of a paper co-authored by myself, Steffen Lauritzen and Jesper Møller (Forbes et al., 2014). Specifically Professor Møller contributed the formal stepwise Poisson point process formulation in §2.1 to §2.3.

Chapter 10 has been published in *Biometrika* (Forbes, 2012). While I am the only author listed on this paper, Professor Lauritzen provided many suggestions that improved the paper's clarity and cohesion.

Chapter 11 originates from a paper co-authored by myself and Steffen Lauritzen (Forbes and Lauritzen, 2014) and published in the journal *Linear Algebra and its Applications*.

Appendix A is the result of a paper written with Kanti Mardia (Forbes and Mardia, 2014) and published in the *Journal of Statistical Computation and Simulation*. Professor Mardia contributed suggestions to improve the paper's exposition and the literature review in §A.1.

Quantifying the Strength of Evidence in Forensic Fingerprints
Submitted for the degree of Doctor of Philosophy in Trinity 2014
Peter George MacEwan Forbes, Somerville College

Abstract

Part I presents a model for fingerprint matching using Bayesian alignment on unlabelled point sets. An efficient Monte Carlo algorithm is developed to calculate the marginal likelihood ratio between the hypothesis that an observed fingerprint and fingermark pair originate from the same finger and the hypothesis that they originate from different fingers. The model achieves good performance on the NIST-FBI fingerprint database of 258 matched fingerprint pairs, though the computed likelihood ratios are implausibly extreme due to oversimplification in our model.

Part II moves to a more theoretical study of proper scoring rules. The chapters in this section are designed to be independent of each other. Chapter 9 uses proper scoring rules to calibrate the implausible likelihood ratios computed in Part I. Chapter 10 defines the class of compatible weighted proper scoring rules. Chapter 11 derives new results for the score matching estimator, which can quickly generate point estimates for a parametric model even when the normalization constant of the distribution is intractable. It is used to find an initial value for the iterative maximization procedure in §3.3.

Appendix A describes a novel algorithm to efficiently sample from the posterior of a von Mises distribution. It is used within the fingerprint model sampling procedure described in §5.6. Appendix B includes various technical results which would otherwise disrupt the flow of the main dissertation.

Contents

Glossary

This glossary contains a list of the nomenclature used throughout Part I. For brevity, symbols which are only used in one location (e.g., the intermediate steps of the point processes in chapter 2 and the test statistics in chapter 7) are not listed. Similarly, since each chapter in Part II is self-contained with its own definitions, the symbols from Part II are not listed.

A: the set of minutiae observed in the fingerprint, $A \subset \mathbb{M}$. p.6
α_ι: the shape parameter for the distribution of the variable ι, $\iota \in \{\delta, \sigma\omega, \kappa\}$. p.32
β_ι: the rate parameter for the distribution of the variable ι, $\iota \in \{\delta, \sigma\omega, \kappa\}$. p.32
B: the set of minutiae observed in the fingermark, $B \subset \mathbb{M}$. p.6
$\mathbb{C}$: the complex plane, used to represent the position of a minutia. p.6
$\tilde{c}(M)$: the normalizing constant for a MPPP density with respect to ζ, which depends only on the observed marked point set M. p.12, 20
χ: the probability that a latent minutia is a bifurcation. p.17
c_ω: the lower bound on $\tilde{\omega}$, $1 < c_\omega < \tilde{\omega}$. p.21
δ_A: the probability that a latent minutia is observed in A. Similarly δ_B. p.18
ϵ: the probability that a observed minutia m has an unclassified type, $t_m = 0$. The value of ϵ does not affect the likelihood ratio. p.18
$\hat{\mathbb{E}}_S$: the sample expectation operator for the sample S. p.35
${}_1F_1$: the confluent hypergeometric function. p.34
fingermark: a residual finger ridge pattern left accidentally, typically low quality. Mathematically represented by a finite set of points $B \subset \mathbb{M}$. p.2
fingerprint: a residual finger ridge pattern left intentionally, typically high quality. Mathematically represented by a finite set of points $A \subset \mathbb{M}$. p.2
$g(s,t)$: the density of the latent minutia marks. p.10
$\Gamma(a,z)$: the upper incomplete Gamma function for real variables. p.33
$\Gamma(z)$: the Gamma function for real variables. p.32
H: the binary variable $H \in \{H_d, H_p\}$, used for reversible jump MCMC. p.41
H_d: the defence hypothesis, that the fingermark and fingerprint originate from different fingers. p.6, 9
H_p: the prosecution hypothesis, that the fingermark and fingerprint originate from the same finger. p.6, 9
$I_0(\kappa)$: the modified Bessel function of the first kind and order zero. p.8
$\mathbb{1}(x)$: the indicator function for the event x. p.12

κ: the precision parameter for the orientation distortion. p.18
k_τ: concentration parameter for the distributions of the variables τ_A, τ_B. p.32
Λ: the marginal likelihood ratio between H_p and H_d. p.31
$\lambda_\kappa(\alpha_\kappa, \beta_\kappa)$: the normalization constant for the distribution of κ. p.33
M: an unobserved set of latent minutiae, $M \subset \mathbb{M}$. p.9
$\mathbb{M}$: the product space $\mathbb{M} = \mathbb{C} \times \mathbb{S}^1 \times \mathbb{T}$, used to represent a minutia. p.6
minutia: a feature of a finger ridge pattern, where a ridge ends or bifurcates. Each minutia is mathematically represented by a point in $\mathbb{M}$. p.5
MPPP(ρ, g): a marked Poisson point process with point intensity ρ and mark density g. p.10
$\mu_\mathbb{C}$: the Lebesgue measure on $\mathbb{C}$. p.7
$\mu_\mathbb{R}$: the Lebesgue measure on $\mathbb{R}$. p.33
$\mu_{\mathbb{S}^1}$: the uniform probability measure on $\mathbb{S}^1$. p.8
$\mu_\mathbb{T}$: the counting measure on $\mathbb{T}$. p.10
μ_θ: the product measure for the variables in θ. p.33
$\mu_{\Xi(A,B)}$: the counting measure on $\Xi(A,B)$. p.16
n_A: the number of minutiae in A. Similarly n_B. p.16
$n_A^{(t)}$: the number of minutiae of type t in A. Similarly $n_B^{(t)}$. p.19
n_ξ: the number of minutiae in the matching ξ, $n_\xi = |\xi|$. p.16
$n_\xi^{(t)}$: the number of minutia pairs $(a, b) \in \xi$ that satisfy $t_a = t_b = t$. p.21
ω: the standard deviation of the distortion in an observed minutia's location. p.18
$\tilde{\omega}$: a parameter for the inverse-variance of the minutia location distortions. p.19
p_0: the prior odds of H_d, used as a tuning parameter for the reversible jump MCMC algorithm. p.41
ϕ: a fixed but arbitrary point in $\mathbb{M} \setminus (A \cup B)$ which is used to denote an unobserved minutia. p.25, 39
$\varphi(r; r_0, \sigma^2)$: the complex normal density with mean r_0 and variance σ^2. p.7
$\varphi(r)$: see $\varphi(r; r_0, \sigma^2)$ with $r_0 = 0$ and $\sigma^2 = 1$. p.7
$\varphi_2(r; r_0', \Sigma)$: bivariate complex normal density with mean r_0' and covariance Σ. p.7
$\varphi_2(r_1, r_2; r_0', \Sigma)$: see $\varphi_2(r; r_0', \Sigma)$ with $r = (r_1, r_2)^\top$, where $r_1, r_2 \in \mathbb{C}$. p.8
$\varphi_2(r)$: see $\varphi_2(r; r_0', \Sigma)$ with $r_0' = 0$ and Σ equal to the identity matrix. p.8
Π_A: the projector $(a, b) \mapsto a$, where $a, b \in \mathbb{M}$. Similarly, $\Pi_B(a, b) = b$. p.15
$\Pi_{A,b}(\xi)$: the minutia $a \in A$ which is matched to $b \in \mathbb{M}$. Similarly $\Pi_{B,a}(\xi)$. p.39
profile likelihood ratio: the likelihood ratio where the matching ξ is marginalized by maximization rather than by summation. p.23
ψ: the relative rotation between between A and B. p.19
$q(\theta, \xi)$: a normalized density which approximates of the posterior $p(\theta, \xi \,|\, A, B, H_p)$. p.36, 43
$R_{f(A,B,\xi)}$: various sums of the observed minutia locations, defined in (3.5). p.24
R_n: $n = 1$–5. various functions of data and parameters; see (3.14),(3.18). p.27, 28
Re(z): the real part of a complex number $z \in \mathbb{C}$. p.8
ρ_0: the expected number of latent minutiae in M, $\rho_0 = \int_\mathbb{C} \rho(r)\, \mathrm{d}\mu_\mathbb{C}(r)$. p.9
$\rho(r)$: the intensity function of the MPPP for latent minutiae. p.9

r_m: the projection of $m \in \mathbb{M} = \mathbb{C} \times \mathbb{S}^1 \times \mathbb{T}$ onto the location space $\mathbb{C}$. p.6
rvM(ν, κ): the root von Mises distribution on $\mathbb{S}^1$. p.8
$S_{f(A,B,\xi)}$: various functions of the observed data, defined in (3.5). p.24
S: a sample from the posterior $p(\theta, \xi | A, B, H_p)$. p.35
$\mathbb{S}^1$: the complex unit circle, $\{s \in \mathbb{C} : |s| = 1\}$, used to represent the orientation of a minutia. p.6
$\mathbb{T}$: the discrete space $\{-1, 0, 1\}$, used to represent the type of a minutia (ending$= -1$, bifurcation$= 1$, unclassified$= 0$). p.6
σ_A: the standard deviation of the scaled minutia locations. Similarly σ_B. p.19
Σ_{AB}: the covariance matrix of the paired minutia locations. p.20
σ_P: the reparameterization $\sigma_P = 1/(\sigma_A \sigma_B)$. p.37
σ_Q: the reparameterization $\sigma_Q = \sigma_B / \sigma_A$. p.37
s_m: the projection of $m \in \mathbb{M} = \mathbb{C} \times \mathbb{S}^1 \times \mathbb{T}$ onto the orientation space $\mathbb{S}^1$. p.6
Θ: the set of all parameters in our model. p.11, 33
θ: the set of parameters which vary from finger to finger, and must be marginalized by either integration or maximization. p.21
t_m: the projection of $m \in \mathbb{M} = \mathbb{C} \times \mathbb{S}^1 \times \mathbb{T}$ onto the type space $\mathbb{T}$. p.6
vM(ν, κ): the von Mises distribution on $\mathbb{S}^1$. p.8
X_i: a generic MPPP over $\mathbb{M}$ with intensity φ. p.70
ξ: the matching, i.e. the unobserved set of edges $\langle a, b \rangle$ which specifies which $a \in A$ and $b \in B$ correspond to the same latent minutia in M. Under H_d we have $\xi = \emptyset$. p.15
$\check{\xi}_i$: the manually-specified "true" matching for the observed fingerprint/fingermark pair (A_i, B_i). p.54
$\xi_{\leq \beta}$: the subset of ξ with $b \leq \beta$. Similarly $\xi_\beta, \xi_{<\beta}, \xi_{>\beta}$. p.39
$\Xi(A, B)$: the space of possible values for the matching ξ. p.16
ζ: the density of a MPPP$(\varphi, 1/3)$ on $\mathbb{M}$. p.12
ζ_2: the density of a MPPP$(\varphi_2, 1/9)$ on $\mathbb{M} \times \mathbb{M}$. p.14
$\tilde{\zeta}$: the base measure of $p(A, B, \xi \,|\, \Theta, H_p)$. p.16

Part I

Fingerprint Modelling

Chapter One

Preliminaries

Societies have used the ridge patterns located on human fingers as a method of identification since at least 300 BC (Herschel, 1916). In recent times these ridge patterns have been primarily used for either user authentication (for instance, smartphone lock screens and airport immigration control) or forensics (providing evidence that a suspect touched an item of interest). The forensic application exploits the fact that fingers transfer marks to almost all surfaces they touch. When these transfers are accidental we call them *fingermarks*: these are usually of low quality and permit only a partial reconstruction of the source finger's ridge pattern. Conversely, intentional transfers, such as those taken at a police station, are called *fingerprints* and usually permit near-complete construction of the source finger's ridge pattern. We use the term *fingerprint evidence* to refer to both fingerprints and fingermarks.

§1.1 History

The first extensive research into fingerprint evidence for identification was conducted by Sir Francis Galton. In his book *Finger Prints* (Galton, 1892), he posited two key assumptions:

- Persistence: the ridge patterns on an individual's finger are invariant over time
- Individuality: a complete ridge pattern uniquely identifies an individual.

This chapter is an elaborated version of the introduction in Forbes et al. (2014).

To this day these assumptions remain undisputed by mainstream forensic scientists (Peterson et al., 2009). Furthermore, Galton used a crude model to estimate the probability of two fingers having identical ridge patterns as 2^{-36}, or around 1 in 69 billion (Galton, 1892, p. 110), though the rationale for this number does not stand up to any scrutiny. In addition to his book, Galton was an avid promoter of fingerprint evidence. He published seventeen articles in *Nature*, *Scientific American*, the *Proceedings of the Royal Society*, the London newspaper *The Times*, and the popular science magazine *Nineteenth Century* from 1888 to 1896.

Galton's work came at an opportune time. England had recently stopped exiling criminals to Australia, choosing instead to deter repeat offenders through progressively tougher sentences (Specter, 2002). Criminals would often use aliases, so police departments required a reliable method to identify repeat offenders. Galton directly influenced a 1894 parliamentary report that paved the way for fingerprint evidence to be used in English courts (Troup, 1894). The use of fingerprint evidence quickly spread to other nations.

§1.2 Motivation

Over the years fingerprint evidence developed an extremely strong reputation. For instance, the Federal Bureau of Investigation's 1963 training manual reads "Of all the methods of identification, fingerprinting alone has proved to be both infallible and feasible" (Hoover, 1963, p. iv). However, whether or not a given fingermark is of sufficient quality to uniquely identify an individual is a subjective judgement: "An individual examiner's threshold for sufficiency is predicated on the examiner's education, knowledge, experience, and training" (Peterson et al., 2009). This subjective judgement of sufficiency is difficult to quantify and may lead to inconsistent judgements across different examiners (Cole, 2005).

There have been many attempts to objectively determine when a partial or distorted fingermark should be considered sufficient for unique identification. These

studies have been met with widespread criticism for their untested assumptions:

> "From a statistical viewpoint, the scientific foundation for fingerprint individuality is incredibly weak. Beginning with Galton and extending through Meagher, Budowle, and Ziesig, there have been a dozen or so statistical models proposed. These vary considerably in their complexity, but in general there has been much speculation and little data…None of the models has been subjected to testing, which is of course the basic element of the scientific approach" (Stoney, 2001, p. 395).

The issue of determining whether a fingermark is sufficient for unique identification is exasperated by the strict rules governing its courtroom presentation. A fingerprint expert must issue one of three mutually exclusive categorical conclusions: the fingermark definitively came from the suspect, the fingermark definitively did not come from the suspect, or the fingermark is of insufficient quality to draw a conclusion. To issue a probabilistic statement ("the fingermark most likely came from the suspect") was considered unbecoming conduct and was grounds for the examiner to be removed from the field's professional association until 2010 (IAI, 2010).[†] Thus potentially useful evidence is discarded whenever the examiner considers the quality of a fingermark insufficient to draw a decisive conclusion (Neumann et al., 2012a).

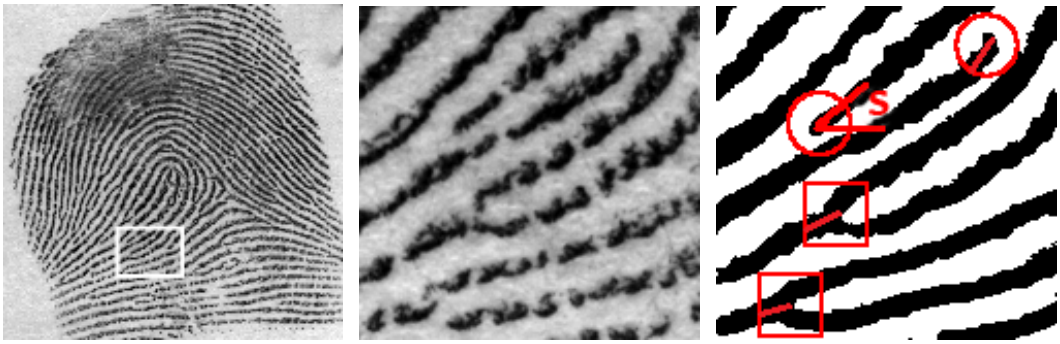
One solution is to abandon the notion of sufficiency and present courtroom fingerprint evidence *probabilistically* rather than categorically. Then fingerprint evidence could be synthesized by the judge or jury, alongside any other evidence, when determining whether or not a suspect is guilty beyond a reasonable doubt. This approach necessitates the development of a model to accurately compute the relevant probability. One model was proposed in a recent paper by members of the now-defunct Forensic Science Service (Neumann et al., 2012a). The fact that this paper was presented to the Royal Statistical Society rather than a forensic journal is significant and represents "a major step on the path to paradigm change in the operational world of fingerprint identification" (Neumann et al., 2012a, p. 394).

[†]Though probabilistic statements are now permitted by the IAI, they are still prohibited in most nation's legal systems. In court, fingerprint evidence must still be presented categorically.

In a response to that paper, Lauritzen et al. (2012) noted the similarity between the proposed model for evidence and the alignment problems often studied in bioinformatics. We follow that idea here by adapting the hierarchical Bayesian model for unlabelled point set matching of Green and Mardia (2006) to the problem of fingerprint/fingermark matching. We develop an efficient Monte Carlo algorithm to estimate the likelihood ratio for the prosecution hypothesis that the observed fingerprint and fingermark originate from the same finger against the defence hypothesis that they originate from different fingers.

§1.3 Likelihood representation of fingerprint evidence

In England, fingerprint evidence has been used for criminal identification since the late nineteenth century (Faulds, 1880). Fingerprint evidence is based on the similarity of two or more pictures (see figure 1.1(a) for an example). It is difficult to represent all the information from these pictures into a mathematically convenient form. Thus most models consider only a subset of the information, namely, the points on the image where a ridge either ends or bifurcates. These points, called *minutiae* (see figure 1.1(c)), contain sufficient information to uniquely identify an individual (Maltoni et al., 2009; Yager and Amin, 2004b).



(a) Exemplar fingerprint (b) Zoomed section (c) Enhanced & labelled

Figure 1.1: A typical exemplar quality fingerprint from Garris and McCabe (2000). The highlighted points in (c) are minutiae: circles are ridge endings and squares are bifurcations.

As in Neumann et al. (2012a), we discuss the situation where we wish to compare

a high-quality fingerprint A, taken under controlled circumstances, with a fingermark B found on a crime scene. A typical fingerprint contains 100–200 minutiae, while a fingermark may contain fewer than ten (Garris and McCabe, 2000).

We consider two hypotheses:

$$\begin{aligned} H_p &: A \text{ and } B \text{ originate from the same finger,} \\ H_d &: A \text{ and } B \text{ originate from different fingers,} \end{aligned} \tag{1.1}$$

where H_p is the *prosecution hypothesis* and H_d is the *defence hypothesis*. Following a tradition that goes at least back to Lindley (1977), we follow standards in modern evaluation of DNA evidence (Balding, 2005) and other types of forensic evidence (Aitken and Taroni, 2004) by quantifying the weight-of-evidence through a *likelihood ratio* between H_p and H_d. This likelihood ratio will be based on probabilistic models for the generation of the fingerprint and fingermark that shall be developed below.

§1.4 Representation of fingerprints

Each minutia m consists of a location, an orientation, and a type classification. We represent the location with a point in the complex plane $\mathbb{C}$. The orientation is represented with a point on the complex unit circle $s \in \mathbb{S}^1$ such that the phase of s is the counter-clockwise angle between the horizontal axis and the ridge which is ending or bifurcating (see figure 1.1(c)). The type is represented by a number in $\mathbb{T} = \{-1, 0, 1\}$, where -1 denotes a ridge ending, 1 denotes a bifurcation, and 0 denotes an unclassified type (see figure 1.1(c)). Thus m lies in the product space $\mathbb{M} = \mathbb{C} \times \mathbb{S}^1 \times \mathbb{T}$. We let r_m, s_m, and t_m denote the projection of m onto the location space, orientation space, and type space respectively.

A fingerprint A (or a fingermark B) is represented by a finite set of points in $\mathbb{M}$.[†] We call this representation a *minutia configuration*. Since A and B are observed in arbitrary and different coordinate systems, the observed minutiae are subjected to arbitrary

[†]We refrain from ordering the elements within A and B in order to avoid extra permutation factors in the subsequent probability distributions.

similarity transformations, which consist of translations, rotations, and scalings. Due to our choice of representation for minutiae, we can simply represent these similarity transformations with a translation parameter $\tau \in \mathbb{C}$ and a rotation/scale parameter $\psi \in \mathbb{C} \setminus \{0\}$ via the mapping

$$m = (r_m, s_m, t_m) \mapsto (\psi r_m + \tau, \psi s_m / |\psi|, t_m) \tag{1.2}$$

from $\mathbb{M}$ to $\mathbb{M}$.

§1.5 Overview of distributions

We will need various probability distributions in the remainder of this dissertation, some of which may not be well known to the reader. We define these distributions here for easy reference. They are also listed in the glossary at the beginning of this dissertation.

The complex normal distribution (Goodman, 1963) will be used to model random variables related to the minutia locations. It is parametrized by its mean $r_0 \in \mathbb{C}$ and its variance $\sigma^2 > 0$, and has density

$$\varphi(r; r_0, \sigma^2) = \exp(-|r - r_0|^2/\sigma^2)/(\pi\sigma^2) \tag{1.3}$$

with respect to $\mu_{\mathbb{C}}$, the Lebesgue measure on $\mathbb{C}$. Here $|r - r_0|$ is the modulus of the complex number $r - r_0$, defined by $|r - r_0|^2 = \overline{(r - r_0)}(r - r_0)$, where the overline represents complex conjugation. The standard case of $r_0 = 0$ and $\sigma^2 = 1$ is denoted by $\varphi(r)$.

The bivariate complex normal distribution will be used for paired minutia locations. It is parametrized by its mean $r_0' \in \mathbb{C}^2$ and a Hermitian positive definite complex 2×2 covariance matrix Σ. It has density

$$\varphi_2(r; r_0', \Sigma) = \exp\{-(\overline{r - r_0'})^\top \Sigma^{-1}(r - r_0')\}/(\pi^2|\Sigma|) \tag{1.4}$$

with respect to $\mu_{\mathbb{C}} \times \mu_{\mathbb{C}}$, where $|\Sigma|$ denotes the determinant and $^\top$ denotes the vector transpose. The standard case of $r_0 = 0$ and Σ equal to the identity matrix is denoted

$\varphi_2(r)$. When we wish to make the two arguments explicit we will write $\varphi_2(r_1, r_2; r_0', \Sigma)$ for $r_1, r_2 \in \mathbb{C}$.

The *von Mises distribution* (Mardia and Jupp, 1999) will be used to model random variables related to the minutia orientations. It has support on the complex unit circle $\mathbb{S}^1$, and it is parametrized by its mean $\nu \in \mathbb{S}^1$ and its precision $\kappa > 0$. Letting $I_0(\kappa)$ denote the modified Bessel function of the first kind and order zero (Olver et al., 2010, §10), and letting $\mathrm{Re}(z) = (z+\overline{z})/2$ denote the real part of z, the density of a von Mises random variable $s \sim \mathrm{vM}(\nu, \kappa)$ can be written as

$$p_{\mathrm{vM}}(s; \nu, \kappa) = \frac{1}{I_0(\kappa)} \exp\{\kappa \,\mathrm{Re}(s\overline{\nu})\} \tag{1.5}$$

with respect to the uniform probability measure $\mu_{\mathbb{S}^1}$ on $\mathbb{S}^1$. Informally, the von Mises distribution can be understood as the result of restricting a univariate complex normal distribution $\varphi(s; \nu, 2/\kappa)$ onto the unit circle $|s| = 1$.

To facilitate the later mathematics we will also need a *root von Mises* distribution, $\mathrm{rvM}(\nu, \kappa)$, which we define by

$$XY \sim \mathrm{vM}(\nu, \kappa) \text{ whenever } X, Y \text{ are independent and } X, Y \sim \mathrm{rvM}(\nu, \kappa). \tag{1.6}$$

This distribution is well defined because the von Mises distribution is infinitely divisible on $\mathbb{S}^1$ (Kent, 1977). The density of the root von Mises distribution is determined by a series expansion; we refrain from giving the details as we shall not need them.

Chapter Two

Fingerprint models

We consider the observed minutia configurations $A, B \subset \mathbb{M}$ as thinned and displaced copies of a latent minutia configuration.[†] We assume that different fingers have independent latent minutiae configurations, whether those fingers come from the same individual or different individuals. Thus we can rephrase our two model hypotheses (1.1) as

H_p : A and B originate from a common latent minutia configuration $M \subset \mathbb{M}$,

H_d : A and B originate from independent latent minutia configurations $M, M' \subset \mathbb{M}$.

In the notation of marked point processes, each minutia $m \in \mathbb{M}$ is a marked point. The projection of m onto the location space $\mathbb{C}$, r_m, is called a *point* and the projection onto $\mathbb{S}^1 \times \mathbb{T}$, (s_m, t_m), is called a *mark*.

We model a latent minutia configuration $M \subset \mathbb{M}$, which consists of a finite number of marked points in $\mathbb{M}$, as a finite *marked Poisson point process* (MPPP) on $\mathbb{C}$ with intensity function $\rho : \mathbb{C} \to [0, \infty)$ (see, e.g., Møller and Waagepetersen (2004)). This implies that the cardinality $|M|$ of M is Poisson distributed with mean $\rho_0 = \int_{\mathbb{C}} \rho(r)\,\mathrm{d}r$, where $\mathrm{d}r = \mathrm{d}\mu_{\mathbb{C}}(r)$ denotes the Lebesgue measure. We assume that ρ_0 is both positive and finite. Conditional on $|M|$, the points are independently and identically distributed (i.i.d.) over $\mathbb{C}$ with density ρ/ρ_0 with respect to $\mu_{\mathbb{C}}$.

This chapter is a slight extension of Forbes et al. (2014).

[†]We use the word *latent* as a synonym for *unobservable*, and thus neither A nor B are latent. This contrasts with a common usage in fingerprints forensics: there a *latent fingerprint* refers to fingermark which is difficult to see with the naked eye, but can be observed via specialized techniques.

The marks are assumed to be i.i.d. and independent of the points. The marks have density g with respect to the product measure $\mu_{\mathbb{S}^1} \times \mu_{\mathbb{T}}$, where $\mu_{\mathbb{T}}$ is the counting measure on $\mathbb{T}$. We write the resulting distribution for M succinctly as $M \sim \mathrm{MPPP}(\rho, g)$.

§2.1 A generic marked point process model

We now describe a framework to model fingerprint and fingermark data. Section 2.3 will use this framework to specify a concrete model by choosing particular parametric families for each of the distributions mentioned in this section.

In our framework, the observed fingerprint A is modelled as a MPPP which is obtained from the latent minutia configuration M through three basic operations as follows:

A1: thinning Only a subset of the latent minutiae are observed, resulting in $M_{A1} = \{m \in M : I_A(m) = 1\}$, where the indicators I_A are Bernoulli variables whose success probabilities $\delta_A(r_m)$ depend only on the location r_m. We then have

$$M_{A1} \sim \mathrm{MPPP}(\rho_{A1}, g_{A1}) \text{ where } \rho_{A1}(r) = \rho(r)\delta_A(r), \quad g_{A1} = g. \tag{2.1}$$

A2: displacement The locations r_m in M_{A1} are subjected to additive errors $e_m \in \mathbb{C}$ with density f_A, the orientations s_m are subjected to multiplicative errors $v_m \in \mathbb{S}^1$ with density h_A, and the types are subjected to classification errors $c_m \in \mathbb{T}$ with distribution d_A. This results in $M_{A2} = \{(r_m + e_m, v_m s_m, c_m t_m) : m \in M_{A1}\}$. Consequently, $M_{A2} \sim \mathrm{MPPP}(\rho_{A2}, g_{A2})$, where

$$\rho_{A2}(r) = f_A * \rho_{A1}(r) = \int_{\mathbb{C}} f_A(e)\rho_{A1}(r-e)\,\mathrm{d}e \tag{2.2}$$

is obtained by usual convolution in $\mathbb{C}$. The mark density is

$$g_{A2}(s,t) = \sum_{u \in \mathbb{T}} d_A(ut) \int_{\mathbb{S}^1} h_A(v) g_{A1}(s\overline{v}, u)\,\mathrm{d}\mu_{\mathbb{S}^1}(v). \tag{2.3}$$

A3: mapping Finally, the marked points are subjected to a similarity transformation parametrized by $\tau_A \in \mathbb{C}$ and $\psi_A \in \mathbb{C} \setminus \{0\}$ to obtain

$$A = \{(\psi_A r_m + \tau_A, \psi_A s_m / |\psi_A|, t_m) : m \in M_{A2}\}. \tag{2.4}$$

Thus $A \sim \text{MPPP}(\rho_{A3}, g_{A3})$ where $\rho_{A3}(r) = \rho_{A2}\{(r-\tau_A)/\psi_A\}/|\psi_A|^2$ and $g_{A3}(s,t) = g_{A2}(s\overline{\psi_A}/|\psi_A|, t)$.

The model for B is specified analogously: B is the MPPP derived from a latent minutia configuration M' by three similar steps B1–B3 obtained by replacing A with B everywhere in A1–A3. That is, $B \sim \text{MPPP}(\rho_{B3}, g_{B3})$ with the intensity function and the mark density defined as above but using a new function δ_B, new indicators $I_B(m)$, new distributions f_B, h_B, d_B, new error terms e'_m, v'_m, c'_m, and new parameters τ_B, ψ_B.

Finally, we make the following independence assumptions. Under H_d we have M and M' are i.i.d., while under H_p, $M = M'$. In both cases they have distribution $\text{MPPP}(\rho, g)$. Conditional on M and M', all the variables $I_A(m), e_m, v_m, c_m$ for $m \in M$, and $I_B(m), e'_m, v'_m, c'_m$ for $m \in M'$ are mutually independent with distributions which do not depend on M and M'.

In §2.3 we specify parametric models for $\rho, g, \delta_A, \delta_B, f_A, f_B, h_A, h_B, d_A$, and d_B. We use Θ as generic notation for the set of all relevant parameters, including τ_A, τ_B, ψ_A, and ψ_B. The intensity functions and mark densities depend on Θ, but for simplicity we suppress this dependence in the notation.

§2.2 Densities under the two hypotheses

We are interested in computing a likelihood ratio between the two hypotheses H_p and H_d. Thus we must find the normalized densities of the MPPPs A and B under both H_p and H_d. In order to rigorously define these densities, we must briefly detour into measure theory. Let $Z \sim \text{MPPP}(\varphi, 1/3)$ be a marked Poisson point process on $\mathbb{M}$ whose intensity is given by a standard complex normal density and whose marks are

uniformly distributed over $\mathbb{S}^1 \times \mathbb{T}$.[†] Then Z takes values on the space of *finite point configurations*, $N_f = \{M \subset \mathbb{M} : |M| < \infty\}$ (Møller and Waagepetersen, 2004, p.82). The probability that Z lies in some set $\mathscr{Z} \subseteq N_f$ is

$$P(Z \in \mathscr{Z}) = \sum_{n=0}^{\infty} \frac{\exp(-1)}{n!} \int_{\mathbb{M}} \cdots \int_{\mathbb{M}} \mathbb{1}(\{m_1, \ldots, m_n\} \in \mathscr{Z}) \left\{ \prod_{i=1}^{n} \frac{\varphi(r_{m_i})}{3} \right\} \mathrm{d}m_1 \ldots \mathrm{d}m_n, \tag{2.5}$$

where $\mathbb{1}$ is the indicator function, and the integral over $\mathbb{M}$ is a shorthand for integrating over the minutia location and orientation and summing over the type, $\int_{\mathbb{M}} f(m)\,\mathrm{d}m = \int_{\mathbb{C}} \int_{\mathbb{S}^1} \sum_{t \in \mathbb{T}} f(r,s,t)\,\mathrm{d}\mu_{\mathbb{S}^1}(s)\,\mathrm{d}r$. We let ζ be the measure defined by $\zeta(\mathscr{Z}) = P(Z \in \mathscr{Z})$; we will use ζ as a dominating measure for the densities of A and B.

Now, using the fact that

$$\int_{\mathbb{C}} \rho_{A3}(r)\,\mathrm{d}r = \int_{\mathbb{C}} \rho_{A2}(r)\,\mathrm{d}r = \int_{\mathbb{C}} \rho_{A1}(r)\,\mathrm{d}r = \int_{\mathbb{C}} \rho(r)\delta_A(r)\,\mathrm{d}r, \tag{2.6}$$

we see that the marginal density of A with respect to ζ is

$$p(A \mid \Theta) = c(A) \exp\left\{ -\int_{\mathbb{C}} \rho(r)\delta_A(r)\,\mathrm{d}r \right\} \prod_{m \in A} \{ g_{A3}(s_m, t_m) \rho_{A3}(r_m) \}, \tag{2.7}$$

where

$$c(A) = 3^{|A|} \exp(1) \prod_{a \in A} \varphi(r_a)^{-1} \tag{2.8}$$

depends only on the data, see for example Møller and Waagepetersen (2004, p.25).

The density of B with respect to ζ is obtained by replacing A by B everywhere in (2.7). Notice the marginal density of A is identical under both H_d and H_p, $p(A \mid \Theta, H_d) = p(A \mid \Theta, H_d)$. This is also true for B.

[†]Note that the uniform distribution on $\mathbb{S}^1$ has density $p(s) = 1$ with respect to $\mu_{\mathbb{S}^1}$. This is because $\mu_{\mathbb{S}^1}$ is the uniform probability measure on $\mathbb{S}^1$, rather than the more commonly seen Lebesgue measure. This choice of measure greatly reduces the number of 2π factors appearing in later computations.

§2.2.1 Density under the defence hypothesis

Under H_d the fingerprint A and fingermark B are independent and thus the density with respect to $\zeta \times \zeta$ is simply the product

$$p(A,B \mid \Theta, H_d) = c(A)c(B) \exp\left\{ -\int_{\mathbb{C}} \rho(r)\delta_A(r)\,\mathrm{d}r - \int_{\mathbb{C}} \rho(r)\delta_B(r)\,\mathrm{d}r \right\} \times \left\{ \prod_{a\in A} \rho_{A3}(r_a) g_{A3}(s_a,t_a) \right\} \left\{ \prod_{b\in B} \rho_{B3}(r_b) g_{B3}(s_b,t_b) \right\}. \quad (2.9)$$

§2.2.2 Density under the prosecution hypothesis

In order to find the desired density $p(A,B \mid \Theta, H_p)$ we need to account for missing information, namely which minutia pairs $(a,b) \in A \times B$ correspond to the same latent minutia. To handle this, we first decompose M into four parts

$$\begin{aligned} M_{11} &= \{m \in M : I_A(m) = 1, I_B(m) = 1\}, & M_{10} &= \{m \in M : I_A(m) = 1,\ I_B(m) = 0\}, \\ M_{01} &= \{m \in M : I_A(m) = 0,\ I_A(m) = 1\}, & M_{00} &= \{m \in M : I_A(m) = 0, I_B(m) = 0\}, \end{aligned} \quad (2.10)$$

which are independent and disjoint MPPPs, and hence all four have known densities. Their intensity functions are

$$\begin{aligned} \rho_{11}(r) &= \rho(r)\delta_A(r)\delta_B(r), & \rho_{10}(r) &= \rho(r)\delta_A(r)\{1-\delta_B(r)\}, \\ \rho_{01}(r) &= \rho(r)\{1-\delta_A(r)\}\delta_B(r), & \rho_{00}(r) &= \rho(r)\{1-\delta_A(r)\}\{1-\delta_B(r)\}, \end{aligned} \quad (2.11)$$

and all four MPPPs have the same mark density g. Note that $M_{A1} = M_{11} \cup M_{10}$ and $M_{B1} = M_{11} \cup M_{01}$, so M_{00} will play no role in the sequel. This partitioning is visualized in figure 2.1.

Applying steps A2–A3 to M_{10} yields M_{103} with distribution MPPP(ρ_{103}, g_{A3}), where

$$\rho_{103}(r) = f_A * \rho_{10}\{(r-\tau_A)/\psi_A\}/|\psi_A|^2. \quad (2.12)$$

Similarly, applying steps B2–B3 to M_{01} yields $M_{013} \sim$ MPPP(ρ_{013}, g_{B3}) with

$$\rho_{013}(r) = f_B * \rho_{01}\{(r-\tau_B)/\psi_B\}/|\psi_B|^2. \quad (2.13)$$

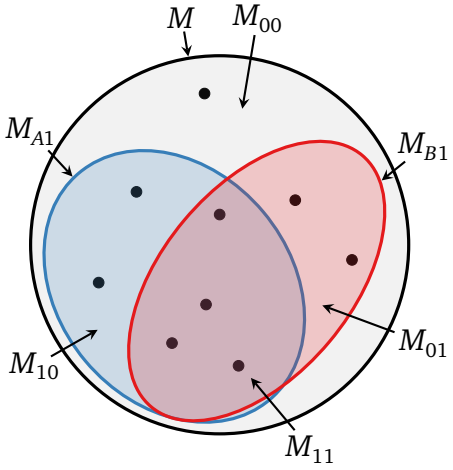


Figure 2.1: Visualization of the selection procedure under H_p. The red and blue ovals correspond to M_{A1} and M_{B1} respectively. The minutiae in their intersection comprise M_{11}: these minutiae are observed twice, once in M_{A1} and again in M_{B1}. The remaining sets are $M_{01} = M_{A1} \setminus M_{11}$, $M_{10} = M_{B1} \setminus M_{11}$, and $M_{00} = M \setminus (M_{A1} \cup M_{B1})$. For simplicity, we have drawn each minutia configuration as a contiguous region of M, however under our model this is unlikely to be the case since each minutia in M is allocated independently to one of $M_{00}, M_{10}, M_{01}, M_{11}$.

Finally, for each $m \in M_{11}$, we apply steps A2–A3 to m yielding a marked point $a(m)$, and separately we apply steps B2–B3 to m yielding a marked point $b(m)$. The set of paired marked points

$$M_{113} = \{(a(m), b(m)) : m \in M_{11}\} \tag{2.14}$$

forms a MPPP on $\mathbb{C} \times \mathbb{C}$. The paired points have the intensity function

$$\rho_{113}(r_a, r_b) = \int_{\mathbb{C}} \rho_{11}(r) f_A\{(r_a - \tau_A)/\psi_A - r\} f_B\{(r_b - \tau_B)/\psi_B - r\}/|\psi_A \psi_B|^2 \, \mathrm{d}r \tag{2.15}$$

and the corresponding paired marks are i.i.d. with density (with respect to $(\mu_{\mathbb{S}^1} \times \mu_{\mathbb{T}})^2$)

$$g_{113}(s_a, t_a, s_b, t_b) = \sum_{u \in \mathbb{T}} d_A(ut_a) d_B(ut_b) \int_{\mathbb{S}^1} g(s,u) h_A\left(\frac{s_a \overline{s\psi_A}}{|\psi_A|}\right) h_B\left(\frac{s_b \overline{s\psi_B}}{|\psi_B|}\right) \mathrm{d}\mu_{\mathbb{S}^1}(s). \tag{2.16}$$

The distribution of M_{113} is dominated by the probability measure $\zeta_2 = \mathrm{MPPP}(\varphi_2, 1/9)$, the MPPP whose points form a Poisson point process on $\mathbb{C} \times \mathbb{C}$ whose intensity function is a bivariate complex normal density and whose marks are uniformly distributed on $(\mathbb{S}^1 \times \mathbb{T})^2$. From (2.15) we have

$$\int_{\mathbb{C}^2} \rho_{113}(r_a, r_b) \, \mathrm{d}r_a \, \mathrm{d}r_b = \int_{\mathbb{C}} \rho_{11}(r) \, \mathrm{d}r, \tag{2.17}$$

and hence from (2.11) we see the density of M_{113} with respect to ζ_2 is

$$p(M_{113}\,|\,\Theta,H_p) = c_2(M_{113})\exp\left\{-\int_{\mathbb{C}}\rho_{11}(r)\,\mathrm{d}r\right\}\prod_{(a,b)\in M_{113}}\rho_{113}(r_a,r_b)g_{113}(s_a,t_a,s_b,t_b), \tag{2.18}$$

where the normalization constant is

$$c_2(M_{113}) = 9^{|M_{113}|}\exp(1)\prod_{(a,b)\in M_{113}}\{\varphi(r_a)\varphi(r_b)\}^{-1}. \tag{2.19}$$

Since $c(M_{103})c(M_{013})c_2(M_{113}) = \exp(1)c(A)c(B)$, the density for $(M_{103},M_{013},M_{113})$ with respect to $\zeta\times\zeta\times\zeta_2$ is

$$\begin{aligned} p(M_{103},M_{013},M_{113}\,|\,\Theta,H_p) = c(A)c(B)\exp\left[1-\int_{\mathbb{C}}\rho(r)\,\{\delta_A(r)+\delta_B(r)-\delta_A(r)\delta_B(r)\}\,\mathrm{d}r\right] \\ \times\left\{\prod_{a\in M_{103}}\rho_{103}(r_a)g_{A3}(s_a,t_a)\right\}\left\{\prod_{b\in M_{013}}\rho_{013}(r_b)g_{B3}(s_b,t_b)\right\} \\ \times\left\{\prod_{(a,b)\in M_{113}}\rho_{113}(r_a,r_b)g_{113}(s_a,t_a,s_b,t_b)\right\}. \end{aligned} \tag{2.20}$$

§2.2.3 Defining the matching

The three marked point processes $(M_{103},M_{013},M_{113})$ can be identified with a bipartite graph (A,B,ξ) of maximum degree one with edge set ξ and vertex sets labelled by A and B respectively. Thus, given the latent matching ξ, we can use (2.20) to find the joint density of A and B under H_p. Specifically, we have the transformation

$$A = M_{103}\cup\Pi_A(M_{113}),\quad B = M_{013}\cup\Pi_B(M_{113}),\quad \xi = \{\langle a,b\rangle : (a,b)\in M_{113}\}, \tag{2.21}$$

where Π_A,Π_B project from point sets on $\mathbb{M}\times\mathbb{M}$ to point sets on $\mathbb{M}$ via

$$\begin{aligned} \Pi_A(M_{113}) &= \{a : (a,b)\in M_{113} \text{ for some } b\in\mathbb{M}\}, \\ \Pi_B(M_{113}) &= \{b : (a,b)\in M_{113} \text{ for some } a\in\mathbb{M}\}. \end{aligned} \tag{2.22}$$

The inverse transformation is given by

$$M_{103} = A\setminus\Pi_A(\xi),\quad M_{013} = B\setminus\Pi_B(\xi),\quad M_{113} = \{(a,b) : \langle a,b\rangle\in\xi\}, \tag{2.23}$$

where now we slightly abuse notation and write Π_A, Π_B for the functions which send ξ to point sets on $\mathbb{M}$ via

$$\Pi_A(\xi) = \{a \in A : \langle a, b\rangle \in \xi \text{ for some } b \in \mathbb{M}\},$$
$$\Pi_B(\xi) = \{b \in B : \langle a, b\rangle \in \xi \text{ for some } a \in \mathbb{M}\}. \tag{2.24}$$

Notice that we use the notation $\langle a, b\rangle$ for the edge between the vertex labelled a and the vertex labelled b. Conversely, the notation (a, b) denotes a pair of marked points and lies in $\mathbb{M} \times \mathbb{M}$.

Now that we have a bijection between $(M_{103}, M_{013}, M_{113})$ and (A, B, ξ), our goal is to sum over the unknown matching ξ and thus find the desired density $p(A, B \mid \Theta, H_p)$. This sum will be over the support of ξ, which we now derive.

Let $n_A = |A|, n_B = |B|$, and $n_\xi = |\xi| = |M_{113}|$ denote the cardinalities of A, B, and ξ respectively. Let $\Xi(A, B, n_\xi)$ denote the space of all possible values for ξ with cardinality n_ξ, where $0 \le n_\xi \le \min(n_A, n_B)$. That is, $\Xi(A, B, n_\xi)$ contains all possible edge sets of size n_ξ for the vertex sets labelled by A and B. It has cardinality

$$|\Xi(A, B, n_\xi)| = \frac{n_A!}{n_\xi!(n_A - n_\xi)!} \frac{n_B!}{n_\xi!(n_B - n_\xi)!} n_\xi!. \tag{2.25}$$

This corresponds to choosing n_ξ points each from A and B to be matched and considering all $n_\xi!$ edge sets between those points. Let

$$\Xi(A, B) = \bigcup_{n_\xi=0}^{\min(n_A, n_B)} \Xi(A, B, n_\xi) \tag{2.26}$$

be the support of ξ, which consists of the union of the disjoint sets $\Xi(A, B, n_\xi)$. Note that the cardinalities $|\Xi(A, B, n_\xi)|$ and $|\Xi(A, B)|$ depend on A and B only through n_A and n_B. We let $\mu_{\Xi(A,B)}$ be the counting measure on $\Xi(A, B)$.

In order to sum over $\xi \in \Xi(A, B)$, we must first change the base measure of (2.20) from $\zeta \times \zeta \times \zeta_2$ to $\tilde{\zeta}$, which is defined by

$$\mathrm{d}\tilde{\zeta}(A, B, X) = \mu_{\Xi(A,B)}(X)\, \mathrm{d}\zeta(A)\, \mathrm{d}\zeta(B), \tag{2.27}$$

where $X \subseteq \Xi(A, B)$. The Radon-Nikodym derivative for this change of measure is shown to be $\exp(-1)$ in §B.1. Thus

$$p(A, B, \xi \mid \Theta, H_p) = \exp(-1) p(M_{103}, M_{013}, M_{113} \mid \Theta, H_p) \tag{2.28}$$

with respect to $\tilde{\zeta}$, where $p(M_{103}, M_{013}, M_{113} \,|\, \Theta, H_p)$ is given in (2.20). Furthermore, by summing over ξ we can compute

$$p(A, B \,|\, \Theta, H_p) = \sum_{\xi \in \Xi(A,B)} p(A, B, \xi \,|\, \Theta, H_p), \tag{2.29}$$

which is our desired density with respect to $\zeta \times \zeta$. Since $\zeta \times \zeta$ is also the base measure of $p(A, B \,|\, \Theta, H_d)$ in (2.9), we have found our desired likelihood ratio between (2.29) and (2.9).

§2.3 Parametric models

To complete the specification of our basic point process model we need to specify parametric models for the basic elements $\rho, g, \delta_A, \delta_B, f_A, f_B, h_A, h_B, d_A, d_B$ in §2.1 that define our MPPPs and the corresponding likelihood ratios. Clearly there are many possibilities. Below we specify a simple choice with the purpose of illustrating and investigating the methodology. We shall evaluate the fit of this simple model in chapter 7 and suggest various model enhancements in chapter 8. A visual overview of the parametric model is provided in figure 2.2, and a summary of the parameters is provided in §2.3.3. The final parameters are also listed in the glossary at the beginning of this dissertation for easier reference.

We model the latent minutia location intensity function $\rho(r)$ as proportional to a complex normal density, with intensity $\rho_0 > 0$, mean $\tau_0 \in \mathbb{C}$ and variance σ_0^2. We model the latent orientations as uniform over $\mathbb{S}^1$. We assume that the latent types are fully classified[†]: the probability of a bifurcation is $\chi \in (0, 1)$, and the probability of an ending is $1 - \chi$. Thus we have

$$\begin{aligned} \rho(r) &= \rho_0 \varphi(r; \tau_0, \sigma_0^2), \\ g(s, t) &= |t| \sqrt{\chi^{|t|+t} (1-\chi)^{|t|-t}} = \begin{cases} \chi & \text{if } t = 1, \\ 1 - \chi & \text{if } t = -1, \\ 0 & \text{if } t = 0. \end{cases} \end{aligned} \tag{2.30}$$

[†]The assumption that the latent types are all classified is without loss of generality, since any observed minutia may still have an unclassified type due to step A2.

Without loss of generality we assume that $\tau_0 = 0$, since this parameter can be absorbed into τ_A and τ_B. Similarly, we assume that $\sigma_0 = 1$, since this parameter can be absorbed into ψ_A and ψ_B. Due to the latent mark distribution $g(s,t)$ being uniform over s, we have

$$g_{A1}(s,t) = g(s,t), \quad g_{A2}(s,t) = g_{A3}(s,t) = d_A(t)\chi + d_A(-t)(1-\chi), \tag{2.31}$$

and similarly for B.

A1: thinning We assume the thinning probabilities are constant with $\delta_A(r) = \delta_A \in (0,1)$ and $\delta_B(r) = \delta_B \in (0,1)$ so that the intensities after thinning become

$$\rho_{A1}(r) = \rho_0\delta_A\varphi(r), \quad \rho_{B1}(r) = \rho_0\delta_B\varphi(r). \tag{2.32}$$

A2: displacement We assume the error distributions of the minutia locations have complex normal distributions with mean zero and variance ω^2. The type errors are Bernoulli distributed on $\{0,1\} \subset \mathbb{T}$, where ϵ is the probability that a type is unclassified. This assumes that there are no type misclassifications: all types are either correctly classified or else unclassified. Thus we have

$$\begin{gathered} f_A(r) = f_B(r) = \varphi(r;0,\omega^2), \quad d_A(c) = \mathbb{1}(c=0)\epsilon + \mathbb{1}(c=1)(1-\epsilon), \\ \rho_{A2}(r) = \rho_0\delta_A\varphi(r;0,1+\omega^2), \quad \rho_{B2}(r) = \rho_0\delta_B\varphi(r;0,1+\omega^2), \\ g_{A2}(s,t) = g_{B2}(s,t) = (1-|t|)\epsilon + |t|(1-\epsilon)g(s,t). \end{gathered} \tag{2.33}$$

The orientations errors $h_A = h_B = h$ are assumed to be root von Mises distributions as defined in §1.5, with mean 1 (i.e., no distortion) and precision κ.

A3: mapping After similarity transformations parametrized by τ_A, ψ_A and τ_B, ψ_B respectively, the observed locations in A and B have intensity functions

$$\begin{aligned} \rho_{A3}(r) &= \rho_0\delta_A\varphi\{r;\tau_A,(1+\omega^2)|\psi_A|^2\}, \\ \rho_{B3}(r) &= \rho_0\delta_B\varphi\{r;\tau_B,(1+\omega^2)|\psi_B|^2\}. \end{aligned} \tag{2.34}$$

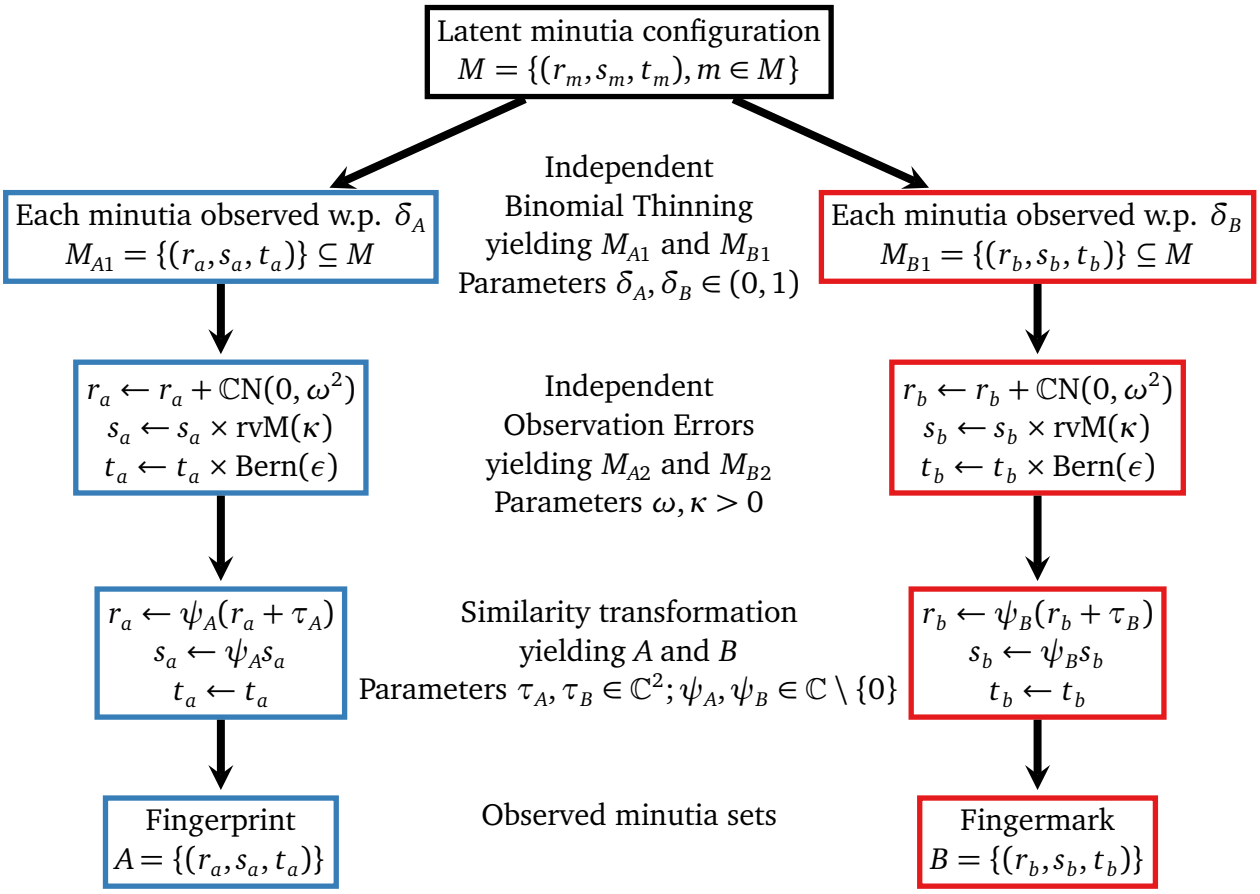


Figure 2.2: Illustration of the generative procedure for A and B under H_p. The procedure under H_d is the same, except that A and B originate from separate, independent latent minutia configurations M and M'.

For notational convenience, in the rest of this dissertation we will use the transformed variables

$$\begin{aligned} \sigma_A^2 &= (1+\omega^2)|\psi_A|^2, & \sigma_B^2 &= (1+\omega^2)|\psi_B|^2, \\ \psi &= \psi_A\overline{\psi_B}/(|\psi_A||\psi_B|), & \tilde{\omega} &= (\omega^2+1)^2/\{(\omega^2+1)^2-1\}. \end{aligned} \tag{2.35}$$

§2.3.1 Density under the defence hypothesis

By letting $n_A^{(t)} = \sum_{a\in A} \mathbb{1}(t_a = t)$ be the number of minutiae of each type $t \in \mathbb{T}$, and similarly for $n_B^{(t)}$, we can write the defence likelihood (2.9) as

$$\begin{aligned} p(A,B \mid \Theta, H_d) = \tilde{c}(A)\tilde{c}(B)\exp\{-\rho_0(\delta_A+\delta_B)\}\,\rho_0^{n_A+n_B}\delta_A^{n_A}\delta_B^{n_B} \\ \times \chi^{n_A^{(1)}+n_B^{(1)}}(1-\chi)^{n_A^{(-1)}+n_B^{(-1)}}\left\{\prod_{a\in A}\varphi(r_a;\tau_A,\sigma_A^2)\right\}\left\{\prod_{b\in B}\varphi(r_b;\tau_B,\sigma_B^2)\right\}. \end{aligned} \tag{2.36}$$

The normalization constant $\tilde{c}(A)$ is given by $c(A)\varepsilon^{n_A^{(0)}}(1-\varepsilon)^{n_A^{(-1)}+n_A^{(1)}}$, and similarly for $\tilde{c}(B)$. We shall see that $\tilde{c}(A)$ and $\tilde{c}(B)$ appear under the prosecution hypothesis as well, hence they will cancel in the likelihood ratio, and the value of ϵ will not affect the likelihood ratio.

§2.3.2 Density under the prosecution hypothesis

The transformed intensities for the unmatched minutiae become

$$\rho_{103}(r_a) = \rho_0\delta_A(1-\delta_B)\varphi(r_a;\tau_A,\sigma_A^2), \quad \rho_{013}(r_b) = \rho_0(1-\delta_A)\delta_B\varphi(r_b;\tau_B,\sigma_B^2). \tag{2.37}$$

The matched density (2.15) becomes

$$\rho_{113}(r_a,r_b) = \rho_0\delta_A\delta_B\varphi_2(r_a,r_b;\tau_A,\tau_B,\Sigma_{AB}), \tag{2.38}$$

where the covariance matrix is

$$\Sigma_{AB} = \begin{pmatrix} \sigma_A^2 & \sigma_A\sigma_B\psi\sqrt{1-1/\tilde{\omega}} \\ \sigma_A\sigma_B\overline{\psi}\sqrt{1-1/\tilde{\omega}} & \sigma_B^2 \end{pmatrix}. \tag{2.39}$$

The mark density (2.16) becomes

$$\begin{aligned} g_{113}(s_a,t_a,s_b,t_b) &= g_{A2}(s_a,t_a)g_{B2}(s_b,t_b)T(t_a,t_b)\exp\{\kappa\,\mathrm{Re}(s_a\overline{s_b\psi})\}/I_0(\kappa) \\ &= \epsilon^2\left(\frac{1-\epsilon}{\epsilon}\right)^{|t_a|+|t_b|}\frac{T(t_a,t_b)}{I_0(\kappa)}\sqrt{\chi^{|t_a|+t_a+|t_b|+t_b}(1-\chi)^{|t_a|-t_a+|t_b|-t_b}}\exp\{\kappa\,\mathrm{Re}(s_a\overline{s_b\psi})\}, \end{aligned} \tag{2.40}$$

where $T(t_a,t_b)$ is given by

$$T(t_a,t_b) = \begin{cases} 1/\chi & \text{if } t_a = t_b = 1, \\ 1/(1-\chi) & \text{if } t_a = t_b = -1, \\ 1 & \text{if } t_a t_b = 0, \\ 0 & \text{if } t_a t_b = -1. \end{cases} \tag{2.41}$$

By substituting these into (2.20) and using (2.28), we have

$$\begin{aligned} p(A,B,\xi \mid \Theta, H_p) = \tilde{c}(A)\tilde{c}(B)\exp\{-\rho_0(\delta_A+\delta_B-\delta_A\delta_B)\}\,\rho_0^{n_A+n_B-n_\xi}\delta_A^{n_A}(1-\delta_A)^{n_B-n_\xi} \\ \times \delta_B^{n_B}(1-\delta_B)^{n_A-n_\xi}\chi^{n_A^{(1)}+n_B^{(1)}-n_\xi^{(1)}}(1-\chi)^{n_A^{(-1)}+n_B^{(-1)}-n_\xi^{(-1)}} \\ \times \left\{\prod_{a\in A\setminus\Pi_A(\xi)}\varphi(r_a;\tau_A,\sigma_A^2)\right\}\left\{\prod_{b\in B\setminus\Pi_B(\xi)}\varphi(r_b;\tau_B,\sigma_B^2)\right\} \\ \times \left[\prod_{\langle a,b\rangle\in\xi}\varphi_2(r_a,r_b;\tau_A,\tau_B,\Sigma_{AB})\frac{\exp\{\kappa\,\mathrm{Re}(s_a\overline{s_b\psi})\}}{I_0(\kappa)}\,\mathbb{1}(t_at_b\geq 0)\right], \end{aligned} \tag{2.42}$$

where $n_\xi^{(t)} = \sum_{\langle a,b\rangle\in\xi}\mathbb{1}(t_a=t)\,\mathbb{1}(t_b=t)$.

§2.3.3 Estimating or marginalizing the parameters

The densities in the parametric models specified above depend on the following set of eleven variation independent parameters

$$\{\rho_0,\chi,\delta_A,\delta_B,\tau_A,\tau_B,\sigma_A,\sigma_B,\psi,\tilde{\omega},\kappa\}, \tag{2.43}$$

where $\rho_0>0$, $\chi,\delta_A,\delta_B\in(0,1)$, $\tau_A,\tau_B\in\mathbb{C}$, $\sigma_A,\sigma_B>0$, $\psi\in\mathbb{S}^1$, $\tilde{\omega}>1$, and $\kappa>0$. As τ_A and τ_B are complex numbers there are thirteen real parameters in total. Of these, ρ_0 and χ relate to the latent minutiae and are common to all fingerprints and fingermarks under consideration. We treat both ρ_0 and χ as known constants; their estimation is addressed in §6.2.

The remaining parameters

$$\theta = \{\delta_A,\delta_B,\tau_A,\tau_B,\sigma_A,\sigma_B,\psi,\tilde{\omega},\kappa\} \tag{2.44}$$

vary from one fingerprint or fingermark to the next, and must be estimated or marginalized for each fingerprint/fingermark pair. We shall assume that $\tilde{\omega}$ has some lower bound $c_\omega>1$. This is equivalent to putting an upper bound on location noise variance ω^2. This increases the stability of our estimates: it prevents the undesirable scenario where our estimate of ω^2 becomes very large and thus all matchings between minutiae become plausible.

There are two likelihood ratios of potential interest. The first maximizes the likelihoods over θ and ξ and takes their ratio. This method has the advantage of a relatively easy, computationally inexpensive implementation. Furthermore, since courtrooms are already used to seeing maximized likelihood ratios in relation to DNA evidence (Balding, 2005), this method may be preferred by practising forensic scientists. However, it takes considerable care to ensure this likelihood ratio is well defined. The details of this method are presented in chapter 3.

The second method assigns prior distributions to θ, sums $p(A, B, \theta, \xi \mid H_p)$ over ξ, integrates both $p(A, B, \theta \mid H_p)$ and $p(A, B, \theta \mid H_d)$ over θ, and takes their ratio. The prior distributions must be chosen with care, since they will affect the interpretation of the resultant likelihood ratio. This method has the advantage of being independent of the choice of base measure for our parameters. Furthermore, it more naturally accounts for the high dimensionality of the matching ξ, and it is likely to be preferred by Bayesian-inclined statisticians. However, this method is difficult and computationally expensive to implement. Our approach to estimating this likelihood ratio is presented in chapter 4 and chapter 5.

Many parts of this report can be simplified by making additional assumptions. Please refer to Forbes et al. (2014) for the special case where the scale parameters are constrained to be equal, $\sigma_A = \sigma_B$, and the variability parameters $\tilde{\omega}, \kappa$ are treated as fixed and known constants.

Chapter Three

Profile likelihood ratio

The standard maximized likelihood ratio is usually written as

$$\frac{\max_{\theta_p,\xi} p(A,B \mid \theta_p, \xi, H_p)}{\max_{\theta_d} p(A,B \mid \theta_d, H_d)}. \tag{3.1}$$

In our case the numerator is not even well defined: we showed in (2.26) that the support of the matching ξ depends on the observed minutia configurations A and B, and thus it does not make sense to condition the probability of A and B on ξ. One alternative is to consider the ratio after summing over the matching:

$$\frac{\max_{\theta_p} p(A,B \mid \theta_p, H_p)}{\max_{\theta_d} p(A,B \mid \theta_d, H_d)} = \frac{\max_{\theta_p} \left\{\sum_{\xi \in \Xi(A,B)} p(A,B,\xi \mid \theta_p, H_p)\right\}}{\max_{\theta_d} p(A,B \mid \theta_d, H_d)}. \tag{3.2}$$

However, this sum is too large to compute, and thus we cannot find the maximum over θ_p marginal of ξ. We proceed by replacing the sum with its largest summand. Letting

$$\breve{\xi}(\theta_p) = \operatorname*{argmax}_{\xi} p(A,B,\xi \mid \theta_p, H_p), \tag{3.3}$$

we define the *profile likelihood ratio* as the ratio of (2.42) and (2.36),

$$\frac{\max_{\theta_p} p\{A,B,\breve{\xi}(\theta_p) \mid \theta_p, H_p\}}{\max_{\theta_d} p(A,B \mid \theta_d, H_d)}. \tag{3.4}$$

Since we have replaced the sum with a single term, the profile likelihood ratio is more favourable to the defence hypothesis than (3.2). If the data A and B admit a single likely matching, then $\breve{\xi}(\theta_p)$ will dominate the sum and the profile likelihood ratio will be similar to (3.2). Conversely, if there is not a single likely matching, then the profile

likelihood ratio will be much smaller than (3.2); however, this is actually a desirable property since we do not wish to have a high likelihood ratio if there is no single good matching between the fingerprint and the fingermark.

We shall need various functions of the observed data. We define all such terms now for easy reference:

$$\begin{aligned}
R_{A\setminus\xi} &= \sum_{a\in A\setminus\Pi_A(\xi)} r_a, & R_{B\setminus\xi} &= \sum_{b\in B\setminus\Pi_B(\xi)} r_b,\\
R_{\xi A} &= \sum_{\langle a,b\rangle\in\xi} r_a, & R_{\xi B} &= \sum_{\langle a,b\rangle\in\xi} r_b,\\
S_{A\setminus\xi} &= \sum_{a\in A\setminus\Pi_A(\xi)} |r_a|^2, & S_{B\setminus\xi} &= \sum_{b\in B\setminus\Pi_B(\xi)} |r_b|^2,\\
S_{\xi A} &= \sum_{\langle a,b\rangle\in\xi} |r_a|^2, & S_{\xi B} &= \sum_{\langle a,b\rangle\in\xi} |r_b|^2,\\
S_{\xi AB} &= \sum_{\langle a,b\rangle\in\xi} \overline{r_a} r_b, & S_{\xi} &= \sum_{\langle a,b\rangle\in\xi} \overline{s_a} s_b.
\end{aligned} \tag{3.5}$$

These functions depend on the matching ξ. Under H_d we set ξ equal to the empty set $\emptyset$; this means $R_{\xi A} = R_{\xi B} = S_{\xi A} = S_{\xi B} = S_{\xi AB} = S_{\xi} = 0$ under H_d.

§3.1 Maximizing the likelihood under H_d

The parameters $\psi, \tilde{\omega}$, and κ do not enter into the likelihood under H_d. We can explicitly maximize $p(A,B \mid \theta, H_d)$, as given in (2.36), over the remaining parameters, yielding

$$\begin{aligned}
\delta_A &= \min(1, n_A/\rho_0), \qquad \tau_A = \frac{R_{A\setminus\xi}}{n_A}, \qquad \sigma_A^2 = \frac{S_{A\setminus\xi} - 2\,\mathrm{Re}(\overline{R_{A\setminus\xi}}\tau_A)}{n_A} + |\tau_A|^2,\\
\delta_B &= \min(1, n_B/\rho_0), \qquad \tau_B = \frac{R_{B\setminus\xi}}{n_B}, \qquad \sigma_B^2 = \frac{S_{B\setminus\xi} - 2\,\mathrm{Re}(\overline{R_{B\setminus\xi}}\tau_B)}{n_B} + |\tau_B|^2.
\end{aligned} \tag{3.6}$$

§3.2 Iteratively maximizing the likelihood under H_p

Note that $\max_{\theta_p} p\{A,B,\breve{\xi}(\theta_p) \mid \theta_p, H_p\} = \max_{\theta_p,\xi} p(A,B,\xi \mid \theta_p, H_p)$, and thus we can find the former by iteratively maximizing (2.42) over $\xi \mid \theta$ and $\theta \mid \xi$. The first

maximization can be done exactly, as described in §3.2.1. The latter maximization is itself done iteratively, as described in §3.2.2.

Since our likelihood is multimodal this iterative approach may converge to a local maximum rather than the global maximum.[†] In practice we can often find the global maximum by running the Markov Chain Monte Carlo (MCMC) procedure described in chapter 5 to find a good initial position for ξ. The MCMC procedure is often able to identify the global mode because it is usually much larger than any secondary modes: when two minutia configurations match, there are typically a large number of matching minutia pairs, and the posterior distribution is heavily concentrated around the true matching. Given an initial value for ξ, we can use the score matching estimator (see §3.3) to quickly find a good starting value for θ.

§3.2.1 Maximizing over the matching $\xi \mid \theta$

We can find the maximizer over ξ conditional on θ using the *Hungarian algorithm* (Kuhn, 1955). Formally, the Hungarian algorithm finds a bijection $\pi : A_1 \to B_1$ between two finite sets of the same cardinality which minimizes $\sum_{a \in A_1} \text{cost}\{a, \pi(a)\}$, where $\text{cost} : A_1 \times B_1 \to \mathbb{R}$ is some loss function. The bijection can be computed in a time proportional to $|A_1|^3$.

We cast our problem into this form by setting $A_1 = A \cup \{\phi_{i,1} : 1 \le i \le n_B\}$ and $B_1 = B \cup \{\phi_{i,2} : 1 \le i \le n_A\}$, where $\phi_{i,j} \in \mathbb{M}$ are dummy elements whose only relevant property is $\phi_{i,j} \notin A \cup B$ for all i, j. These dummy elements are necessary because the Hungarian algorithm can only find a bijection between two sets of the same cardinality, but we wish to consider partial matchings between A and B. The

[†]For pathological A and B the procedure may not converge at all. Indeed, the likelihood may not even have a unique global maximum; see Wu (1983) for examples of what may go wrong. However, for the data we have considered this has not been an issue.

cost function is $\text{cost}(a,b) = -w(a,b \mid \theta)$, where

$$w(a,b \mid \theta) = \mathbb{1}(a \in A)\,\mathbb{1}(b \in B)\Bigg[\text{Re}\left(\kappa s_a \overline{\psi s_b} + 2\sqrt{\tilde{\omega}^2 - \tilde{\omega}\overline{\psi}}\,\frac{r_a - \tau_A}{\sigma_A}\,\overline{\frac{r_b - \tau_B}{\sigma_B}}\right) \\ -(\tilde{\omega}-1)\left(\frac{|r_a-\tau_A|^2}{\sigma_A^2} + \frac{|r_b-\tau_B|^2}{\sigma_B^2}\right) + \log\left\{\frac{T(t_a,t_b)\tilde{\omega}}{\rho_0 I_0(\kappa)(1-\delta_A)(1-\delta_B)}\right\}\Bigg], \quad (3.7)$$

and the function $T(t_a,t_b)$ is defined in (2.41). We convert the optimal bijection π into a matching via

$$\xi = \bigcup_{a \in A:\pi(a)\in B} \{\langle a, \pi(a)\rangle\}. \quad (3.8)$$

§3.2.2 Maximizing over $\theta \mid \xi$

The product structure of the likelihood $p(A,B \mid \theta, H_p)$ means that we can maximize over δ_A and δ_B directly. The remaining parameters in θ are maximized one at a time, and for fixed values of all other variables: first we maximize (τ_A,τ_B), then (σ_A,σ_B), then ψ, then $\tilde{\omega}$, and finally κ. We iterate this procedure until the parameters converge. If this procedure is to converge to the global maximum over $\theta \mid \xi$ we must find a good initial position for θ, which will depend on the value of ξ. We describe our method of finding this initial position in §3.3.

Maximizing over the thinning probabilities δ_A, δ_B

As a function of δ_A and δ_B, the log-likelihood under H_p is proportional to

$$-\rho_0(\delta_A+\delta_B-\delta_A\delta_B)+n_A\log\delta_A+n_B\log\delta_B+(n_B-n_\xi)\log(1-\delta_A)+(n_A-n_\xi)\log(1-\delta_B). \quad (3.9)$$

We find the maximizing value by setting the derivatives with respect to δ_A and δ_B to zero. Solving for δ_B yields the cubic equation

$$\rho_0 n_A {\delta_B}^3 - \rho_0(n_A+n_B){\delta_B}^2 + n_B(\rho_0 + n_A + n_B - n_\xi)\delta_B - n_B^2 = 0. \quad (3.10)$$

One of the roots of this cubic is the maximizing value δ_B. The second equation yields

$$\delta_A = \max\left[0, \min\left\{1, 1 - \frac{n_B - (n_A+n_B-n_\xi)\delta_B}{\rho_0\delta_B(1-\delta_B)}\right\}\right]. \quad (3.11)$$

Maximizing over the translation parameters τ_A, τ_B

We maximize the likelihood over τ_A and τ_B conditional on $\sigma_A, \sigma_B, \psi, \tilde{\omega}$, and ξ. The maximizing values are

$$\begin{pmatrix}\tau_A\\ \tau_B\end{pmatrix} = \left\{\begin{pmatrix} n_A\sigma_A^{-2} & 0\\ 0 & n_B\sigma_B^{-2}\end{pmatrix} + n_\xi \Sigma_{AB}^{-1}\right\}^{-1}\left\{\begin{pmatrix}\sigma_A^{-2}R_{A\backslash\xi}\\ \sigma_B^{-2}R_{B\backslash\xi}\end{pmatrix} + \Sigma_{AB}^{-1}\begin{pmatrix}R_{\xi A}\\ R_{\xi B}\end{pmatrix}\right\}, \tag{3.12}$$

where Σ_{AB} is given in (2.39).

Maximizing over the scale parameters σ_A, σ_B

As a function of σ_A^{-2} and σ_B^{-2}, the log-likelihood (2.42) is proportional to

$$n_A \log(\sigma_A^{-2}) + n_B \log(\sigma_B^{-2}) - R_1\sigma_A^{-2} - R_2\sigma_B^{-2} - 2R_3\sqrt{\sigma_A^{-2}\sigma_B^{-2}}, \tag{3.13}$$

where

$$\begin{aligned} R_1 &= S_{A\backslash\xi} + \tilde{\omega}S_{\xi A} - 2\,\mathrm{Re}\{(R_{A\backslash\xi} + \tilde{\omega}R_{\xi A})\overline{\tau_A}\} + \{n_A + n_\xi(\tilde{\omega}-1)\}|\tau_A|^2,\\ R_2 &= S_{B\backslash\xi} + \tilde{\omega}S_{\xi B} - 2\,\mathrm{Re}\{(R_{B\backslash\xi} + \tilde{\omega}R_{\xi B})\overline{\tau_B}\} + \{n_B + n_\xi(\tilde{\omega}-1)\}|\tau_B|^2,\\ R_3 &= -\sqrt{\tilde{\omega}^2-\tilde{\omega}}\,\mathrm{Re}\left\{\psi\left(S_{\xi AB} - \overline{R_{\xi A}}\tau_B - R_{\xi B}\overline{\tau_A} + n_\xi\overline{\tau_A}\tau_B\right)\right\}. \end{aligned} \tag{3.14}$$

Taking the derivative with respect to σ_A^{-2} and σ_B^{-2} and setting it to zero, we see the maximizers are

$$\begin{aligned} \sigma_A^2 &= \frac{2R_1(R_1R_2 - R_3^2)}{(n_B-n_A)R_3^2 + 2n_AR_1R_2 + |R_3|\sqrt{\{(n_B-n_A)R_3\}^2 + 4n_An_BR_1R_2}},\\ \sigma_B^2 &= R_3^2\sigma_A^2/\left(n_A\sigma_A^2 - R_1\right)^2. \end{aligned} \tag{3.15}$$

Maximizing over the rotation parameter ψ

Optimizing the likelihood over ψ yields $\psi = \nu/|\nu|$, where

$$\nu = \kappa\overline{S_\xi} + 2\sqrt{\tilde{\omega}^2-\tilde{\omega}}(\overline{S_{\xi AB}} - \overline{R_{\xi B}}\tau_A - R_{\xi A}\overline{\tau_B} + n_\xi\tau_A\overline{\tau_B})/(\sigma_A\sigma_B). \tag{3.16}$$

Maximizing over the location distortion parameter $\tilde{\omega}$

As a function of $\tilde{\omega}$, the log-likelihood is proportional to

$$n_\xi \log\tilde{\omega} - R_4\tilde{\omega} + 2R_5\sqrt{\tilde{\omega}^2-\tilde{\omega}}, \tag{3.17}$$

where

$$R_4 = \frac{S_{\xi A} - 2\,\mathrm{Re}(R_{\xi A}\overline{\tau_A}) + n_\xi|\tau_A|^2}{\sigma_A^2} + \frac{S_{\xi B} - 2\,\mathrm{Re}(R_{\xi B}\overline{\tau_B}) + n_\xi|\tau_B|^2}{\sigma_B^2},$$
$$R_5 = \frac{1}{\sigma_A\sigma_B}\,\mathrm{Re}\{\psi(S_{\xi AB} - R_{\xi B}\overline{\tau_A} - \overline{R_{\xi A}}\tau_B + n_\xi\overline{\tau_A}\tau_B)\}. \tag{3.18}$$

The maximizing value for $\tilde{\omega}$ is either its boundary value c_ω or a real root of

$$(R_4^2 - 4R_5^2)\tilde{\omega}^3 + (4R_5^2 - 2n_\xi R_4 - R_4^2)\tilde{\omega}^2 + (n_\xi^2 + 2n_\xi R_4 - R_5^2)\tilde{\omega} - n_\xi^2 = 0. \tag{3.19}$$

This can be solved using the cubic formula.

Maximizing over the orientation distortion parameter κ

The maximizing value for κ depends on ψ and ξ; it solves

$$n_\xi \frac{I_1(\kappa)}{I_0(\kappa)} = \mathrm{Re}\left(\psi S_\xi\right), \tag{3.20}$$

where I_1 is the modified Bessel function of the first kind and first order (Olver et al., 2010, §10). This equation can be solved numerically using the Newton–Raphson based technique described in §B.3.

§3.3 Find an initial value for $\theta \mid \xi, H_p$

In § 3.2.2 we described a method for iteratively maximizing the likelihood $p(A,B \mid \theta, \xi, H_p)$ over θ for fixed ξ. This iterative procedure can only produce a locally optimal value for θ. In order for the procedure to be effective we must start the iterations at a value of θ close to the global maximum, which will depend on the current value of ξ. We now describe a method to find such initial values.

The score matching estimator (see chapter 11; also Hyvärinen (2005, 2007); Forbes and Lauritzen (2014)) provides a computationally efficient and statistically consistent estimate of a model's parameters. We use it here to find initial positions for $\tau_A, \tau_B, \sigma_A, \sigma_B, \psi, \tilde{\omega}$, and κ.

The score matching estimate, which will be described in detail inchapter 11, is defined as the minimizer of the objective function (11.8). In this case, the objective

function is

$$\sum_{m\in A\cup B}\left\{\frac{1}{2}\left|\frac{\mathrm{d}\ell}{\mathrm{d}r_m}\right|^2+\frac{1}{2}\left|\frac{\mathrm{d}\ell}{\mathrm{d}s_m}\right|^2+\frac{\mathrm{d}^2\ell}{\mathrm{d}r_m^2}+\frac{\mathrm{d}^2\ell}{\mathrm{d}s_m^2}\right\}, \tag{3.21}$$

where $\ell = \log p(A,B,\xi\,|\,\Theta,H_p)$ is given by (2.42). Note that this estimator is not invariant under reparameterization of the observed data, nor is it invariant under change of base measure.

By defining $k_A = \sigma_A^{-2}, k_B = \sigma_B^{-2}, \tilde{\psi} = \sqrt{\tilde{\omega}^2-\tilde{\omega}}\psi/(\sigma_A\sigma_B)$, and $\tilde{\kappa} = \kappa\sigma_A\sigma_B/\sqrt{\tilde{\omega}^2-\tilde{\omega}}$, we can write all the derivatives as polynomials in the positive real variables $(k_A, k_B, \tilde{\omega}, \tilde{\kappa})$ and the complex variables $(\tau_A, \tau_B, \tilde{\psi})$. Note that this is an overparameterization of the parameter space and we have the constraint $|\tilde{\psi}|^2 = k_A k_B(\tilde{\omega}^2-\tilde{\omega})$.

For the locations, the derivatives are

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}r_a} &= 2r_a - 2k_A(r_a-\tau_A)\tilde{\omega}^{\mathbb{1}\{a\in\Pi_A(\xi)\}} + 2\sum_{b:\langle a,b\rangle\in\xi}\tilde{\psi}(r_b-\tau_B),\\
\frac{\mathrm{d}\ell}{\mathrm{d}r_b} &= 2r_b - 2k_B(r_b-\tau_B)\tilde{\omega}^{\mathbb{1}\{b\in\Pi_B(\xi)\}} + 2\sum_{a:\langle a,b\rangle\in\xi}\overline{\tilde{\psi}}(r_a-\tau_A),\\
\frac{\mathrm{d}^2\ell}{\mathrm{d}^2r_a} &= 2-2k_A\tilde{\omega}^{\mathbb{1}\{a\in\Pi_A(\xi)\}}, \qquad \frac{\mathrm{d}^2\ell}{\mathrm{d}^2r_b} = 2-2k_B\tilde{\omega}^{\mathbb{1}\{b\in\Pi_B(\xi)\}}.
\end{aligned} \tag{3.22}$$

The first terms $2r_a$ and $2r_b$ appear due to $\tilde{c}(A)$ and $\tilde{c}(B)$, which in turn appear due to our choice of base measure. Note that the above summations each contain at most one term, and they contain no terms if the relevant minutia is unmatched.

The derivatives of our orientations over $\mathbb{S}^1$ are

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}s_a} &= \sum_{b:\langle a,b\rangle\in\xi}\tilde{\kappa}\,\mathrm{Im}(\overline{s_a}\tilde{\psi}s_b), \quad \frac{\mathrm{d}^2\ell}{\mathrm{d}^2s_a} = \sum_{b:\langle a,b\rangle\in\xi}-\tilde{\kappa}\,\mathrm{Re}(\overline{s_a}\tilde{\psi}s_b),\\
\frac{\mathrm{d}\ell}{\mathrm{d}s_b} &= \sum_{a:\langle a,b\rangle\in\xi}\tilde{\kappa}\,\mathrm{Im}(s_a\overline{s_b\tilde{\psi}}), \quad \frac{\mathrm{d}^2\ell}{\mathrm{d}^2s_b} = \sum_{a:\langle a,b\rangle\in\xi}-\tilde{\kappa}\,\mathrm{Re}(s_a\overline{s_b\tilde{\psi}}),
\end{aligned} \tag{3.23}$$

where $\mathrm{Im}(z) = (z-\overline{z})/2$ is the imaginary part of z. Substituting these derivatives into (3.21) yields a sixth degree polynomial. To this polynomial we add the Lagrangian constraint term $\lambda\{|\tilde{\psi}|^2 - k_A k_B(\tilde{\omega}^2-\tilde{\omega})\}$, yielding another sixth degree polynomial to minimize.

This polynomial can be globally minimized via efficient semidefinite programming methods (see, e.g., chapter 18 of Anjos 2012). In practice, even a basic Newton-

Raphson approach suffices to quickly find the minimum. We use this minimizing value as an initial value in our iterative maximization procedure.

Chapter Four

Marginal likelihood ratio

We are interested in calculating the (marginal) *likelihood ratio*

$$\Lambda = p(A, B \mid H_p)/p(A, B \mid H_d) \tag{4.1}$$

for assessing the strength of the evidence for H_p. Though some may prefer to call Λ a Bayes factor, integrated likelihood ratio, or marginal likelihood ratio, we use the term likelihood ratio to conform with standard terminology in forensic science. In this chapter we will investigate methods to approximate Λ. We first specify distributions for $p(\theta \mid H_p)$ and $p(\theta \mid H_d)$.

§4.1 Distributions on parameters

When specifying the distribution of θ, we require that it is uninformative for the model choice. Formally, we follow Dawid and Lauritzen (2000) and ensure that we use compatible distributions for the competing models H_d and H_p. This leads to the condition that the marginal distributions of the print and mark are independent of the model: $p(A \mid H_p) = p(A \mid H_d)$ and $p(B \mid H_p) = p(B \mid H_d)$. This leads to the constraint $\mathbb{E}_\theta\{p(A \mid \theta) \mid H_p\} = \mathbb{E}_\theta\{p(A \mid \theta) \mid H_d\}$ for any value of A, where the expectations are taken under the model H_p and H_d respectively. For the parametric model described in

Sections 4.1, 4.2 and 4.3.2 are based off Forbes et al. (2014).

§2.3, the constraint becomes

$$\int \{p(\delta_A, \tau_A, \sigma_A | H_p) - p(\delta_A, \tau_A, \sigma_A | H_d)\} \times \left(\frac{\delta_A}{\sigma_A^2}\right)^{n_A} \exp\left(-\rho_0 \delta_A - \frac{|\tau_A|^2 - 2|\tau_A r_1| + |r_2|^2}{\sigma_A^2}\right) \mathrm{d}(\delta_A, \tau_A, \sigma_A) = 0 \quad (4.2)$$

for all $r_1, r_2 \in \mathbb{C}$, and all non-negative integers n_A. The fundamental lemma of the calculus of variations then implies $p(\delta_A, \tau_A, \sigma_A | H_p) = p(\delta_A, \tau_A, \sigma_A | H_d)$ almost everywhere. The remaining parameters $\psi, \tilde{\omega}, \kappa$ do not enter under H_d and are thus unconstrained by this consideration. These parameters will be given priors which are estimated using our dataset by empirical Bayes (see §6.1).

The print thinning probability δ_A is given a conjugate beta distribution with parameters $(\alpha_\delta, \beta_\delta)$; assuming that our dataset is representative for the number of minutiae in a fingerprint, these parameters can be estimated reliably. Conversely, the mark thinning probability δ_B is given a uniform distribution on $(0, 1)$. This is because our dataset is stratified on the fingermark qualities (see §6.1), and therefore it is not representative for the number of minutiae in a generic fingermark.

The translation and scale parameters are given conjugate distributions. That is, the scale parameters σ_A^{-2} and σ_B^{-2} both have gamma distributions with shape α_σ and scale β_σ, and conditional on these, the translations τ_A and τ_B have complex normal distributions with mean zero and variance σ_A^2/k_τ and σ_B^2/k_τ respectively for some $k_\tau > 0$. The densities for τ_A and σ_A are thus

$$p(\tau_A | \sigma_A) = \frac{k_\tau}{\pi \sigma_A^2} \exp\left(-\frac{k_\tau |\tau_A|^2}{\sigma_A^2}\right), \quad p(\sigma_A) = \frac{\beta_\sigma^{\alpha_\sigma} \sigma_A^{-2(\alpha_\sigma - 1)}}{\Gamma(\alpha_\sigma)} \exp\left(-\beta_\sigma \sigma_A^{-2}\right) \times 2\sigma_A^{-3}, \quad (4.3)$$

where the final term $2\sigma_A^{-3}$ is a Jacobian, and Γ is the gamma function (Olver et al., 2010, §5). The densities for τ_B and σ_B are analogous.

We use a uniform distribution over $\mathbb{S}^1$ for the rotation parameter ψ. For the location distortion parameter $\tilde{\omega}$ we use a gamma distribution with shape α_ω and rate β_ω, left truncated at c_ω. For the orientation distortion parameter κ we use the

conjugate distribution

$$p(\kappa) = \frac{\lambda_\kappa(\alpha_\kappa, \beta_\kappa)}{I_0(\kappa)^{2(\alpha_\kappa - 1)}} \exp\{(2\alpha_\kappa - 2 - \beta_\kappa)\kappa\}, \tag{4.4}$$

which, for large κ, behaves like a Gamma$(\alpha_\kappa, \beta_\kappa)$ since $I_0(\kappa) \to \exp(\kappa)/\sqrt{2\pi\kappa}$ as $\kappa \to \infty$ (Olver et al., 2010, §10.30). The normalization constant $\lambda_\kappa(\alpha_\kappa, \beta_\kappa)$ depends only on the hyperparameters, so it may be found once by a one-dimensional numeric integration and used for all subsequent likelihood computations.

The joint density of $\theta = \{\delta_A, \delta_B, \tau_A, \tau_B, \sigma_A, \sigma_B, \psi, \tilde{\omega}, \kappa\}$ is thus

$$\begin{aligned} p(\theta) &= p(\theta \,|\, H_p) = p(\theta \,|\, H_d) \\ &= \frac{4\Gamma(\alpha_\delta + \beta_\delta) k_\tau^2 \beta_\sigma^{2\alpha_\sigma} \beta_\omega^{\alpha_\omega} \lambda_\kappa(\alpha_\kappa, \beta_\kappa) \delta_A^{\alpha_\delta - 1} (1 - \delta_A)^{\beta_\delta - 1} (\sigma_A \sigma_B)^{-2\alpha_\sigma - 3} \tilde{\omega}^{\alpha_\omega - 1}}{\pi^2 \Gamma(\alpha_\delta)\Gamma(\beta_\delta)\Gamma(\alpha_\sigma)^2 \Gamma(\alpha_\omega, \beta_\omega c_\omega) I_0(\kappa)^{2(\alpha_\kappa - 1)}} \\ &\quad \times \exp\left\{ -\frac{k_\tau |\tau_A|^2}{\sigma_A^2} - \frac{k_\tau |\tau_B|^2}{\sigma_B^2} - \beta_\sigma \sigma_A^{-2} - \beta_\sigma \sigma_B^{-2} - \beta_\omega \tilde{\omega} + (2\alpha_\kappa - 2 - \beta_\kappa)\kappa \right\}, \end{aligned} \tag{4.5}$$

where $\Gamma(\alpha_\omega, \beta_\omega c_\omega)$ is the upper incomplete Gamma function (Olver et al., 2010, §8.2). This density is with respect to the measure $\mu_\theta = \mu_{\mathbb{R}} \times \mu_{\mathbb{R}} \times \mu_{\mathbb{C}} \times \mu_{\mathbb{C}} \times \mu_{\mathbb{R}} \times \mu_{\mathbb{R}} \times \mu_{\mathbb{S}^1} \times \mu_{\mathbb{R}} \times \mu_{\mathbb{R}}$, where $\mu_{\mathbb{R}}$ is the Lebesgue measure on $\mathbb{R}$.

In summary, our model contains the parameters

$$\Theta = \theta \cup \{\rho_0, \chi, k_\tau, c_\omega\} \cup \{\alpha_\iota, \beta_\iota : \iota = \delta, \sigma, \omega, \kappa\}. \tag{4.6}$$

The parameters in θ vary from one fingerprint or fingermark to the next according to the distribution (4.5). The remaining parameters in $\Theta \setminus \theta$ are viewed as known constants which are common to all fingerprints and fingermarks. In practice, their values may be specified by scientific expertise, or alternatively they may be estimated using an appropriate database of fingerprints and fingermarks. The method for determining these constants will affect the way in which the likelihood ratio should be interpreted. We discuss our estimation approach for these fixed parameters in §6.2.

§4.1.1 Invariance under similarity transformations

It would be intuitively pleasing if our likelihoods were invariant under similarity transformations. This can be achieved with a specific choice of prior distributions

for the parameters $\sigma_A, \sigma_B, \tau_A, \tau_B$, and ψ. For our likelihood to be invariant under scale transformations, we require

$$p(A, B, \xi \mid \theta, H_p) p(\lambda_A \tau_A, \lambda_B \tau_B, \lambda_A \sigma_A, \lambda_B \sigma_B) \, \mathrm{d}(\lambda_A \tau_A, \lambda_B \tau_B, \lambda_A \sigma_A, \lambda_B \sigma_B) \tag{4.7}$$

to be independent of the values of $\lambda_A, \lambda_B > 0$. For the likelihood to be invariant under translation and rotation as well, we must have $p(\tau_A, \tau_B, \sigma_A, \sigma_B, \psi) \propto \sigma_A^{-3} \sigma_B^{-3}$. This density is improper; it corresponds to the limit of (4.5) as $k_\tau, \alpha_\sigma, \beta_\sigma \to 0$, after multiplying (4.5) by the divergent term $\{\Gamma(\alpha_\sigma)/(k_\tau \beta_\sigma^{\alpha_\sigma})\}^2$.

Normally such a prior may result in a meaningless likelihood ratio. However, in our case the improper prior is common to both models H_d and H_p under consideration and the marginal likelihood ratio is equal to the limit described above. Nevertheless, in order to avoid other potential difficulties with using improper prior distributions, we do not pursue this invariance approach in this dissertation. The interested reader is directed to Forbes et al. (2014), where the invariance approach is considered in detail.

§4.2 Integrating the likelihood under H_d

Under H_d we can analytically integrate $p(A, B \mid \theta, H_d) p(\theta)$, the product of (2.36) and (4.5), over θ. First, we note that

$$\int_{\mathbb{C}^2} p(\tau_A \mid \sigma_A) \prod_{m_a \in A} \varphi(r_a; \tau_A, \sigma_A^2) \, \mathrm{d}\tau_A = \frac{k_\tau \pi^{-n_A}}{n_A + k_\tau} \sigma_A^{-2n_A} \exp\left\{ \frac{|R_{A\setminus\xi}|^2}{(n_A + k_\tau)\sigma_A^2} - \frac{S_{A\setminus\xi}}{\sigma_A^2} \right\}. \tag{4.8}$$

The integral over τ_B is analogous. Second, we integrate over δ_A, yielding

$$\int_0^1 e^{-\rho_0 \delta_A} \delta_A^{\alpha_\delta + n_A - 1} (1 - \delta_A)^{\beta_\delta - 1} \, \mathrm{d}\delta_A = e^{-\rho_0} \frac{\Gamma(\alpha_\delta + n_A)\Gamma(\beta_\delta)}{\Gamma(\alpha_\delta + \beta_\delta + n_A)} {}_1F_1(\beta_\delta, \alpha_\delta + \beta_\delta + n_A, \rho_0), \tag{4.9}$$

where ${}_1F_1$ is the confluent hypergeometric function (Olver et al., 2010, §13). Similarly, for δ_B, we have

$$\int_0^1 e^{-\rho_0 \delta_B} \delta_B^{n_B} \, \mathrm{d}\delta_B = e^{-\rho_0} \frac{1}{n_B + 1} {}_1F_1(1, n_B + 2, \rho_0). \tag{4.10}$$

Finally, for σ_A we have

$$\int_0^\infty \sigma_A^{-2(\alpha_\sigma+n_A-1)} \exp\left[-\sigma_A^{-2}\left\{S_{A\setminus\xi}+\beta_\sigma-|R_{A\setminus\xi}|^2/(n_A+k_\tau)\right\}\right]\times 2\sigma_A^{-3}\,\mathrm{d}\sigma_A$$
$$= \Gamma(\alpha_\sigma+n_A)\left\{S_{A\setminus\xi}+\beta_\sigma-|R_{A\setminus\xi}|^2/(n_A+k_\tau)\right\}^{-\alpha_\sigma-n_A}. \quad (4.11)$$

The integral over σ_B^{-2} is analogous. Combining these we have

$$p(A,B\,|\,H_d) = \tilde{c}(A)\tilde{c}(B)e^{-2\rho_0}(\rho_0/\pi)^{n_A+n_B}\chi^{n_A^{(1)}+n_B^{(1)}}(1-\chi)^{n_A^{(-1)}+n_B^{(-1)}}\frac{k_\tau^2\beta_\sigma^{2\alpha_\sigma}}{\Gamma(\alpha_\sigma)^2}$$
$$\times\frac{\Gamma(\alpha_\delta+\beta_\delta)\Gamma(\alpha_\delta+n_A)\,{}_1F_1(\beta_\delta,\alpha_\delta+\beta_\delta+n_A,\rho_0)\,{}_1F_1(1,n_B+2,\rho_0)\Gamma(\alpha_\sigma+n_A)\Gamma(\alpha_\sigma+n_B)}{\Gamma(\alpha_\delta)\Gamma(\alpha_\delta+\beta_\delta+n_A)(n_A+k_\tau)(n_B+k_\tau)(n_B+1)}$$
$$\times\left\{S_{A\setminus\xi}+\beta_\sigma-|R_{A\setminus\xi}|^2/(n_A+k_\tau)\right\}^{-\alpha_\sigma-n_A}\left\{S_{B\setminus\xi}+\beta_\sigma-|R_{B\setminus\xi}|^2/(n_B+k_\tau)\right\}^{-\alpha_\sigma-n_B}. \quad (4.12)$$

§4.3 Approximating the marginal likelihood under H_p

The form of $p(A,B\,|\,H_d)$ was readily computed in (4.12), but we cannot analytically obtain $p(A,B\,|\,H_p)$ because the required sums and integrals are intractable. For example, for $n_A = n_B = 100$, the number of possible values in the sum for ξ is approximately 10^{165}. We detail several methods of estimating the sums and integrals here.[†] In all cases, we rewrite $p(A,B\,|\,H_p)$ in terms of an expectation with respect to the distribution $p(\theta,\xi\,|\,A,B,H_p)$, the product of (2.42) and (4.5), and approximate the expectation with its sample average

$$\hat{\mathbb{E}}_S\{f(\theta,\xi)\} = |S|^{-1}\sum_{i=1}^{|S|} f(\theta_i,\xi_i), \quad (4.13)$$

where the sample S is obtained from the MCMC sampler described in chapter 5. We shall compare the performance of the various estimation methods in chapter 6.

Several of the methods we describe require a normalized distribution with density $q(\theta,\xi)$ with respect to the measure $\mu_\theta \times \mu_{\Xi(A,B)}$. We describe such a distribution in

[†]Most of the following methods provide estimators of the marginal likelihood $p(A,B\,|\,H_p)$. Since $p(A,B\,|\,H_d)$ is analytically available through (4.12), the resulting estimates can be immediately translated to estimates of the likelihood ratio Λ. Thus in the rest of this dissertation we shall sometimes refer to the following estimators as estimators for the (log) likelihood ratio.

§4.3.5. The accuracy of the methods depends on how well $q(\theta, \xi)$ approximates the posterior $p(\theta, \xi \,|\, A, B, H_p)$.

§4.3.1 Harmonic means estimate

By far the simplest method is the harmonic means estimate (Newton and Raftery, 1994; Raftery et al., 2007), which uses the identity

$$p(A, B \,|\, H_p) = 1 / \mathbb{E}_{\theta,\xi} \left\{ \frac{q(\theta, \xi)}{p(A, B, \theta, \xi \,|\, H_p)} \,\middle|\, A, B, H_p \right\} \tag{4.14}$$

for any proper normalized density q, where the conditional expectation notation denotes the expectation with respect to $p(\theta, \xi \,|\, A, B, H_p)$. The *naive* harmonic means estimate chooses q to be the prior distribution of the parameters. In our case,

$$\begin{aligned} q_{\text{naive}}(\theta, \xi) &= p(\theta) p(\xi \,|\, \theta, n_A, n_B) \\ &= \frac{p(\theta) p(n_\xi \,|\, \theta)}{|\Xi(A, B, n_\xi)|} = p(\theta) \exp(-\rho_0 \delta_A \delta_B) (\rho_0 \delta_A \delta_B)^{n_\xi} \frac{(n_A - n_\xi)!(n_B - n_\xi)!}{n_A! n_B!}. \end{aligned} \tag{4.15}$$

We use the prior of ξ after conditioning on n_A and n_B because it is uniform over $\Xi(A, B)$, whereas the distribution of ξ marginal of n_A and n_B is much harder to define.

While this simple estimator is appealing, it often has infinite variance (see, for example, the response to Raftery et al. (2007) by Robert and Chopin) and therefore offers poor performance. A better approach sets $q(\theta, \xi)$ to some normalized density which approximates the posterior $p(\theta, \xi \,|\, A, B, H_p)$. This yields the estimate

$$\widehat{p}(A, B \,|\, H_p) = 1 / \hat{\mathbb{E}}_S \{ q(\theta, \xi) / p(A, B, \theta, \xi \,|\, H_p) \} = |S| \left\{ \sum_{i=1}^{|S|} \frac{q(\theta_i, \xi_i)}{p(A, B, \theta_i, \xi_i \,|\, H_p)} \right\}^{-1}, \tag{4.16}$$

where S is a sample from $p(\theta, \xi \,|\, A, B, H_p)$.

§4.3.2 Chib's estimate

To use Chib's method (Chib, 1995; Chib and Jeliazkov, 2001), we notice that

$$p(A, B \,|\, H_p) = \frac{p(A, B, \theta^*, \xi^* \,|\, H_p)}{p(\theta^*, \xi^* \,|\, A, B, H_p)} \tag{4.17}$$

for any fixed values of θ^* of θ and ξ^* of ξ. The numerator of (4.17) can be computed exactly, while the denominator can be rewritten as

$$\begin{aligned} p(\theta^*, \xi^* | A, B, H_p) = {} & p(\delta_A^* | A, B, H_p) \\ & \times p(\delta_B^* \,|\, \delta_A^*, A, B, H_p) \\ & \times p(\tau_A^*, \tau_B^* \,|\, \delta_A^*, \delta_B^*, A, B, H_p) \\ & \times p(\psi^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, A, B, H_p) \\ & \times p(\sigma_Q^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, A, B, H_p) \\ & \times p(\sigma_P^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_Q^*, A, B, H_p) \\ & \times p(\tilde{\omega}^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_A^*, \sigma_B^*, A, B, H_p) \\ & \times p(\kappa^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_A^*, \sigma_B^*, \tilde{\omega}^*, A, B, H_p) \\ & \times p(\xi^* \,|\, \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_A^*, \sigma_B^*, \tilde{\omega}^*, \kappa^*, A, B, H_p), \end{aligned} \tag{4.18}$$

where $\sigma_P = 1/(\sigma_A \sigma_B)$ and $\sigma_Q = \sigma_B/\sigma_A$ (the reason for this parametrization will become clear below). We shall approximate each term using a sample expectation of the relevant normalized full conditional. For these approximations to be accurate, we require that (θ^*, ξ^*) has high posterior probability.

Our method for choosing (θ^*, ξ^*) and performing the approximations is detailed in algorithm 4.1. In the rest of this section we will investigate how to quickly evaluate the (approximate) normalized full conditional densities in (4.18).

In the case of $\delta_A, \delta_B, (\tau_A, \tau_B)$, and ψ, the full conditionals are all standard distributions (beta, bivariate complex normal, and von Mises respectively). The normalization constants for these distributions are known.

The full conditionals for σ_Q and σ_P are generalized inverse Gaussian and gamma respectively, both of which have known normalization constants. This is the reason we used this particular parametrization: the full conditionals of σ_A and σ_B are not standard distributions. The change of variables and the normalization constants are described in §B.2.

In the case of κ and $\tilde{\omega}$, the distributions are non-standard and the normalization

Algorithm 4.1 Chib's method to approximate the marginal posterior $p(A, B \mid H_p)$.

Let the sample sizes be $N = 1,000$ and $N_\xi = 200$.
All samples S_i are generated by holding the starred variables constant while
sampling the non-starred variables as described in algorithm 5.1.
$r \leftarrow 1$
Generate sample S_1 of size N from $\delta_A, \delta_B, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi \mid A, B, H_p$
$\delta_A^* \leftarrow \hat{\mathbb{E}}_{S_1}(\delta_A)$
Generate sample S_2 of size N from $\delta_B, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi \mid \delta_A^*, A, B, H_p$
$\delta_B^* \leftarrow \hat{\mathbb{E}}_{S_2}(\delta_B)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_2}\{p(\delta_A^* \mid \delta_B, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$
Generate sample S_3 of size N from $\tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi \mid \delta_A^*, \delta_B^*, A, B, H_p$
$(\tau_A^*, \tau_B^*) \leftarrow \hat{\mathbb{E}}_{S_3}(\tau_A, \tau_B)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_3}\{p(\delta_B^* \mid \delta_A^*, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$
Generate sample S_4 of size N from $\sigma_A, \sigma_B, \psi, \tilde{\omega}, \kappa, \xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, A, B, H_p$
$\psi^* \leftarrow \hat{\mathbb{E}}_{S_4}(\psi)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_4}\{p(\tau_A^*, \tau_B^* \mid \delta_A^*, \delta_B^*, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$
Generate sample S_5 of size N from $\sigma_A, \sigma_B, \tilde{\omega}, \kappa, \xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, A, B, H_p$
$\sigma_Q^* \leftarrow \hat{\mathbb{E}}_{S_5}(\sigma_Q)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_5}\{p(\psi^* \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \sigma_P, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$
Generate sample S_6 of size N from $\sigma_P, \tilde{\omega}, \kappa, \xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_Q^*, A, B, H_p$
$\sigma_P^* \leftarrow \hat{\mathbb{E}}_{S_6}(\sigma_P)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_6}\{p(\sigma_Q^* \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_P, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$ # See §B.2
Generate sample S_7 of size N from $\tilde{\omega}, \kappa, \xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_A^*, \sigma_B^*, A, B, H_p$
$\tilde{\omega}^* \leftarrow \hat{\mathbb{E}}_{S_7}(\tilde{\omega})$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_7}\{p(\sigma_P^* \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \psi^*, \sigma_Q^*, \tilde{\omega}, \kappa, \xi, A, B, H_p)\}$ # See §B.2
Generate sample S_8 of size N from $\kappa, \xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \sigma_A^*, \sigma_B^*, \psi^*, \tilde{\omega}^*, A, B, H_p$
$\kappa^* \leftarrow \hat{\mathbb{E}}_{S_8}(\kappa)$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_8}\{p(\tilde{\omega}^* \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \sigma_A^*, \sigma_B^*, \psi^*, \kappa, \xi, A, B, H_p)\}$ # See §B.4
Generate sample S_9 of size N from $\xi \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \sigma_A^*, \sigma_B^*, \psi^*, \tilde{\omega}^*, \kappa^*, A, B, H_p$
$r \leftarrow r \times \hat{\mathbb{E}}_{S_9}\{p(\kappa^* \mid \delta_A^*, \delta_B^*, \tau_A^*, \tau_B^*, \sigma_A^*, \sigma_B^*, \psi^*, \tilde{\omega}^*, \xi, A, B, H_p)\}$ # See §B.3
$\xi^* \leftarrow \text{argmax}_\xi\{p(\theta^*, \xi, A, B, H_p)\}$ # See §3.2.1
for $\beta \in B$ **do**
    Generate sample S_β of size N_ξ from $\xi_{>\beta} \mid \theta^*, \xi_{\leq\beta}^*, A, B, H_p$ # See (4.19)
    $r \leftarrow r \times \hat{\mathbb{E}}_{S_\beta}\{p(\xi_\beta^* \mid \xi_{<\beta}^*, \xi_{>\beta}, \theta^*, A, B, H_p)\}$ # See (4.22)
end for
return r, an estimate of $p(A, B \mid H_p)$

constants are intractable. We approximate these normalization constants as detailed in §B.3 and §B.4 respectively.

Finally, we must approximate the final term $p(\xi \mid \theta, A, B, H_p)$. First we need some more notation. Given $\beta \in B$ and $\xi \in \Xi(A,B)$, we define the sub-matches by

$$\begin{aligned} \xi_\beta &= \{\langle a, \beta\rangle \in \xi\}, \quad \xi_{\leq\beta} = \{\langle a, b\rangle \in \xi : b \in B, b \leq \beta\}, \\ \xi_{<\beta} &= \{\langle a, b\rangle \in \xi : b \in B, b < \beta\}, \quad \xi_{>\beta} = \{\langle a, b\rangle \in \xi : b \in B, b > \beta\}, \end{aligned} \tag{4.19}$$

with respect to some arbitrary total ordering on B. Next, let ϕ be a fixed but arbitrary element of $\mathbb{M} \setminus (A \cup B)$ which serves to denote an unobserved minutia. Given $b \in \mathbb{M}$, define $\Pi_{A,b}(\xi) : \Xi(A,B) \to \mathbb{M}$ by[†]

$$\Pi_{A,b}(\xi) = \begin{cases} \phi & \text{if } \{\langle a, b\rangle \in \xi : a \in A\} = \emptyset, \\ a \in A : \langle a, b\rangle \in \xi & \text{otherwise.} \end{cases} \tag{4.20}$$

That is, $\Pi_{A,b}(\xi)$ returns the minutia $a \in A$ which is matched to b, if such an a exists, and otherwise returns ϕ. With this notation, we can write

$$\begin{aligned} p(\xi \mid \theta, A, B, H_p) &= \prod_{\beta \in B} p(\xi_\beta \mid \xi_{<\beta}, \theta, A, B, H_p) \\ &= \prod_{\beta \in B} \mathbb{E}_{\xi_{>\beta}} \left\{ p(\xi_\beta \mid \xi_{<\beta}, \xi_{>\beta}, \theta, A, B, H_p) \,\middle|\, \xi_{\leq\beta}, \theta, A, B, H_p \right\}. \end{aligned} \tag{4.21}$$

We see from (2.42) that

$$p\left(\xi_\beta \mid \xi_{<\beta}, \xi_{>\beta}, \theta, A, B, H_p\right) \propto \exp\left[w\{\Pi_{A,\beta}(\xi), \beta \mid \theta\}\right] \mathbb{1}\{\Pi_{A,\beta}(\xi) \notin \Pi_A(\xi_{<\beta} \cup \xi_{>\beta})\}, \tag{4.22}$$

where w is given in (3.7). The indicator function simply says that β cannot be matched to any $a \in A$ which is already matched in $\xi_{<\beta}$ or $\xi_{>\beta}$. The normalization constant of (4.22) is obtained by summing over the support, which is $\xi_\beta = \emptyset$ and $\xi_\beta = \{\langle a, \beta\rangle\}$ for each $a \in A$.

Thus we can evaluate and normalize (4.22), and therefore we can approximate (4.21) by approximating each expectation with a sample average.

[†] $\Pi_{A,b}(\xi)$ is well-defined because ξ is the edge set of a bipartite graph with maximum degree one, and hence each vertex b is adjacent to at most one edge in ξ.

§4.3.3 Bridge sampling estimate

Bridge sampling (Meng and Wong, 1996) uses the basic identity

$$\frac{p(A,B\,|\,H_p)}{\sum_{\xi\in\Xi(A,B)}\int q(\theta,\xi)\,\mathrm{d}\mu_\theta(\theta)} = \frac{\mathbb{E}_{\theta,\xi\sim q}\{\lambda_{BS}(\theta,\xi)p(A,B,\theta,\xi\,|\,H_p)\}}{\mathbb{E}_{\theta,\xi}\{\lambda_{BS}(\theta,\xi)q(\theta,\xi)\,|\,A,B,H_p\}}, \tag{4.23}$$

where $\mathbb{E}_{\theta,\xi\sim q}$ denotes the expectation of (θ,ξ) with respect to the density $q(\theta,\xi)$, in contrast to the denominator, which is the expectation with respect to $p(\theta,\xi\,|A,B,H_p)$. This identity holds for any function $\lambda_{BS}(\theta,\xi)$ and any proper (potentially unnormalized) density $q(\theta,\xi)$ such that the expectations are finite. The distribution with density q is called the *bridge*. When q is normalized, (4.23) can be used to estimate $p(A,B\,|\,H_p)$:

$$\widehat{p}(A,B\,|\,H_p) = \frac{\hat{\mathbb{E}}_{S_0}[\lambda_{BS}(\theta,\xi)p(A,B,\theta,\xi\,|\,H_p)]}{\hat{\mathbb{E}}_{S}[\lambda_{BS}(\theta,\xi)q(\theta,\xi)]}, \tag{4.24}$$

where S_0 is a sample from $q(\theta,\xi)$ and S is a sample from $p(\theta,\xi\,|A,B,H_p)$. Meng and Wong (1996) show that

$$\lambda_{BS}(\theta,\xi)^{-1} = |S|p(A,B,\theta,\xi\,|\,H_p) + |S_0|q(\theta,\xi)p(A,B\,|\,H_p) \tag{4.25}$$

minimizes the mean-squared error of $\log\widehat{p}(A,B\,|\,H_p)$ for a fixed bridge q. This optimal choice for λ_{BS} depends on the unknown quantity $p(A,B\,|\,H_p)$, so they recommend finding it iteratively via $\lambda_{BS}^{(0)}(\theta,\xi)=1$ and

$$\begin{aligned}\widehat{p}^{(t+1)}(A,B\,|\,H_p) &= \hat{\mathbb{E}}_{S_0}\{\lambda_{BS}^{(t)}(\theta,\xi)p(A,B,\theta,\xi\,|\,H_p)\}/\hat{\mathbb{E}}_{S}\{\lambda_{BS}^{(t)}(\theta,\xi)q(\theta,\xi)\},\\ \lambda_{BS}^{(t+1)}(\theta,\xi)^{-1} &= |S|p(A,B,\theta,\xi\,|\,H_p) + |S_0|q(\theta,\xi)\widehat{p}^{(t+1)}(A,B\,|\,H_p).\end{aligned} \tag{4.26}$$

In practice this procedure converges in fewer than ten iterations.

Mira and Nicholls (2004) show that the bridge sampling estimate reduces to Chib's estimate for a specific choice of bridge and λ_{BS}. Since the bridge sampling estimate optimizes over λ_{BS}, we expect that the bridge sampling estimate corresponding to the Chib bridge will outperform Chib's estimate. However, for ease of implementation, we use the simple bridge described in §4.3.5 rather than the Chib bridge for our bridge sampling estimate. Therefore it is possible that our bridge sampling estimate will be less accurate than our Chib's estimate.

§4.3.4 Reversible jump estimate

Suppose the hypothesis $H \in \{H_d, H_p\}$ is a binary random variable with $p(H_d) = p_0 \in (0,1)$. Furthermore, suppose that the joint density of θ, ξ, H is given by

$$p(\theta, \xi, H \mid A, B) = \frac{1}{p(A,B)} \times \begin{cases} (1-p_0)p(\theta)p(A,B,\xi \mid \theta, H_p) & \text{if } H = H_p, \\ p_0 p(A,B \mid H_d) q(\theta, \xi) & \text{if } H = H_d, \end{cases} \tag{4.27}$$

where q is some normalized probability density for (θ, ξ). Let S_{RJ} be a sample from (4.27). Then, for any value of p_0, we can obtain an estimator for Λ by using S_{RJ} to approximate both $p(H_p \mid A,B)$ and $p(H_d \mid A,B)$, taking the ratio, and correcting for the prior odds $(1-p_0)/p_0$:

$$\widehat{\Lambda}_{RJ1} = \frac{p_0}{1-p_0} \frac{\hat{\mathbb{E}}_{S_{RJ}}\{\mathbb{1}(H = H_p)\}}{\hat{\mathbb{E}}_{S_{RJ}}\{\mathbb{1}(H = H_d)\}}. \tag{4.28}$$

This estimation procedure contrasts with the previous methods, which approximated $p(A,B \mid H_p)$ and divided by the known exact value of $p(A,B \mid H_d)$. Here we must do two approximations, because $p(A,B)$ is intractable and we cannot convert from the known $p(A,B \mid H_d)$ to the desired $p(H_d \mid A,B)$.

We generate S_{RJ} by initializing H to H_p and repeatedly performing the following two steps, which define a Markov chain $\mathscr{M}_{RJ}$:

1. If $H = H_p$, then sample (θ, ξ) using the Metropolis-within-Gibbs sampler as described in algorithm 5.1. Otherwise, generate a sample (θ, ξ) from q.
2. Sample $H \in \{H_d, H_p\}$ by noticing $p(H \mid \theta, \xi, A, B)$ has a Bernoulli distribution with the probability of H_p equal to

$$\frac{p(A,B,\theta,\xi \mid H_p)(1-p_0)}{p(A,B,\theta,\xi \mid H_d)p_0 + p(A,B,\theta,\xi \mid H_p)(1-p_0)}. \tag{4.29}$$

The distribution q can be viewed equally well as a pseudoprior for a Carlin–Chib-type algorithm (Carlin and Chib, 1995), or as a proposal distribution for a reversible-jump algorithm (Green, 1995). Usually, these algorithms are used to sample between two or more models, each of which typically has several parameters. However, in our case, the model under H_d has zero parameters: they were all analytically integrated out in

§4.2. Thus, $\mathscr{M}_{RJ}$ can be viewed as a *regenerative stochastic process* (Smith, 1955), whose regenerative points are the states with $H = H_d$. In effect, a sample from $\mathscr{M}_{RJ}$ can be viewed as a sequence of samples from i.i.d. sub-chains, where each sub-chain is randomly terminated by visiting the state H_d.

As (4.28) holds for any value of p_0, we can use p_0 as a tuning parameter to control the expected length of each sub-chain in $\mathscr{M}_{RJ}$. We use a pilot run to find p_0 such that the marginal posterior probability of H_p is approximately $1/2$; the details are given in §6.3. We have not investigated how different values of p_0 affect the performance of the estimation procedure.

An alternative estimator for Λ can be derived by analytically performing the expectation over H in (4.28). Letting

$$\ell(A,B,\theta,\xi) = \frac{p(A,B,\theta,\xi\,|\,H_p)}{p(A,B,\theta,\xi\,|\,H_d)} = \frac{p_0}{1-p_0}\frac{p(A,B,\theta,\xi,H_p)}{p(A,B,\theta,\xi,H_d)} \tag{4.30}$$

be the likelihood ratio for fixed θ and ξ, we notice from (4.27) that

$$p(H_d\,|\,\theta,\xi,A,B) = \frac{p(A,B,\theta,\xi,H_d)}{p(A,B,\theta,\xi,H_d)+p(A,B,\theta,\xi,H_p)} = \frac{p_0}{p_0+(1-p_0)\ell(A,B,\theta,\xi)}. \tag{4.31}$$

Similarly, we have $p(H_p\,|\,\theta,\xi,A,B) = (1-p_0)/\{1-p_0+p_0/\ell(A,B,\theta,\xi)\}$. Combining the above two equations yields

$$\begin{aligned}\Lambda &= \frac{p_0}{1-p_0}\frac{\mathbb{E}_{\theta,\xi,H}\{\mathbb{1}(H=H_p)\}}{\mathbb{E}_{\theta,\xi,H}\{\mathbb{1}(H=H_d)\}} = \frac{p_0}{1-p_0}\frac{\mathbb{E}_{\theta,\xi}\left\{p(H_p\,|\,\theta,\xi,A,B)\,|\,A,B\right\}}{\mathbb{E}_{\theta,\xi}\{p(H_d\,|\,\theta,\xi,A,B)\,|\,A,B\}} \\ &= \frac{\mathbb{E}_{\theta,\xi}\left[\{1-p_0+p_0/\ell(A,B,\theta,\xi)\}^{-1}\,|\,A,B\right]}{\mathbb{E}_{\theta,\xi}\left[\{p_0+(1-p_0)\ell(A,B,\theta,\xi)\}^{-1}\,|\,A,B\right]}.\end{aligned} \tag{4.32}$$

The corresponding estimator for Λ,

$$\widehat{\Lambda}_{RJ2} = \frac{\hat{\mathbb{E}}_{S_{RJ}}\left[\{1-p_0+p_0/\ell(A,B,\theta,\xi)\}^{-1}\right]}{\hat{\mathbb{E}}_{S_{RJ}}\left[\{p_0+(1-p_0)\ell(A,B,\theta,\xi)\}^{-1}\right]}, \tag{4.33}$$

uses more of the information in the sample S_{RJ} than $\widehat{\Lambda}_{RJ1}$. After performing numerical tests, we believe that $\widehat{\Lambda}_{RJ2}$ provides a more accurate estimate of the likelihood ratio than $\widehat{\Lambda}_{RJ1}$, and hence we use $\widehat{\Lambda}_{RJ2}$ as our representative reversible jump estimate in the rest of this dissertation.

Due to the nature of the Markov chain $\mathscr{M}_{RJ}$ which we use to generate the sample S_{RJ}, we frequently jump from H_d to H_p via the proposal distribution q. Hence the accuracy of our reversible jump estimate depends on how closely q approximates the posterior distribution $p(\theta, \xi \,|\, A, B, H_p)$. We can improve the accuracy by tuning the proposal distribution, as described in §4.3.5. In practice, the tuned proposal tends to be concentrated about a single mode, while the posterior distribution $p(\theta, \xi \,|\, A, B, H_p)$ tends to be highly multimodal. This causes the Markov chain to frequently bounce between H_d and a single mode of H_p, but it tends not to stay in H_p long enough to explore the remainder of the parameter space. We ameliorate this by decreasing p_0 and by reducing the frequency with which we propose model jumps, but it remains a significant issue that impacts the accuracy of our reversible jump estimate.

§4.3.5 Form of $q(\theta, \xi)$

The harmonic means estimate, bridge sampling estimate, and reversible jump estimate are only accurate if the normalized proposal distribution $q(\theta, \xi)$ is close to the posterior $p(\theta, \xi \,|\, A, B, H_p)$. We describe our choice of $q(\theta, \xi)$ here.

For simplicity we consider a $q(\theta)$ which factorizes as

$$q(\theta) = q(\delta_A)q(\delta_B)q(\tau_A, \tau_B)q(\sigma_A)q(\sigma_B)q(\psi)q(\tilde{\omega})q(\kappa). \tag{4.34}$$

We use beta distributions for δ_A and δ_B, a bivariate complex normal distribution for (τ_A, τ_B), gamma distributions for σ_A^{-2} and σ_B^{-2}, a von Mises distribution for ψ, a gamma distribution left-truncated at c_ω for $\tilde{\omega}$, and a gamma distribution for κ. The parameters of these distributions are obtained by maximum likelihood estimation using a pilot sample from $p(\theta, \xi \,|\, A, B)$. In the cases of $\delta_A, \delta_B, \sigma_A, \sigma_B, \tilde{\omega}$, and κ, the maximum likelihood estimates are found numerically.

It remains to define a normalized density $q(\xi \mid \theta)$ which approximates the posterior

distribution of ξ. From (4.22) we have

$$p(\xi \mid \theta, A, B, H_p) = \prod_{\beta \in B} \mathbb{E}_{\xi_{>\beta}} \left(\frac{\exp\left[w\{\Pi_{A,\beta}(\xi), \beta \mid \theta\} \right] \mathbb{1}\{\Pi_{A,\beta}(\xi) \notin \Pi_A(\xi_{<\beta} \cup \xi_{>\beta})\}}{\sum_{a \in A \cup \phi} \exp\{w(a, \beta \mid \theta)\} \, \mathbb{1}\{a \notin \Pi_A(\xi_{<\beta} \cup \xi_{>\beta})\}} \,\middle|\, \xi_{\leq\beta}, \theta, A, B, H_p \right). \quad (4.35)$$

The expectation is difficult to compute. However, if we drop $\xi_{>\beta}$ from the indicators in both the numerator and the denominator, the term inside the expectation becomes independent of $\xi_{>\beta}$ and we are left with

$$q(\xi \mid \theta) = \prod_{\beta \in B} q_\beta(\xi_\beta \mid \xi_{<\beta}), \quad (4.36)$$

where the individual terms for each $\beta \in B$ are

$$q_\beta(\xi_\beta \mid \xi_{<\beta}) = \frac{\exp\left[w\{\Pi_{A,\beta}(\xi), \beta \mid \theta\} \right] \mathbb{1}\left\{\Pi_{A,\beta}(\xi) \notin \Pi_A\left(\xi_{<\beta}\right)\right\}}{\sum_{a \in A \cup \phi} \exp\{w(a, \beta \mid \theta)\} \, \mathbb{1}\left\{a \notin \Pi_A\left(\xi_{<\beta}\right)\right\}}. \quad (4.37)$$

Each term q_β is a normalized distribution function for which minutia $a \in A \cup \phi$ is matched to β, and hence $q(\xi \mid \theta)$ is a normalized distribution function over $\Xi(A, B)$ with respect to $\mu_{\Xi(A,B)}$. The product form of $q(\xi \mid \theta)$ allows it to be quickly evaluated, and furthermore we can easily generate samples from it by sequentially sampling the discrete random variables $q_\beta(\xi_\beta \mid \xi_{<\beta})$ for $\beta \in B$. This makes (4.36) a convenient choice for our proposal function $q(\xi \mid \theta)$.

Chapter Five

Sampling the posterior distribution

The estimators of the likelihood ratio developed in chapter 4 require samples from the distribution $p(\theta, \xi \,|\, A, B, H_p)$. In this chapter we describe a Metropolis-within-Gibbs sampler that generates such samples. Our method is described in pseudocode in algorithm 5.1. To ease the computation of the necessary full conditional distributions, we explicitly compute the product of (2.42) and (4.5) to find the density under H_p:

$$
\begin{aligned}
p(A,B,\theta,\xi \,|\, H_p) &= \tilde{c}(A)\tilde{c}(B)\frac{4k_\tau^2\beta_\sigma^{2\alpha_\sigma}\lambda_\kappa(\alpha_\kappa,\beta_\kappa)\Gamma(\alpha_\delta+\beta_\delta)\beta_\omega^{\alpha_\omega}}{\pi^2\Gamma(\alpha_\delta)\Gamma(\beta_\delta)\Gamma(\alpha_\sigma)^2\Gamma(\alpha_\omega,\beta_\omega c_\omega)}\pi^{-n_A-n_B}\rho_0^{n_A+n_B-n_\xi} \\
&\times \delta_A^{\alpha_\delta+n_A-1}\delta_B^{n_B}(1-\delta_B)^{n_A-n_\xi}(1-\delta_A)^{\beta_\delta+n_B-n_\xi-1}\chi^{n_A^{(1)}+n_B^{(1)}-n_\xi^{(1)}}(1-\chi)^{n_A^{(-1)}+n_B^{(-1)}-n_\xi^{(-1)}} \\
&\times \sigma_A^{-2(\alpha_\sigma+n_A)-3}\sigma_B^{-2(\alpha_\sigma+n_B)-3}\tilde{\omega}^{\alpha_\omega+n_\xi-1}I_0(\kappa)^{2(\alpha_\kappa-1)-n_\xi}\exp\{-\rho_0(\delta_A+\delta_B-\delta_A\delta_B)\} \\
&\times \exp\left\{-(k_\tau|\tau_A|^2+\beta_\sigma)\sigma_A^{-2}-(k_\tau|\tau_B|^2+\beta_\sigma)\sigma_B^{-2}-\beta_\omega\tilde{\omega}+(2\alpha_\kappa-2-\beta_\kappa)\kappa\right\} \\
&\times \exp\left\{-\frac{S_{A\setminus\xi}-2\,\mathrm{Re}(\overline{R_{A\setminus\xi}}\tau_A)+(n_A-n_\xi)|\tau_A|^2}{\sigma_A^2}-\frac{S_{B\setminus\xi}-2\,\mathrm{Re}(\overline{R_{B\setminus\xi}}\tau_B)+(n_B-n_\xi)|\tau_B|^2}{\sigma_B^2}\right\} \\
&\times \exp\left[2\sqrt{\tilde{\omega}^2-\tilde{\omega}}\frac{\mathrm{Re}\left\{\psi\left(S_{\xi AB}-\overline{R_{\xi A}}\tau_B-R_{\xi B}\overline{\tau_A}+n_\xi\overline{\tau_A}\tau_B\right)\right\}}{\sigma_A\sigma_B}+\kappa\,\mathrm{Re}(\psi S_\xi)\right] \\
&\times \exp\left[-\tilde{\omega}\left\{\frac{S_{\xi A}-2\,\mathrm{Re}(\overline{R_{\xi A}}\tau_A)+n_\xi|\tau_A|^2}{\sigma_A^2}+\frac{S_{\xi B}-2\,\mathrm{Re}(\overline{R_{\xi B}}\tau_B)+n_\xi|\tau_B|^2}{\sigma_B^2}\right\}\right].
\end{aligned}
\tag{5.1}
$$

The various R and S terms used in the above equation are defined in (3.5).

To generate our samples, we use Marsaglia and Tsang (2000b) for normally distributed variables, Marsaglia and Tsang (2000a) for gamma distributed variables,

Sections 5.1 – 5.4 and 5.7 are based off Forbes et al. (2014).

Algorithm 5.1 Metropolis-within-Gibbs sampler for the posterior of θ and ξ under H_p

Set $\theta^{(0)}, \xi^{(0)}$ set to some initial value.
for $n = 1, \ldots, N$ **do**
$n_0 \leftarrow (n-1)$
$(\delta_A^{(n)}, \delta_B^{(n)}) \leftarrow \text{Sample}(\delta_A, \delta_B \,|\, A, B, \tau_A^{(n_0)}, \tau_B^{(n_0)}, \sigma_A^{(n_0)}, \sigma_B^{(n_0)}, \psi^{(n_0)}, \tilde{\omega}^{(n_0)}, \kappa^{(n_0)}, \xi^{(n_0)}, H_p)$
$(\tau_A^{(n)}, \tau_B^{(n)}) \leftarrow \text{Sample}(\tau_A, \tau_B \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \sigma_A^{(n_0)}, \sigma_B^{(n_0)}, \psi^{(n_0)}, \tilde{\omega}^{(n_0)}, \kappa^{(n_0)}, \xi^{(n_0)}, H_p)$
$(\sigma_A^{(n)}, \sigma_B^{(n)}) \leftarrow \text{Sample}(\sigma_A, \sigma_B \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \psi^{(n_0)}, \tilde{\omega}^{(n_0)}, \kappa^{(n_0)}, \xi^{(n_0)}, H_p)$
$\psi^{(n)} \leftarrow \text{Sample}(\psi \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \sigma_A^{(n)}, \sigma_B^{(n)}, \tilde{\omega}^{(n_0)}, \kappa^{(n_0)}, \xi^{(n_0)}, H_p)$
$\tilde{\omega}^{(n)} \leftarrow \text{Sample}(\tilde{\omega} \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \sigma_A^{(n)}, \sigma_B^{(n)}, \psi^{(n)}, \kappa^{(n_0)}, \xi^{(n_0)}, H_p)$
$\kappa^{(n)} \leftarrow \text{Sample}(\kappa \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \sigma_A^{(n)}, \sigma_B^{(n)}, \psi^{(n)}, \tilde{\omega}^{(n)}, \xi^{(n_0)}, H_p)$
$\xi^{(n)} \leftarrow \xi^{(n_0)}$
for $j = 1, \ldots, n_A$ **do** # sample repeatedly to reduce autocorrelation
$\xi^{(n)} \leftarrow \text{Sample}(\xi \,|\, A, B, \delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \sigma_A^{(n)}, \sigma_B^{(n)}, \psi^{(n)}, \tilde{\omega}^{(n)}, \kappa^{(n)}, \xi^{(n)}, H_p)$
end for
Save desired statistics from $(\delta_A^{(n)}, \delta_B^{(n)}, \tau_A^{(n)}, \tau_B^{(n)}, \sigma_A^{(n)}, \sigma_B^{(n)}, \psi^{(n)}, \tilde{\omega}^{(n)}, \kappa^{(n)}, \xi^{(n)})$
end for

Dagpunar (1978) for truncated gamma distributed variables, Cheng (1978) for beta distributed variables, and Best and Fisher (1979) for von Mises distributed variables. For those variables whose full conditionals are not one of the above type, we give a detailed sampling algorithm below. All the sampling algorithms use Marsaglia (2003) as a source of pseudo-random numbers.

§5.1 Sampling the thinning probabilities δ_A, δ_B

Define the distribution $D(\alpha, \beta, \lambda)$ to have density

$$f_D(\delta) \propto \delta^{\alpha-1}(1-\delta)^{\beta-1}e^{-\lambda\delta} \tag{5.2}$$

for $\delta \in (0,1)$ and $\alpha > 0, \beta > 0$, and $\lambda \in \mathbb{R}$. Notice that if $\lambda = 0$ this is a beta distribution, and if $\beta = 1$ it is a gamma distribution right-truncated at one. The full conditionals of δ_A and δ_B have $D(\alpha_\delta + n_A, \beta_\delta + n_B - n_\xi, \rho_0 - \rho_0\delta_B)$ and $D(n_B, n_A - n_\xi, \rho_0 - \rho_0\delta_A)$ distributions respectively.

Let δ_0 be the mode of D, which can be easily computed via the quadratic formula applied to $\mathrm{d}\log f_D(\delta)/\mathrm{d}\delta = 0$. To sample from D, first notice that $\log(1-\delta) \approx 1 - \delta_0 - \delta/(1-\delta_0)$ for $\delta_0 \ll 1$, and hence $(1-\delta)^{\beta-1} \approx C\exp\{-(\beta-1)\delta/(1-\delta_0)\}$

where C is a constant independent of δ. Plugging this approximation into $f_D(\delta)$ yields a gamma density with shape α and rate $(\beta-1)/(1-\delta_0)+\lambda$. Thus, when $\delta_0 \leq 0.5$, we use rejection sampling with a gamma proposal right-truncated at one.

Similarly, when $\delta_0 > 0.5$, let $\tilde{\delta} = 1-\delta$ so that $\tilde{\delta} \sim D(\beta-1, \alpha-1, -\lambda)$ with mode $\tilde{\delta}_0 = 1-\delta_0$. Using the same approximation as before, we can use rejection sampling on $\tilde{\delta}$ with a gamma proposal with shape β and rate $(\alpha-1)/(1-\tilde{\delta}_0)-\lambda)$, right-truncated at one.

The full algorithm is detailed in algorithm 5.2. In practice we achieve acceptance rates greater than 0.9.

Algorithm 5.2 Rejection sampler for the thinning parameters $D(\alpha, \beta, \lambda)$

if $\lambda = 0$ **then**
  $\delta \leftarrow \text{Sample}\{\text{Gamma}(\alpha, \beta) \,|\, \delta \leq 1\}$
else
  $\delta_0 \leftarrow \text{argmax}\{x^{\alpha-1}(1-x)^{\beta-1}e^{-\lambda x} : x \in [0,1]\}$
  if $\delta_0 > 0.5$ **and** $\lambda < (\alpha-1)/(1-\delta_0)$ **then**
    $\delta \leftarrow 1 - \text{Sample}\{D(\beta, \alpha, -\lambda)\}$ # Sample from reversed distribution
  else
    repeat
      $\delta \leftarrow \text{Sample}\{\text{Gamma}\{\alpha, \lambda + (\beta-1)/(1-\delta_0)\} \,|\, \delta \leq 1\}$
      $U \leftarrow \text{Sample}\{\text{Uniform}(0,1)\}$
    until $\log(U)/(\beta-1) < \log\{(1-\delta)/(1-\delta_0)\} + (\delta-\delta_0)/(1-\delta_0)$
  end if
end if
return δ

§5.2 Sampling the translation parameters τ_A, τ_B

The full conditional distribution is bivariate complex normal with mean r_0 and inverse variance K, where

$$K = \begin{pmatrix} (n_A + k_\tau)\sigma_A^{-2} & 0 \\ 0 & (n_B + k_\tau)\sigma_B^{-2} \end{pmatrix} + n_\xi \Sigma_{AB}^{-1}, \quad r_0 = K^{-1}\left\{ \begin{pmatrix} R_{A\setminus\xi}/\sigma_A^2 \\ R_{B\setminus\xi}/\sigma_B^2 \end{pmatrix} + \Sigma_{AB}^{-1} \begin{pmatrix} R_{\xi A} \\ R_{\xi B} \end{pmatrix} \right\}. \tag{5.3}$$

The matrix Σ_{AB} is given in (2.39) and the R terms are given in (3.5).

§5.3 Sampling the scale parameters σ_A, σ_B

We make the change of variables $k_A = \sigma_A^{-2}$ and $k_B = \sigma_B^{-2}$ and note that the prior distributions of k_A and k_B have gamma distributions with shape α_σ and scale β_σ. The full conditional posterior of k_A, k_B is proportional to

$$k_A^{\alpha_\sigma+n_A} k_B^{\alpha_\sigma+n_B} \exp\{-(R_1+\beta_\sigma+k_\tau|\tau_A|^2)k_A-(R_2+\beta_\sigma+k_\tau|\tau_B|^2)k_B-2R_3\sqrt{k_Ak_B}\}, \quad (5.4)$$

where R_1, R_2, R_3 are given in (3.14).

We sample k_A and k_B using an auxiliary variable x. By defining

$$p(k_A,k_B,x) \propto k_A^{\alpha_\sigma+n_A} k_B^{\alpha_\sigma+n_B}\ \mathbb{1}(x > 2R_3\sqrt{k_Ak_B}) \\ \times \exp\left\{-(R_1+\beta_\sigma+k_\tau|\tau_A|^2)k_A-(R_2+\beta_\sigma+k_\tau|\tau_B|^2)k_B-x\right\} \quad (5.5)$$

for $x \in \mathbb{R}$, we see that (k_A, k_B), marginal of x, has our desired distribution.

We sample $k_A\,|(k_B,x)$, $k_B\,|(k_A,x)$, and $x\,|(k_A,k_B)$ as in algorithm 5.3. The auxiliary variable $x\,|\,k_A,k_B$ can be sampled by inverting the distribution function,

$$x \leftarrow 2R_3\sqrt{k_Ak_B} - \log U, \quad (5.6)$$

where U is uniform on $(0,1)$. Finally $k_A\,|\,k_B,x$ and $k_B\,|\,k_A,x$ are independent truncated gamma variables and can be sampled using a standard procedure (Dagpunar, 1978).

§5.4 Sampling the rotation parameter ψ

The full conditional of the phase ψ is von Mises with location $\nu/|\nu|$ and concentration $|\nu|$, where ν is given in (3.16).

§5.5 Sampling the location distortion parameter $\tilde{\omega}$

The full conditional of $\tilde{\omega}$ is proportional to

$$\tilde{\omega}^{\alpha_\omega+n_\xi-1} \exp\left\{-(\beta_\omega+R_4)\tilde{\omega}+2R_5\sqrt{\tilde{\omega}^2-\tilde{\omega}}\right\}\ \mathbb{1}(\tilde{\omega} \geq c_\omega), \quad (5.7)$$

Algorithm 5.3 Sampler for σ_A, σ_B

Require: Previous values $\sigma_A^{(0)}, \sigma_B^{(0)}$
  Compute R_1, R_2, R_3 from (3.14)
  if $H = H_d$ **or** $R_3 = 0$ **then**
    $k_A \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_A, R_1 + \beta_\sigma + k_\tau |\tau_A|^2)\}$
    $k_B \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_B, R_2 + \beta_\sigma + k_\tau |\tau_B|^2)\}$
  else
    $U \leftarrow \text{Sample}\{\text{Uniform}(0,1)\}$
    $x \leftarrow 2R_3/(\sigma_A^{(0)}\sigma_B^{(0)}) - \log U$
    if $R_3 > 0$ **and** $x > 0$ **then**
      $k_A \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_A, R_1 + \beta_\sigma + k_\tau |\tau_A|^2) \,|\, k_A \le {\sigma_B^{(0)}}^2 x^2/(4R_3^2)\}$
      $k_B \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_B, R_2 + \beta_\sigma + k_\tau |\tau_B|^2) \,|\, k_B \le x^2/(4R_3^2 k_A)\}$
    else if $R_3 < 0$ **and** $x < 0$ **then**
      $k_A \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_A, R_1 + \beta_\sigma + k_\tau |\tau_A|^2) \,|\, k_A \ge {\sigma_B^{(0)}}^2 x^2/(4R_3^2)\}$
      $k_B \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_B, R_2 + \beta_\sigma + k_\tau |\tau_B|^2) \,|\, k_B \ge x^2/(4R_3^2 k_A)\}$
    else
      $k_A \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_A, R_1 + \beta_\sigma + k_\tau |\tau_A|^2)\}$
      $k_B \leftarrow \text{Sample}\{\text{Gamma}(\alpha_\sigma + n_B, R_2 + \beta_\sigma + k_\tau |\tau_B|^2)\}$
    end if
  end if
  return $(\sigma_A = k_A^{-1/2}, \sigma_B = k_B^{-1/2})$

where R_4 and R_5 are defined in (3.18). We sample this with an auxiliary variable. Define z with support $\mathbb{R}$ by

$$p(\tilde{\omega}, z) \propto \tilde{\omega}^{\alpha_\omega + n_\xi} \exp\{-(\beta_\omega + R_4 - 2z)\tilde{\omega}\}\ \mathbb{1}(\tilde{\omega} \ge c_\omega)\ \mathbb{1}\Big(z < R_5\sqrt{1-\tilde{\omega}^{-1}}\Big). \quad (5.8)$$

By integrating out z this reduces to (5.7). We sample $z \,|\, \tilde{\omega}$ by direct inversion of the distribution function,

$$z \leftarrow \frac{1}{2\tilde{\omega}} \log U + R_5\sqrt{1-\tilde{\omega}^{-1}}, \quad (5.9)$$

where U is uniform on $(0,1)$. We sample $\tilde{\omega} \,|\, z$ by noticing it is a Gamma$(\alpha_\omega + n_\xi + 1, \beta_\omega + R_4 - 2z)$ left-truncated at $\max\Big\{c_\omega, \text{sgn}(z)\Big|1 - (z/R_5)^2\Big|^{-1}\Big\}$.

§5.6 Sampling the orientation distortion parameter κ

The full conditional of κ is proportional to

$$I_0(\kappa)^{-2(\alpha_\kappa - 1) - n_\xi} \exp\Big[\{2\alpha_\kappa - 2 - \beta_\kappa + \text{Re}(\psi S_\xi)\}\kappa\Big], \quad (5.10)$$

where S_ξ is given in (3.5). We sample this using the algorithm in Appendix A.

§5.7 Sampling the matching ξ

Finally, we sample the matching ξ. A possible Metropolis–Hastings sampler for ξ is described in Green and Mardia (2006). They propose creating or breaking a single, random matched pair at each iteration. In contrast, our algorithm 5.4 randomly selects a minutia $\beta \in B$ and then considers creating an edge between β and each $a \in A$, as well as allowing β to be unmatched. Empirically our sampler converges faster than the sampler in Green and Mardia.

We need one more piece of notation. In analogy with $\Pi_{A,b}$ (4.20), for each $a \in \mathbb{M}$, let $\Pi_{B,a}$ be the minutia $b \in B$ which is matched to a:

$$\Pi_{B,a}(\xi) = \begin{cases} \phi & \text{if } \{\langle a,b\rangle \in \xi : b \in B\} = \emptyset, \\ b \in B : \langle a,b\rangle \in \xi & \text{otherwise,} \end{cases} \tag{5.11}$$

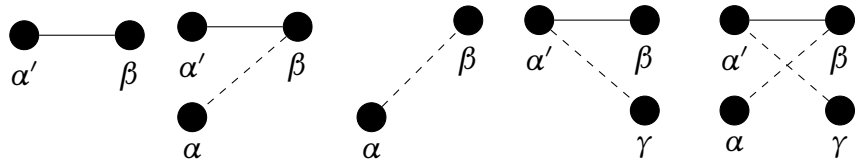
where, as before, ϕ is a fixed but arbitrary point in $\mathbb{M} \setminus (A \cup B)$ which is used to denote the fact that a is not matched to any $b \in B$.

Formally, we sample ξ with the help of an auxiliary random variable β. We use the following transition kernel to move in the augmented state space from (ξ, β) to (ξ', β'):

$$q_\xi(\xi', \beta' \mid \xi, \beta) \propto p(A, B, \theta, \xi' \mid H_p)\, \mathbb{1}\big(\xi' \setminus \{\langle \alpha', \beta\rangle\} = \xi \setminus \{\langle \alpha, \beta\rangle, \langle \alpha', \gamma\rangle\}\big), \tag{5.12}$$

where $\alpha' = \Pi_{A,\beta}(\xi')$ is the new match of β, $\alpha = \Pi_{A,\beta}(\xi)$ is the old match of β, $\gamma = \Pi_{B,\alpha'}(\xi)$ is the old match of α', and $p(A, B, \theta, \xi' \mid H_p)$ is given in (5.1). If any of these minutiae are unmatched, then the corresponding variable (α', α or γ) will be equal to ϕ. Note that $q_\xi(\xi', \beta' \mid \xi, \beta)$ is independent of the value of β', and thus β' is selected uniformly at random over B.

This transition kernel allows transitions to any ξ' which differs from ξ only in its matches for α' and β. The states ξ' which are accessible from the state ξ are illustrated in figure 5.1. We can move from any state ξ to any other state ξ' in at most n_B steps, so the Markov chain with this transition kernel is irreducible. Clearly it is also aperiodic and therefore ergodic. Its stationary distribution is $p(\xi \mid \theta, A, B, H_p)$ as desired.



(a) Add (b) Swap on A (c) Remove (d) Swap on B (e) Add/remove (f) No change

Figure 5.1: Illustration of which states for ξ' are accessible from a given state ξ when the auxiliary variable is equal to β, using the same notation as (5.12). The dashed edges are the removed matches: $\langle\alpha,\beta\rangle \in \xi$ and $\langle\alpha',\gamma\rangle \in \xi$. The solid edge is the new match, $\langle\alpha',\beta\rangle \in \xi'$. Edges that are common to both ξ' and ξ are not shown.

The densities of the allowed states have many terms in common. By ignoring these common terms, we have

$$q_\xi(\xi',\beta' \mid \xi,\beta) \propto \exp[w(\alpha',\beta \mid \theta) - w\{\alpha',\gamma \mid \theta\}] \times \mathbb{1}\big(\xi' \setminus \{\langle\alpha',\beta\rangle\} = \xi \setminus \{\langle\alpha,\beta\rangle,\langle\alpha',\gamma\rangle\}\big), \tag{5.13}$$

where w is given in (3.7). Thus the proposal function can be computed very quickly, and it can be normalized over ξ' by simply summing over the permitted moves. There are $n_A + 1$ such moves, one for each possible value of $a \in A \cup \phi$. The full algorithm is described in algorithm 5.4.

Algorithm 5.4 Sampler for ξ using the auxiliary variable β

Require: Previous value ξ
  $\beta \leftarrow$ Sample from the uniform distribution on B
  $\alpha \leftarrow \Pi_{A,\beta}(\xi)$ # find the old match of β
  $\xi' \leftarrow \xi \setminus \{\langle\alpha,\beta\rangle\}$ # remove the old match, $\langle\alpha,\beta\rangle$
  $\alpha' \leftarrow$ Sample$\big(p(a) \propto \exp[w(a,\beta \mid \theta) - w\{a,\Pi_{B,a}(\xi) \mid \theta\}]$ for $a \in A \cup \phi\big)$
  if $\alpha' \neq \phi$ **then**
    $\xi' \leftarrow \xi' \setminus \{\langle\alpha',\Pi_{B,\alpha'}(\xi)\rangle\}$ # remove the old match of α'
    $\xi' \leftarrow \xi' \cup \{\langle\alpha',\beta\rangle\}$ # add the new match, $\langle\alpha',\beta\rangle$
  end if
  return ξ'

Chapter Six

Results

In this chapter we compare the performance of our various methods to estimate the likelihood ratio between H_p and H_d. We test our methods on both simulated data (§6.3.1) and real data (§6.3.2).

§6.1 Our dataset

To test the algorithm on real world data we use a small database provided by the National Institute for Standards and Technology and the Federal Bureau of Investigation (Garris and McCabe, 2000). This database consists 258 fingerprint/fingermark pairs, for a total of 516 digital images. The fingerprints (the *A*s) are all of high quality, and the fingermarks (the *B*s) are of significantly lower quality. The fingerprint/fingermark pairs are partitioned into three sets based on the quality of the fingermarks: 88 are "good", 85 are "bad", and 85 are "ugly" (see figure 6.1). All fingermarks and fingerprint images have their minutiae hand-labelled by expert fingerprint examiners. Most fingermark minutia types are classified as bifurcations or ridge endings, but almost all fingerprint minutia types are unclassified. This dataset is used for estimation of unknown parameters, for model validation, and for evaluating the performance of the calculated likelihood ratio.

Section 6.2 is based off Forbes et al. (2014).

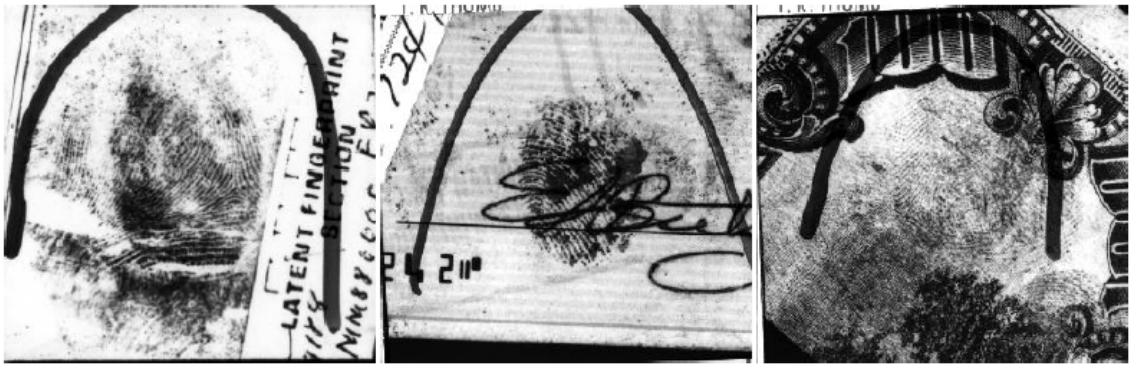


Figure 6.1: Example fingermarks from Garris and McCabe (2000). From left to right, the fingermark qualities are good, bad and ugly.

§6.2 Fixed parameter estimation

We must find point estimates for the fixed parameters $\rho_0, \chi, c_\omega, k_\tau$, and $\{\alpha_\iota, \beta_\iota : \iota = \delta, \sigma, \omega, \kappa\}$. We adopt an empirical Bayes approach and estimate the fixed parameters by maximizing the likelihood of the observed data in our dataset. This implicitly assumes that our dataset is a representative sample for the type of fingerprint evidence which we wish to evaluate using our model. As our real dataset contains matched fingerprint/fingermark pairs which conform with the prosecution hypothesis, we estimate all parameters under H_p.

The parameter estimates in this section are specific to our dataset (Garris and McCabe, 2000). If the model described here is used in a different context, the parameters should be chosen appropriately: either by performing a similar estimation procedure on a relevant dataset, or by exploiting prior scientific knowledge, or by maximizing some desired objective function such as the discrimination between ground-truth cases of H_p and H_d. The interpretation of the resulting likelihood ratio will change depending on how the fixed parameters were chosen: while the first two methods yield likelihood ratios which might reasonably be interpreted as the ratio of two probabilities as computed by a scientific model, the latter method yields "likelihood ratio" which should only be interpreted as discrimination score. Such a discrimination score would have to be calibrated appropriately before it could be viewed as a likelihood ratio (see chapter 9).

We found it difficult to achieve reliable empirical Bayes estimates of the fixed parameters without knowledge of the correct matching ξ. Unfortunately our dataset contains only the 258 paired minutia configurations without matching the corresponding minutiae within a configuration; that is, it contains A_i and B_i but not ξ_i for $i = 1 \ldots 258$. Previous research (Mikalyan and Bigun, 2012) attempted to ameliorate this by running an automated matching algorithm on the dataset Garris and McCabe (2000). However, we found the quality of these matchings to be poor. Instead we manually found and recorded what we believe to be the correct minutia matchings $\check{\xi}_i$ for each of the 258 fingerprint/fingermark pairs in the dataset. With the "true" matching $\check{\xi}_i$ fixed, we proceeded with the parameter estimation. We emphasize that this "true" matching was only used for estimation of the unknown parameters.

We estimate the fixed parameters by maximizing

$$\prod_{i=1}^{258} \left\{ \int p(A_i, B_i, \check{\xi}_i, \theta_i \,|\, H_p) \, \mathrm{d}\mu_\theta(\theta_i) \right\} = \prod_{i=1}^{258} \left\{ p(A_i, B_i, \check{\xi}_i \,|\, \rho_0, \chi, c_\omega, k_\tau, \{\alpha_\iota, \beta_\iota : \iota = \delta, \sigma, \omega, \kappa\}, H_p) \right\}, \quad (6.1)$$

where $p(A_i, B_i, \check{\xi}_i, \theta \,|\, H_p)$ is given in (5.1), and where the fixed parameters have been suppressed on the left-hand side of this equation. Each integrand on the left-hand side further factorizes into

$$p(A_i, B_i, \check{\xi}_i, \theta \,|\, H_p) = f_0(A_i, B_i, \check{\xi}_i) \times f_1(A_i, B_i, \check{\xi}_i, \delta_A, \delta_B; \alpha_\delta, \beta_\delta, \rho_0) \times f_2(A_i, B_i, \check{\xi}_i; \chi) \times f_3(A_i, B_i, \check{\xi}_i, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa; c_\omega, k_\tau, \{\alpha_\iota, \beta_\iota : \iota = \sigma, \omega, \kappa\}), \quad (6.2)$$

where f_0 is independent of the parameters and can be ignored. The remaining functions f_1, f_2, f_3 are

$$f_1(A, B, \check{\xi}, \delta_A, \delta_B; \alpha_\delta, \beta_\delta, \rho_0) = \exp\{-\rho_0(\delta_A + \delta_B - \delta_A\delta_B)\} \frac{\Gamma(\alpha_\delta + \beta_\delta)}{\Gamma(\alpha_\delta)\Gamma(\beta_\delta)} \times \rho_0^{n_A + n_B - n_\xi} \delta_A^{\alpha_\delta + n_A - 1} \delta_B^{n_B} (1 - \delta_A)^{\beta_\delta + n_B - n_\xi - 1} (1 - \delta_B)^{n_A - n_\xi}, \quad (6.3)$$

$$f_2(A, B, \check{\xi}; \chi) = \chi^{n_A^{(1)} + n_B^{(1)} - n_\xi^{(1)}} (1 - \chi)^{n_A^{(0)} + n_B^{(0)} - n_\xi^{(0)}}, \quad (6.4)$$

$$
\begin{aligned}
&f_3(A_i, B_i, \check{\xi}_i, \tau_A, \tau_B, \psi, \sigma_A, \sigma_B, \tilde{\omega}, \kappa; c_\omega, k_\tau, \{\alpha_\iota, \beta_\iota : \iota = \sigma, \omega, \kappa\}) \\
&= \frac{k_\tau^2 \beta_\sigma^{2\alpha_\sigma} \lambda_\kappa(\alpha_\kappa, \beta_\kappa) \beta_\omega^{\alpha_\omega}}{\Gamma(\alpha_\sigma)^2 \Gamma(\alpha_\omega, \beta_\omega c_\omega)} \sigma_A^{-2(\alpha_\sigma + n_A) - 3} \sigma_B^{-2(\alpha_\sigma + n_B) - 3} \tilde{\omega}^{\alpha_\omega + n_\xi - 1} I_0(\kappa)^{2(\alpha_\kappa - 1) - n_\xi} \\
&\times \exp\left\{-(k_\tau |\tau_A|^2 + \beta_\sigma)\sigma_A^{-2} - (k_\tau |\tau_B|^2 + \beta_\sigma)\sigma_B^{-2} - \beta_\omega \tilde{\omega} + (2\alpha_\kappa - 2 - \beta_\kappa)\kappa\right\} \\
&\times \exp\left\{-\frac{S_{A\setminus\xi} - 2\,\mathrm{Re}(\overline{R_{A\setminus\xi}}\tau_A) + (n_A - n_\xi)|\tau_A|^2}{\sigma_A^2} - \frac{S_{B\setminus\xi} - 2\,\mathrm{Re}(\overline{R_{B\setminus\xi}}\tau_B) + (n_B - n_\xi)|\tau_B|^2}{\sigma_B^2}\right\} \\
&\times \exp\left[2\sqrt{\tilde{\omega}^2 - \tilde{\omega}} \frac{\mathrm{Re}\left\{\psi\left(S_{\xi AB} - \overline{R_{\xi A}}\tau_B - R_{\xi B}\overline{\tau_A} + n_\xi \overline{\tau_A}\tau_B\right)\right\}}{\sigma_A \sigma_B} + \kappa\,\mathrm{Re}(\psi S_\xi)\right] \\
&\times \exp\left[-\tilde{\omega}\left\{\frac{S_{\xi A} - 2\,\mathrm{Re}(\overline{R_{\xi A}}\tau_A) + n_\xi |\tau_A|^2}{\sigma_A^2} + \frac{S_{\xi B} - 2\,\mathrm{Re}(\overline{R_{\xi B}}\tau_B) + n_\xi |\tau_B|^2}{\sigma_B^2}\right\}\right].
\end{aligned}
\tag{6.5}
$$

Since $(\alpha_\delta, \beta_\delta, \rho_0)$ only enter into f_1, the estimates for these parameters are the maximizers of

$$
\prod_{i=1}^{258}\left\{\int f_1(A_i, B_i, \check{\xi}_i, \delta_{Ai}, \delta_{Bi}; \alpha_\delta, \beta_\delta, \rho_0)\,\mathrm{d}(\delta_{Ai}, \delta_{Bi})\right\}. \tag{6.6}
$$

The integral over δ_B can be done analytically as in §4.2. The integral over δ_A can be done numerically, and the resulting function can be maximized numerically. We used the CRAN package **pracma** for the integrals and the base R function **optim** for the optimization. The resulting estimates are $\widehat{\alpha_\delta} = 14.7, \widehat{\beta_\delta} = 3.30$, and $\widehat{\rho_0} = 133$.

Similarly, χ only enters into f_2 and can be found by directly maximizing $\sum_{i=1}^{258} \log f_2(A_i, B_i, \check{\xi}_i; \chi)$. This yields a linear equation for χ with solution $\widehat{\chi} = 0.384$.

We use a stochastic expectation-maximization approach to estimate the remaining fixed parameters. Specifically, let $\alpha_\sigma, \beta_\sigma, \alpha_\omega, \beta_\omega, \alpha_\kappa, \beta_\omega, c_\omega$, and k_τ be some initial values and for $i = 1, \ldots, 258$ generate a sample $\theta_i = (\delta_{Ai}, \delta_{Bi}, \tau_{Ai}, \tau_{Bi}, \sigma_{Ai}, \sigma_{Bi}, \psi_i, \tilde{\omega}_i, \kappa_i)$ from the distribution

$$
p(\theta \,|A_i, B_i, \check{\xi}_i, \widehat{\rho_0}, \widehat{\chi}, c_\omega, k_\tau, \widehat{\alpha_\delta}, \widehat{\beta_\delta}, \{\alpha_\iota, \beta_\iota : \iota = \sigma, \omega, \kappa\}, H_p) \tag{6.7}
$$

using algorithm 5.1. We update $\alpha_\sigma, \beta_\sigma, \alpha_\omega, \beta_\omega, \alpha_\kappa, \beta_\omega, c_\omega, k_\tau$ to the maximizers of

$$
\sum_{i=1}^{258} \log \int_{c_\omega}^{\infty} f_3(A_i, B_i, \check{\xi}_i, \tau_{Ai}, \tau_{Bi}, \sigma_{Ai}, \sigma_{Bi}, \psi_i, \tilde{\omega}, \kappa_i; c_\omega, k_\tau, \{\alpha_\iota, \beta_\iota : \iota = \sigma, \omega, \kappa\})\,\mathrm{d}\tilde{\omega}. \tag{6.8}
$$

Both the integration and the maximizations are done numerically. Note that the integral over $\tilde{\omega}$ is necessary to get a reasonable estimate for c_ω: if we were to simply use the sampled value $\tilde{\omega}_i$, the estimate would always be $c_\omega = \min_i \tilde{\omega}_i$.

We use the values for $c_\omega, k_\tau, \alpha_\sigma, \beta_\sigma, \alpha_\omega, \beta_\omega, \alpha_\kappa, \beta_\kappa$ to generate a new θ_i. By iterating this procedure the sequence of values will stabilize. After stabilization we generate 500 more values and set our estimates of to the average of these samples. Our estimates are $\widehat{c_\omega} = 65, k_\tau = 0.25, \widehat{\alpha_\sigma} = 12.6, \widehat{\beta_\sigma} = 345,000, \widehat{\alpha_\omega} = 2.0, \widehat{\beta_\omega} = 0.0079, \widehat{\alpha_\kappa} = 4.4$, and $\widehat{\beta_\kappa} = 0.087$.

§6.3 Marginalized likelihood computations

We compute the likelihood ratio for each fingerprint/fingermark pair independently. When the fingerprint and fingermark originate from the same finger we call them a *true match*, otherwise they are a *false match*. We compute each likelihood ratio using each of the four methods described in §4.3.1 to §4.3.4: harmonic means, bridge sampling, Chib's method, and reversible jump.

The code is written in C# version 4.51. We use this language due to its multi-thread support and advanced data plotting capabilities. Our algorithm generates approximately 1,500 joint samples of θ and ξ per thread per second on a 3GHz Intel Xeon processor.

When computing a specific likelihood ratio, we start by setting the initial value of ξ to the empty match. Within 10,000 iterations the variable traces (figure 6.2) appear to be stationary. We burn a further 5,000 samples to remove any transient effects of our initial position. Next, we generate 5,000 samples which we use to estimate the parameters for the proposal distribution $q(\theta)$, as discussed in §4.3.5.

With our initialization completed, we generate a set S of 10,000 samples from the model under H_p. We use S to approximate the expectations with respect to $p(A, B \mid H_p)$ in the harmonic means estimate (§4.3.1) and the bridge sampling estimate (§4.3.3). We complete the bridge sampling estimate by generating S_0, which consist of 10,000

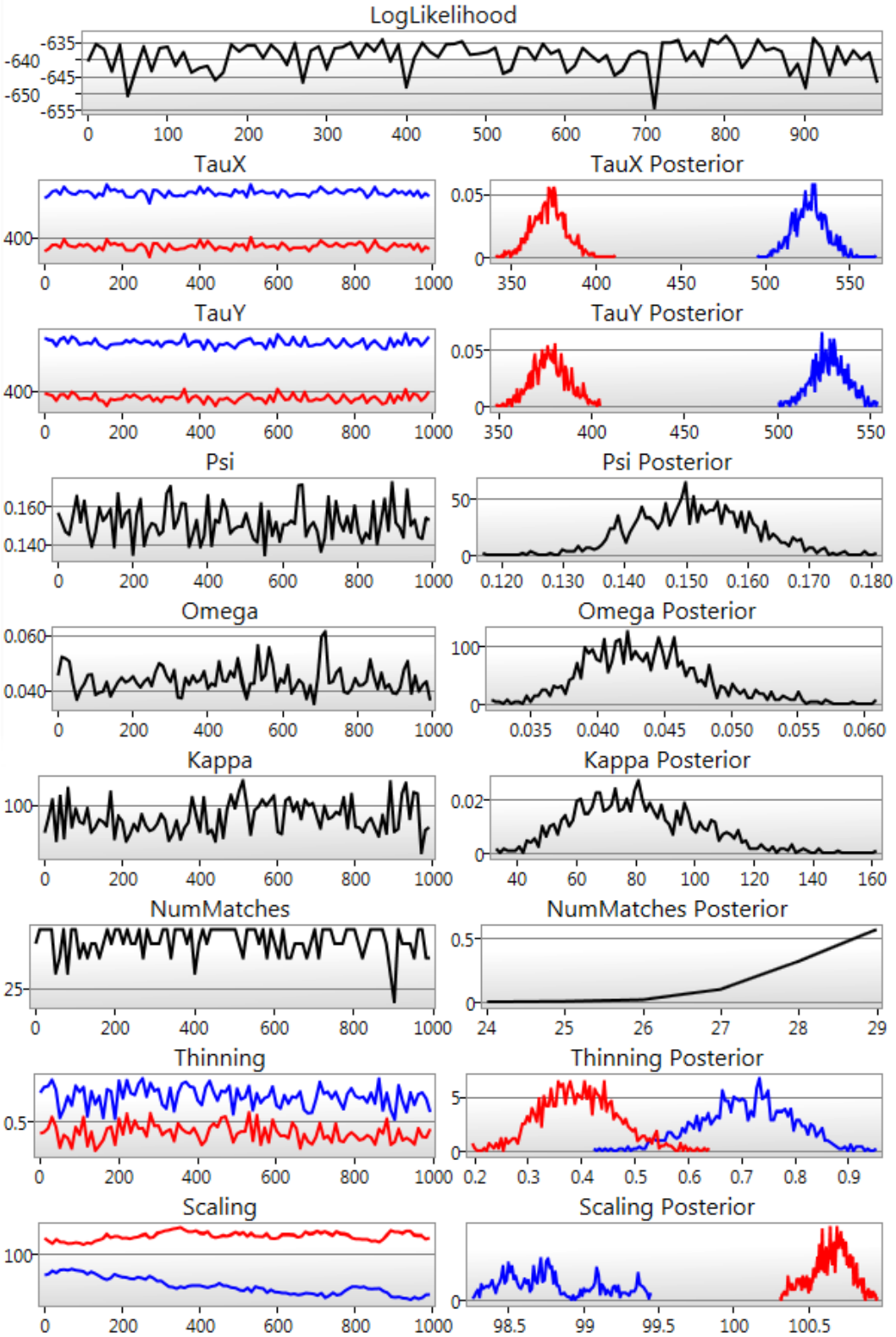


Figure 6.2: Graphical user interface showing partial traces and the posterior distribution of the model parameters after an initial burn-in. These traces are typical when the fingerprint and fingermark originate from the same finger, and hence the posterior likelihood has a single prominent mode.

samples from the distribution $q(\theta, \xi)$. We have not investigated the extent of the autocorrelation in these samples, though we expect it to be significant due to the high-dimensionality of the matching ξ. Despite this autocorrelation, $10,000$ samples seems sufficient to reduce the Monte Carlo error of the bridge sampling estimate of the log-likelihood ratio estimate to less than 0.2. The Monte Carlo error of the weighted harmonic means estimate is also usually below 0.2.

Let $\widehat{p}(A, B \mid H_p)$ be the bridge sampling estimate of the likelihood under H_p. We choose the reversible jump tuning parameter p_0 so that

$$\frac{p_0}{1-p_0} = \frac{\widehat{p}(A, B \mid H_p)}{p(A, B \mid H_d)}, \tag{6.9}$$

where the denominator of the right-hand side is given in (4.12). This ensures that $p(A, B, H_d) \approx p(A, B, H_p)$ in the density (4.27) so that our chain mixes between the two models H_p and H_d. We then generate $20,000$ samples from a Markov chain with the equilibrium distribution (4.27), which we use to approximate the expectations in (4.33) and compute a reversible-jump estimate of the likelihood ratio. On average half of these samples will have $H = H_p$.

We compute the Chib's estimate of the likelihood ratio using algorithm 4.1. In total, with $N = 1,000$ and $N_\xi = 200$, this algorithm requires $9,000 + 200 n_B \approx 25,000$ samples per likelihood ratio estimated. The Monte Carlo error of the log-likelihood ratio estimate is nearly always less than 0.2.

We compare these estimates of the marginalized likelihood ratio with an estimate of the profile likelihood ratio, which is computed by following the iterative procedure of chapter 3. We use a custom graphical user interface to enable visualization of the maximizing parameters $\widehat{\theta}$ and $\widehat{\xi}$ and qualitatively evaluate the model fit under H_p; an example screenshot of this visualization is shown in figure 6.3.

To assess the relative performance of these estimation methods, we frame our likelihood estimation problem as a binary classification problem and examine receiver operating characteristic (ROC) curves. We classify each fingerprint/fingermark pair with estimated likelihood ratio $\widehat{\Lambda}$ as a match if $\widehat{\Lambda} > \Lambda_{\text{threshold}}$ for some $\Lambda_{\text{threshold}} \in \mathbb{R}$,

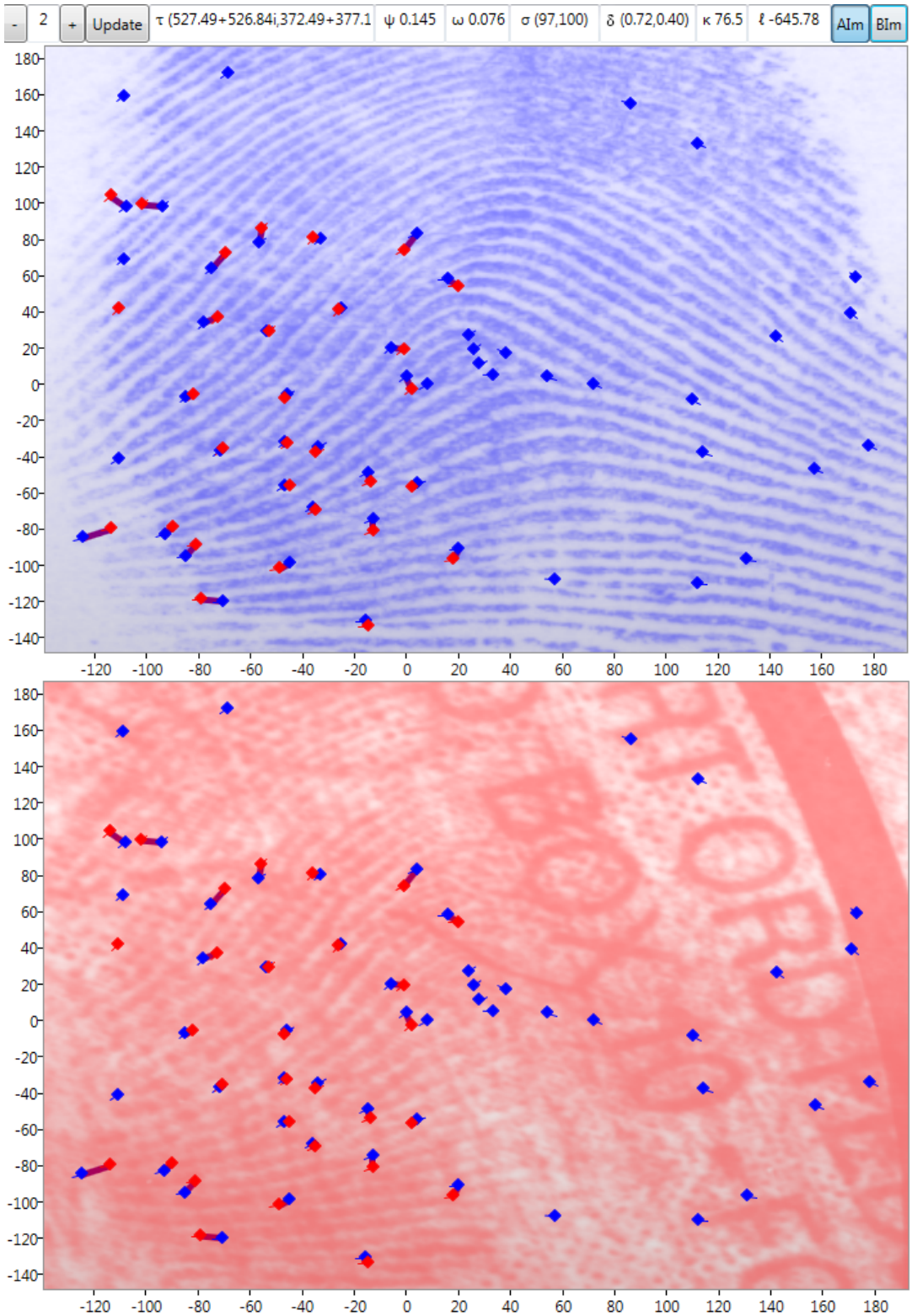


Figure 6.3: Graphical user interface showing the fingerprint (top) and fingermark (bottom) images, with the minutia configurations overlaid. Purple lines are drawn between matched minutiae for the estimated matching $\widehat{\xi}$.

and otherwise we classify it as a non-match. The shape of the ROC curve as $\Lambda_{\text{threshold}}$ ranges from negative infinity to infinity allows us to evaluate our model's ability to discriminate between H_p and H_d. Thus the ROC curves can be used to quantify the predictive accuracy of our model when it is used as a classifier for true and false matches.

§6.3.1 Results on simulated dataset

Before we compute the results for our real dataset, we first validate our algorithm implementation by considering a simulated data set. We first generated several fingerprint/fingermark pairs of four minutiae each: these were small enough to compute the likelihood ratio by brute force, which involved summing over the matching $\xi \in \Xi(A, B)$ and numerically integrating over the parameters θ for each match. We confirmed that our algorithm computed the correct likelihood ratio for these small test pairs.

Next, we generated 258 fingerprint/fingermark pairs $\{(A_i, B_i) : 1 \leq i \leq 258\}$ according to the density

$$\int p(A, B \mid \theta, H_d) p(\theta) \, \mathrm{d}\mu_\theta(\theta), \tag{6.10}$$

where the integrand is the product of (2.36) and (4.5). The integrand depends on the fixed parameters $\rho_0, \chi, k_\tau, \alpha_\delta, \beta_\delta, \alpha_\sigma, \beta_\sigma$; we used the values estimated in §6.2, which resulted in many of the simulated fingerprints having in excess of one hundred minutiae. We will call the generated fingerprint/fingermark pair (A_i, B_j) a true match if $i = j$, otherwise we call it a false match.

To ease the comparison with the real database, we also partitioned the simulated data into a good set consists of those 88 pairs with the highest number of fingermark minutiae n_B, a bad set containing the next 85 pairs, and an ugly set containing those 85 pairs with the lowest n_B. We then compute the estimated log-likelihood ratio for all true matches and a random 5% subset of the false matches using each method described in §4.3.

The resulting ROC curves are shown in figure 6.4.[†] A perfect classifier's curve would look like the letter "Γ": the point located at the top-left corner of the plot represents perfect discrimination between true and false matches. Conversely, a method with no discrimination between true and false matches would have a straight line from the bottom-left to the top-right corner. Thus, we can evaluate the discriminatory power of each method by inspecting how closely the ROC curve approaches the upper-left corner. As expected, we see that the discrimination is best for good quality fingermarks and worst for ugly quality fingermarks. However, even for ugly fingermarks, we have almost complete discrimination between H_p and H_d. This provides some evidence that our algorithm is performing as expected.

To investigate further, we plot the histograms of the computed log-likelihood ratios in figure 6.5. Reassuringly, we see that all the methods give similar results except for naive harmonic means. The failure of the naive harmonic means estimate is to be expected because it often has infinite variance (see, for example, the response to Raftery et al. (2007) by Robert and Chopin). Out of the remaining marginalization methods, Chib's estimate has the best performance: it is the only method without a long tail of false matches with large positive log-likelihood ratios. This is likely because our implementation of Chib's method (algorithm 4.1) estimates the posterior distributions $p(\theta \mid A, B, H_p)$ and $p(\xi \mid \theta, A, B, H_p)$ separately, whereas the other methods deals with the joint distribution $p(\theta, \xi \mid A, B, H_p)$. Due to the high dimensionality of ξ, the former approach is easier.

As noted in §4.3.3, Chib's method can be seen as a special case of bridge sampling. Thus we would expect that an improved bridge sampling estimate, which uses a more sophisticated bridge to estimate $p(\theta \mid A, B, H_p)$ and $p(\xi \mid \theta, A, B, H_p)$ separately, would dominate the performance of Chib's estimate. We have not attempted to design or implement such a bridge.

Finally, we note that our implementation of reversible jump has the largest tail of

[†]The figures 6.4–6.7 have been printed side-by-side at the end of this chapter.

false matches with positive log-likelihood ratios. We suspect this is due to our proposal distribution $q(\theta, \xi)$ not being sufficiently close to the posterior distribution of (θ, ξ). In particular, our proposal distribution tends to be concentrated about a single mode of the highly multimodal H_p parameter space, which sometimes causes our Markov chain to become stuck in a local mode of H_p. This problem was also discussed near the end of §4.3.4.

§6.3.2 Results on real dataset

We estimate the log-likelihood ratios for all 258 true matches and a random 5% subset of the 258 × 257 false matches, for a total of 3,574 computed likelihood ratios. Figure 6.4 shows the resulting ROC curves for each of the three subsets described in §6.1. We immediately note that the discrimination is much worse than for simulated data, which strongly implies that our model does not match the actual data very well. We investigate this in more detail in chapter 7. Furthermore, while we expect the discrimination between true and false matches to be lower for lower quality fingermarks, we see that the model has almost no discriminatory power for ugly fingermarks.

We investigate more closely by looking at the histograms of log-likelihood ratios in figure 6.6. As in the simulated case, we note that every method (except for naive harmonic means) yields similar estimates of the log-likelihoods, which lends some credibility to our implementation's correctness. We note many of the same features exist here as in the simulated data. In particular, the relatively poor performance of the naive harmonic means estimate can be explained by that method's often infinite variance, and the large tail of false matches with positive log-likelihood ratios can be explained by poor mixing of the Markov chain over the highly multimodal parameter space under H_p.

Notice that the log-likelihood ratios for false matches are positive on average. This is a striking contrast to the simulated data, where almost all of the false matches had negative log-likelihood ratios. This shift is caused by a poor model fit. In particular,

our model assumes that the minutia positions and orientations within a given minutia configuration should be independent, but in fact they are highly correlated for real-world data. Thus minutia configurations from distinct fingers share many common features which are unexplained under H_d, but can be partially explained under H_p.

We also notice that the true log-likelihood ratios are less positive for real data than simulated data. Once again, this can be explained by a poor model fit. In particular, the model under H_p assumes that the latent minutia are subjected to independent distortions, but in fact the distortions between nearby minutia are correlated (see §7.10). This causes the distortion variance to be inflated, which in turn decreases the likelihood of observed data under H_p.

The left-shift of the true log-likelihood ratios is strongest for the ugly quality fingermarks, which have the most distortion and hence suffer most from the assumption of independent distortions. Indeed, by considering only the good quality fingermarks, we observe a substantial shift to the right for the positive log-likelihood ratios (see figure 6.7). This implies that our current model may be adequate in a situation where the fingermarks under consideration have relatively little distortion.

§6.3.3 Discussion

The likelihood ratios found in this chapter are truly astronomical, exceeding 10^{80} in many cases. For comparison, a typical DNA identification has a likelihood ratio around 10^{14} (Balding, 1999). Such extreme likelihood ratios are cause for concern, especially since they are not supported by the relatively poor discrimination between true and false matches observed in figure 6.6. Our inflated likelihood ratios are the direct result of modelling inadequacies. The likelihood ratios must be calibrated against a training dataset before they can be interpreted as an accurate measure of the strength of evidence. We investigate methods of calibration and present the calibrated results in chapter 9.

Notice from figure 6.4 that the profile likelihood ratios often provide better discrimination than any of the marginalized likelihood ratios. The profile likelihood ratios are

also easier to compute, since they do not require any MCMC techniques. Furthermore, since the profile likelihood ratio is closely related to the standard maximized likelihood ratio, it may be easier to explain in court than a marginalized likelihood ratio. In fact, maximized likelihood ratios have already been used in many courts worldwide in connection to DNA evidence (Balding, 2005; Cowell et al., 2014). Thus, based on the results in this dissertation, the author advocates the use of profile likelihood ratios for the courtroom presentation of fingerprint evidence. These profile likelihood ratios should be based on a reasonable model for fingerprints and fingermarks. Since the profile likelihood ratios are prone to inflation due to modelling inadequacies, they should be calibrated against a large real-world database of fingerprints and fingermarks. Example calibration procedures are given in chapter 9.

We note that for both simulated and real data, the discrimination between false matches and true matches observed in this chapter is notably worse than the discrimination observed in Forbes et al. (2014). That paper uses a simpler model than the one described here. In particular, it assumes that the fingerprint and fingermark have the same scale, $\sigma_A = \sigma_B$, and furthermore that the location distortion parameter $\tilde{\omega}$ and the orientation distortion parameter κ are fixed constants rather than random variables. The loss of discrimination is likely because there is insufficient information in a given fingerprint/fingermark pair to accurately infer the joint distribution of the variability terms $\tilde{\omega}$ and κ with the matching ξ.

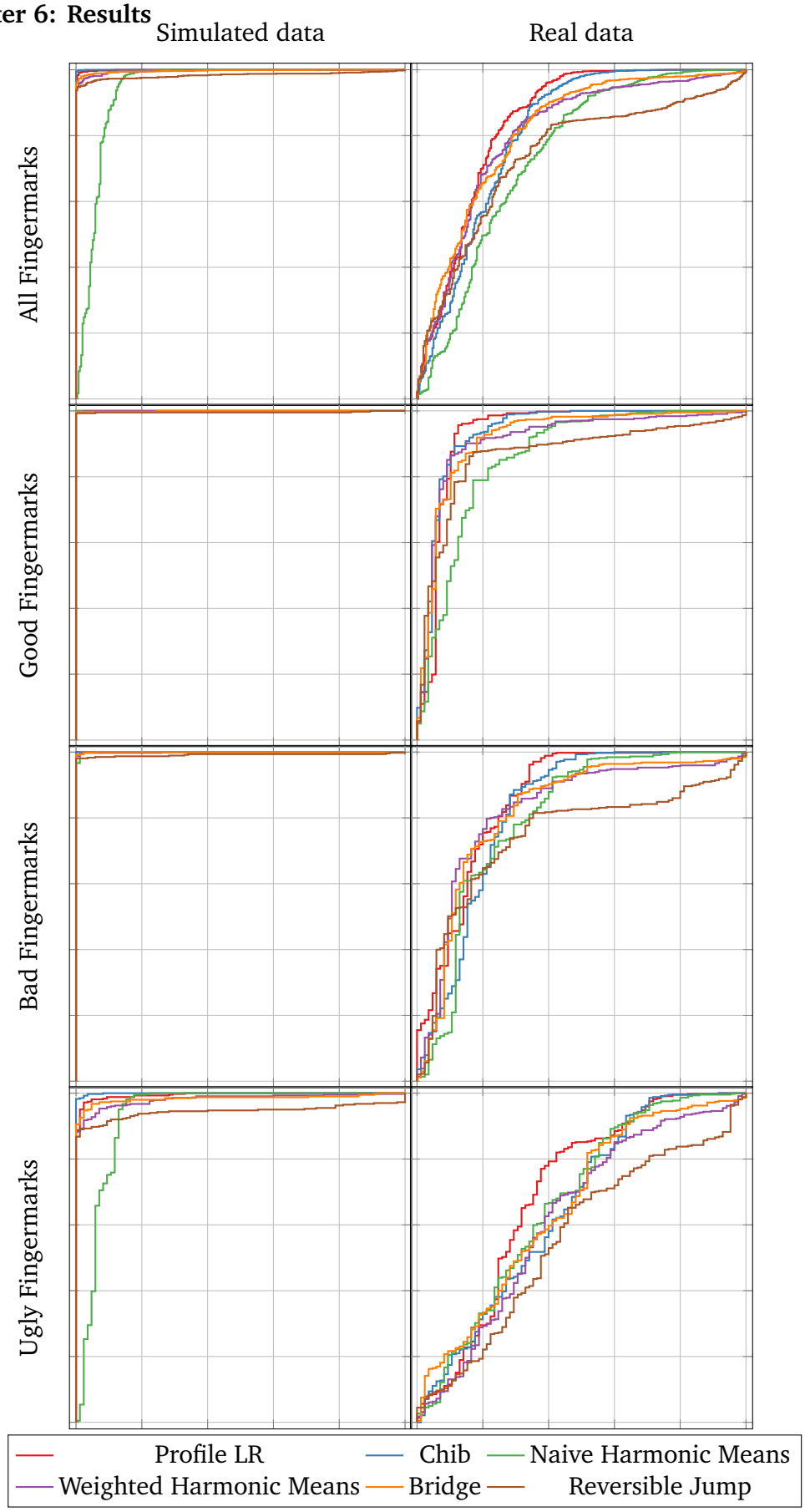


Figure 6.4: Receiver Operating Characteristic (ROC) curves for the dataset (Garris and McCabe, 2000), the simulated data, and their three subsets (good, bad and ugly). The x-axis represents the proportion of false matches classified as matches, the y-axis represents the proportion of true matches classified as matches.

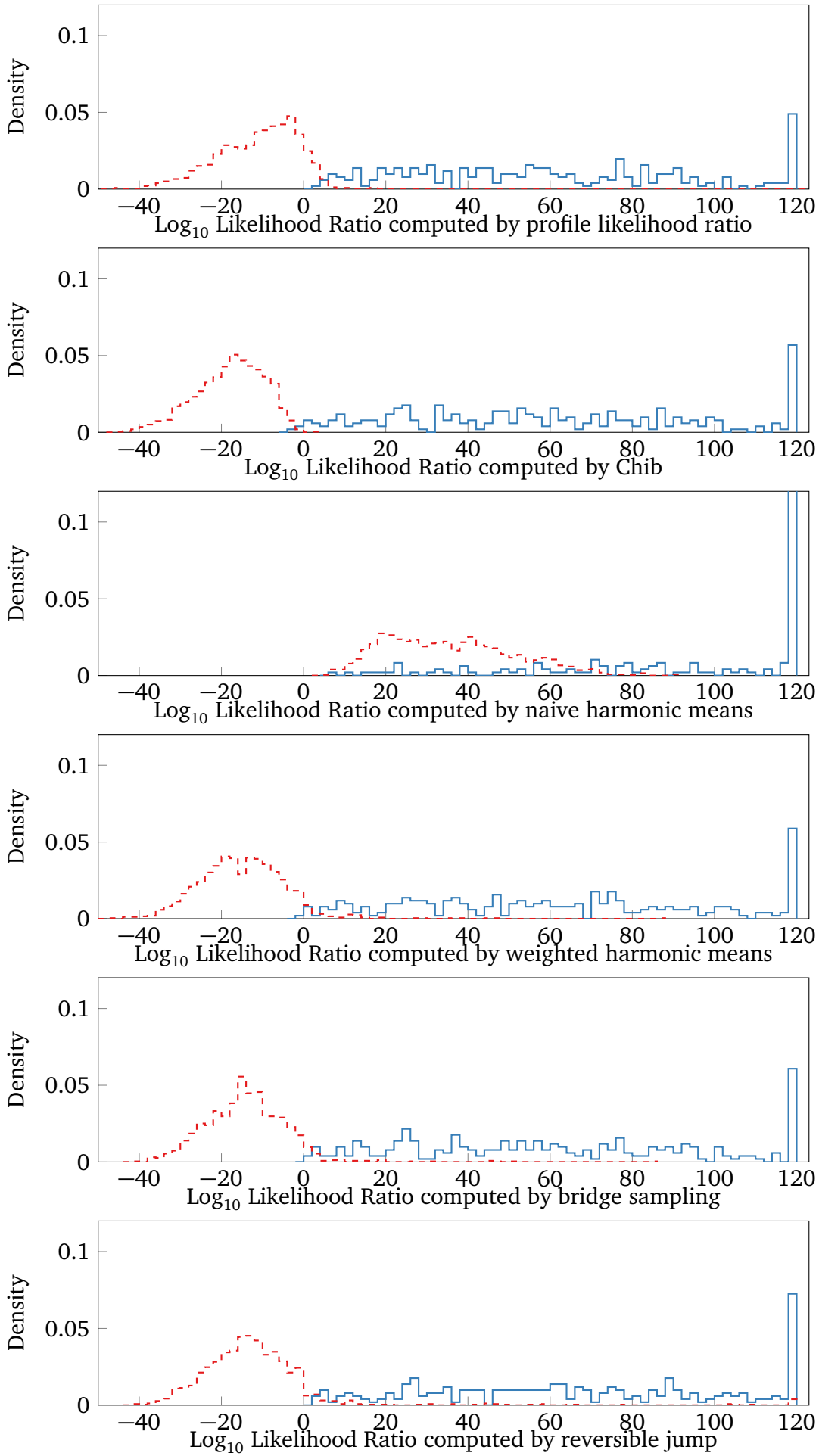


Figure 6.5: Histogram of the log-likelihood ratios computed on simulated data. Log-likelihood ratios corresponding to false matches are dashed and red, and true matches are solid and blue.

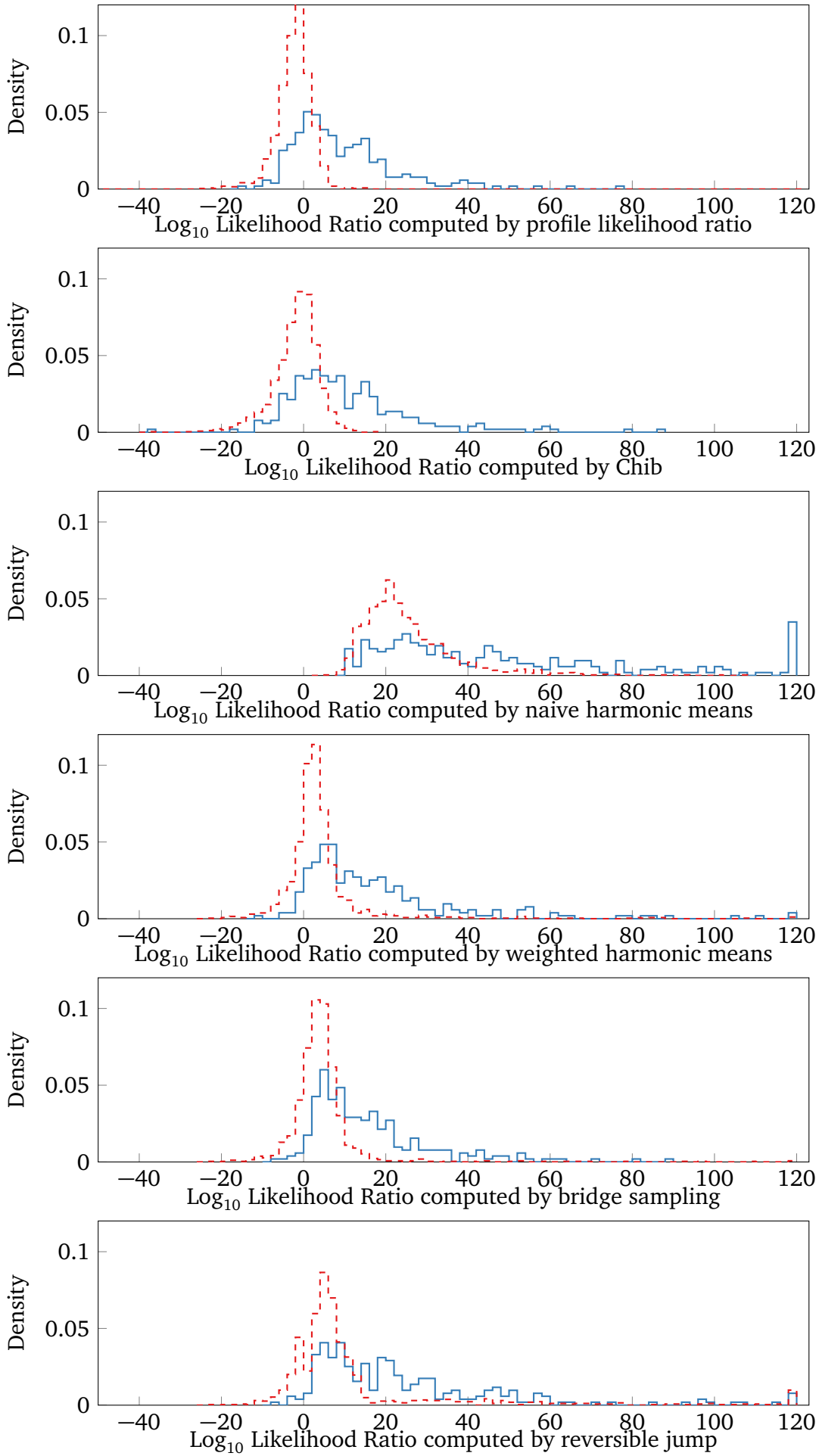


Figure 6.6: Histogram of the log-likelihood ratios computed on the dataset Garris and McCabe (2000). Log-likelihood ratios corresponding to false matches are dashed and red, and true matches are solid and blue.

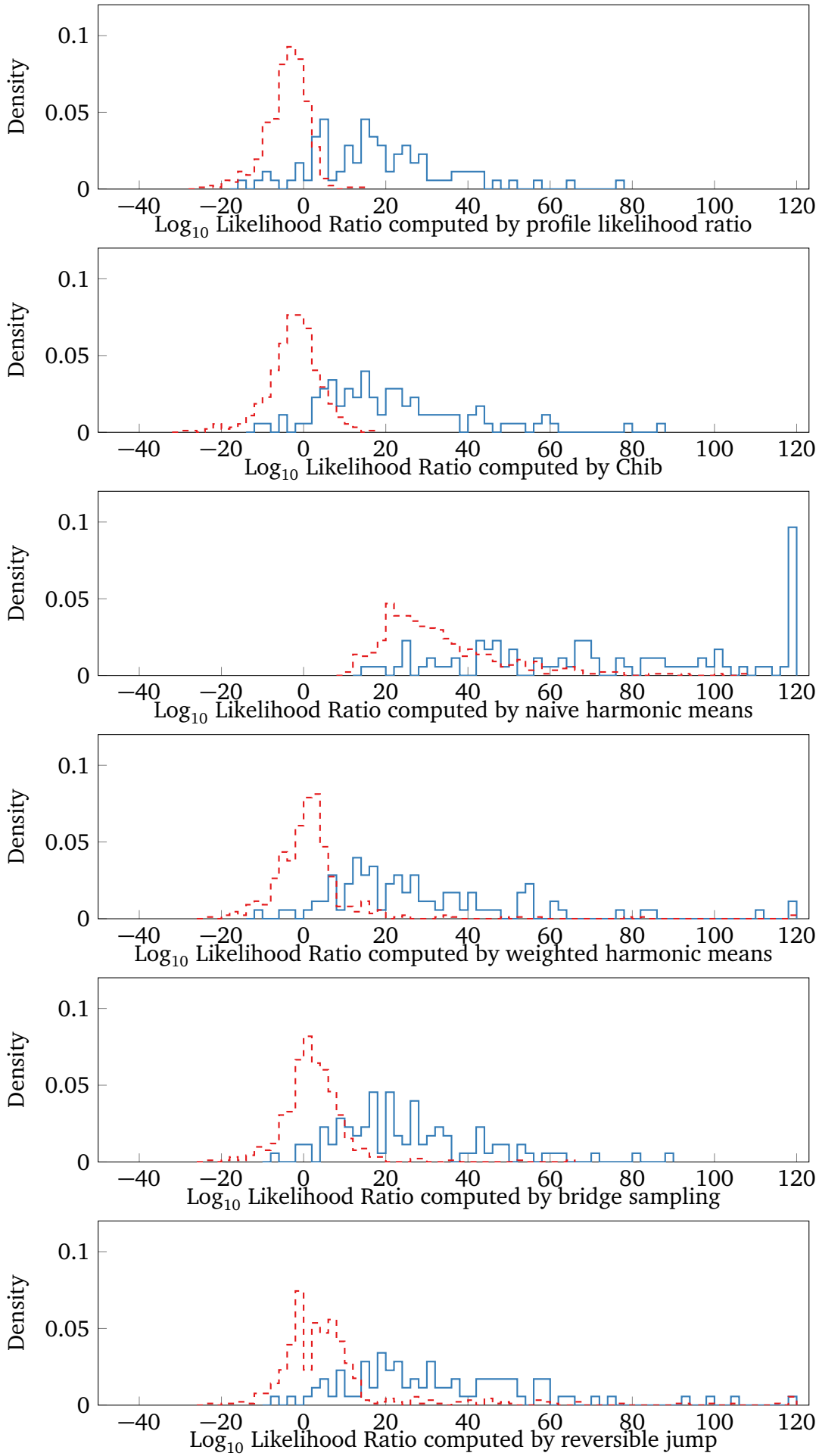


Figure 6.7: Histogram of the log-likelihood ratios computed on **only the good subset of** Garris and McCabe (2000). Log-likelihood ratios corresponding to false matches are dashed and red, and true matches are solid and blue.

Chapter Seven

Model validation

In this chapter we will attempt to validate or refute our modelling assumptions. Our observed data consists of 258 fingermark/fingerprint pairs (Garris and McCabe, 2000), $\{A_i, B_i : 1 \leq i \leq 258\}$. We augment the observed data with our manually-chosen matchings $\{\check{\xi}_i : 1 \leq i \leq 258\}$, described in §6.2. We can then use the bijection (2.23) and represent the observed data as the three independent marked Poisson point processes, $\{M_{103,i}, M_{013,i}, M_{113,i} : 1 \leq i \leq 258\}$.

All of the observed point sets are modelled as MPPPs with intensities proportional to complex normals. By estimating the unknown parameters in these intensity functions, we can use a similarity transformation on the observed MPPPs so the intensity functions become proportional to the standard complex normal:

$$\begin{gathered}
\widetilde{A}_i = \left\{ \left(\frac{r_a - \widehat{\tau}_{A,i}^{(H_d)}}{\widehat{\sigma}_{A,i}^{(H_d)}}, s_a, t_a \right) : a \in A_i \right\}, \quad \widetilde{B}_i = \left\{ \left(\frac{r_b - \widehat{\tau}_{B,i}^{(H_d)}}{\widehat{\sigma}_{B,i}^{(H_d)}}, s_b, t_b \right) : b \in B_i \right\}, \\
\widetilde{M_{103,i}} = \left\{ \left(\frac{r_a - \widehat{\tau}_{A,i}^{(H_p)}}{\widehat{\sigma}_{A,i}^{(H_p)}}, s_a, t_a \right) : a \in M_{103,i} \right\}, \\
\widetilde{M_{013,i}} = \left\{ \left(\widehat{\psi}_i^{(H_p)} \frac{r_b - \widehat{\tau}_{B,i}^{(H_p)}}{\widehat{\sigma}_{B,i}^{(H_p)}}, \widehat{\psi}_i^{(H_p)} s_b, t_b \right) : b \in M_{013,i} \right\}.
\end{gathered} \tag{7.1}$$

We similarly transform $M_{113,i}$ so that its projections $\Pi_A(\widetilde{M_{113,i}})$ and $\Pi_B(\widetilde{M_{113,i}})$ are

proportional to standard complex normals:

$$\widetilde{M_{113,i}} = \left\{ \left(\left(\frac{r_a - \widehat{\tau_{A,i}}^{(H_p)}}{\widehat{\sigma_{A,i}}^{(H_p)}}, s_a, t_a \right), \left(\widehat{\psi_i}^{(H_p)} \frac{r_b - \widehat{\tau_{B,i}}^{(H_p)}}{\widehat{\sigma_{B,i}}^{(H_p)}}, \widehat{\psi_i}^{(H_p)} s_b, t_b \right) \right) : (a,b) \in M_{113,i} \right\}. \tag{7.2}$$

The necessary parameter estimates are found by maximizing the likelihood of the observed MPPP. The estimates under H_d, $\widehat{\tau_{A,i}}^{(H_d)}, \widehat{\tau_{B,i}}^{(H_d)}, \widehat{\sigma_{A,i}}^{(H_d)}$ and $\widehat{\sigma_{B,i}}^{(H_d)}$, are given explicitly in § 3.1. The estimates under H_p are found using the iterative maximization procedure described in §3.2.

We will use the notation X_i to refer to a generic MPPP over $\mathbb{C}$ with intensity φ and marks on $\mathbb{S}^1 \times \mathbb{T}$. Thus any result derived for X_i is applicable to $\widetilde{A}_i, \widetilde{B}_i, \widetilde{M_{103,i}}, \widetilde{M_{013,i}}, \Pi_A(\widetilde{M_{113,i}})$, and $\Pi_B(\widetilde{M_{113,i}})$. We let $X_\bullet = \cup_{i=1}^{258} X_i$ be the union of the point sets $X_i, i = 1, \ldots, 258$.

We will validate our model by considering some of its testable consequences. Some of these follow directly from the model distributions:

- T1 Each $|X_i|$ has a Poisson distribution;
- T2 Conditional on $|X_i|$, the minutiae locations within each X_i have standard complex normal distributions;
- T3 The orientations are uniformly distributed on $\mathbb{S}^1$;
- T4 The location distortions $\{r_a - r_b : (a,b) \in \widetilde{M_{113,i}}\}$ have complex normal distributions;
- T5 The orientation distortions $\{s_a \overline{s_b} : (a,b) \in \widetilde{M_{113,i}}\}$ have von Mises distributions.

Other consequences follow from independence assumptions in our model:

- T6 The minutia locations within each X_i are independent (when combined with T1, this implies X_i is a Poisson point process);
- T7 $\widetilde{M_{103,i}}, \widetilde{M_{013,i}}$, and $\widetilde{M_{113,i}}$ are independent for each i (equivalently, the thinning of latent minutiae into observed and unobserved sets is independent of each minutia's location);
- T8 The orientations are independent of the minutia locations and other orientations;

T9 The minutia types $t_m \in \mathbb{T}$ are independent of minutia locations, orientations, and other types;

T10 The location distortions $\{r_a - r_b : (a,b) \in \widetilde{M_{113,i}}\}$ are independent of their locations and independent of other location distortions;

T11 The orientation distortions $\{s_a\overline{s_b} : (a,b) \in \widetilde{M_{113,i}}\}$ are independent of minutia locations, orientations, types, and all other distortions.

We shall attempt to test each of T1–T11 below. However, by testing each consequence independently we are implicitly assuming that the remainder of the model is correct. In particular, since we shall see strong evidence against T1–T5, our tests for independence T6–T11 must be viewed with suspicion.

§7.1 Test the number of observed points

Under both H_d and H_p we expect $n_A \mid \delta_A$ to have a Poisson distribution with mean $\rho_0\delta_A$ where $\delta_{A,i} \sim \text{Beta}(\alpha_\omega, \beta_\omega)$. The distribution of n_B is expected to be Poisson with mean $\rho_0\delta_B$ where δ_B is uniform on $(0,1)$. Figure 7.1 plots the empirical histograms against these theoretical distributions. These plots must be interpreted carefully. Recall that the distribution of δ_B was not estimated from the dataset, but was rather assumed to be uniform on $(0,1)$: this is because we do not wish to assume that our dataset is representative of future fingermarks. Thus the prior distribution over δ_B (and therefore also n_B) represents our subjective opinion, and we should not expect our subjective opinion of the distribution for n_B to match the empirical distribution of n_B. Conversely, the parameters for the distribution of δ_A were found by empirical Bayes, so we expect the theoretical distribution of n_A to roughly match the empirical distribution.

The value of n_ξ enters into the likelihood only under H_p. Because of the strong dependence of n_ξ with n_A and n_B, it is not fruitful to investigate the marginal distribution of n_ξ in a manner similar to figure 7.1. Rather, in order to investigate how well

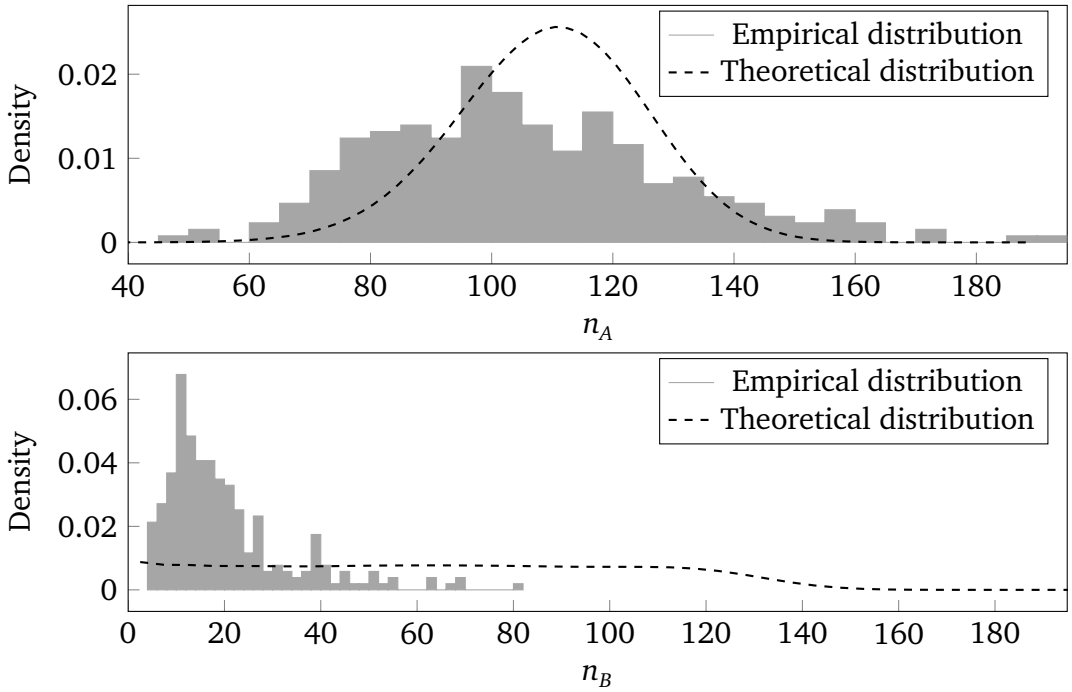


Figure 7.1: Histogram of n_A and n_B. The theoretical distributions under our model are plotted for reference.

the model fits the empirical distribution of n_ξ, we consider the statistic

$$\begin{aligned} T_i &= n_{A,i}n_{B,i} - (\rho_0+1)n_{\xi,i} \\ &= |M_{113,i}|^2 + (|M_{103,i}| + |M_{013,i}| - \rho_0 - 1)|M_{113,i}| + |M_{103,i}||M_{013,i}|, \end{aligned} \tag{7.3}$$

where in the second line we use $|M_{103,i}| = n_{A,i} - n_{\xi,i}, |M_{013,i}| = n_{B,i} - n_{\xi,i}$, and $|M_{113,i}| = n_{\xi,i}$. Recalling that $|M_{103,i}|, |M_{013,i}|$, and $|M_{113,i}|$ have independent Poisson distributions with means $\rho_0\delta_{A,i}(1-\delta_{B,i}), \rho_0\delta_{B,i}(1-\delta_{A,i})$, and $\rho_0\delta_{A,i}\delta_{B,i}$ respectively, we see that T_i has mean zero. A positive T_i implies that we have observed fewer matched minutiae than expected under our model, while a negative T_i implies we have observed more matched minutiae than expected.

A lengthy but straight-forward calculation shows that the conditional variance is

$$\begin{aligned} \mathbb{V}(T_i \mid \delta_{A,i}, \delta_{B,i}) &= \mathbb{E}(T_i^2 \mid \delta_{A,i}, \delta_{B,i}) \\ &= \rho_0^2\delta_{A,i}\delta_{B,i}\{1 + \delta_{A,i}\delta_{B,i} + \rho_0(1 + 2\delta_{A,i}\delta_{B,i} - \delta_{A,i} - \delta_{B,i})\}. \end{aligned} \tag{7.4}$$

Using the law of total variance we see that the variance of T_i is

$$\mathbb{V}(T_i) = \mathbb{E}\{\mathbb{V}(T_i \mid \delta_{A,i}, \delta_{B,i})\} = \frac{\alpha_\omega\rho_0^2}{6(\alpha_\omega+\beta_\omega)}\left\{\rho_0 + 3 + \frac{(\alpha_\omega+1)(\rho_0+2)}{\alpha_\omega+\beta_\omega+1}\right\}. \tag{7.5}$$

Thus the random variables $\{\tilde{T}_i = T_i/\sqrt{\mathbb{V}(T_i)} : 1 \leq i \leq 258\}$ should be independent with mean zero and unit variance under H_p. The histogram of $\tilde{T}_i$ is shown in figure 7.2. We see that the empirical distribution of $\tilde{T}_i = T_i/\sqrt{\mathbb{V}(T_i)}$ has a slightly shorter left tail than the theoretical distribution, which implies that we observe more matched minutiae than expected. However, the discrepancy is small, and we do not expect it to significantly hamper our model's ability to explain the data.

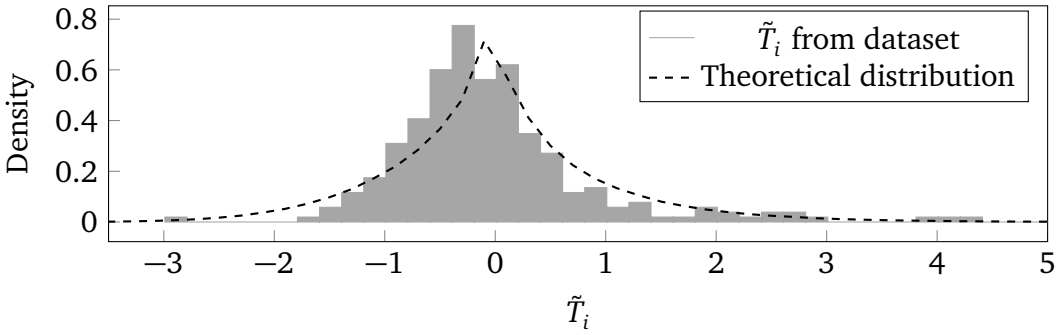
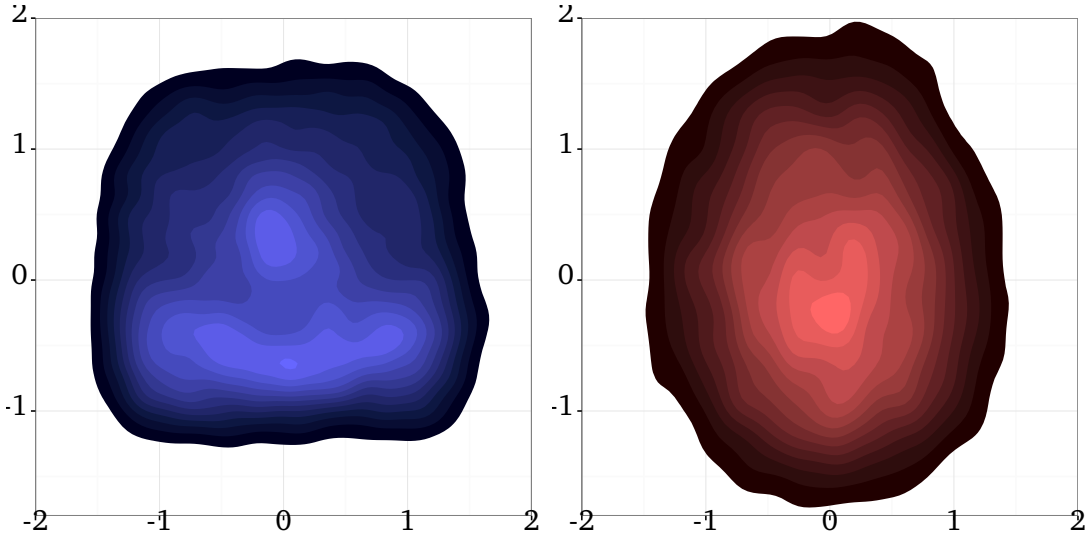


Figure 7.2: Histogram of $\tilde{T}_i \propto n_{A,i}n_{B,i} - (\rho_0 + 1)n_{\xi,i}$. Under our model these should have mean zero and unit variance. The theoretical model distribution is plotted for reference; it was obtained numerically.

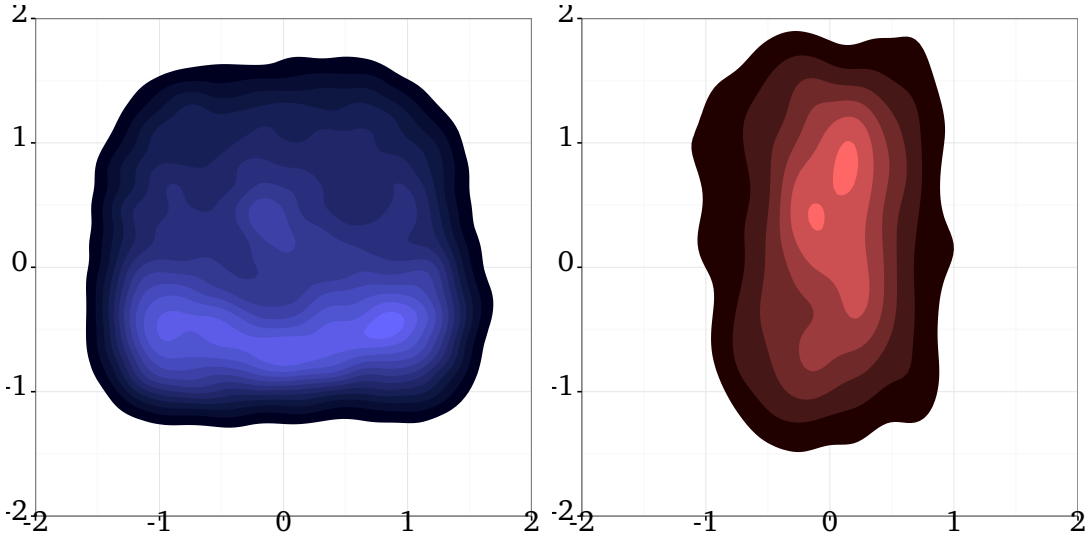
§7.2 Test the location distributions

We have assumed the latent intensity function is proportional to a standard complex normal density. When combined with the independent uniform thinning assumption, we expect all of $\widetilde{A}_i, \widetilde{B}_i, \widetilde{M_{103,i}}, \widetilde{M_{013,i}}, \Pi_A(\widetilde{M_{113,i}})$, and $\Pi_B(\widetilde{M_{113,i}})$ to have intensities proportional to standard complex normals.

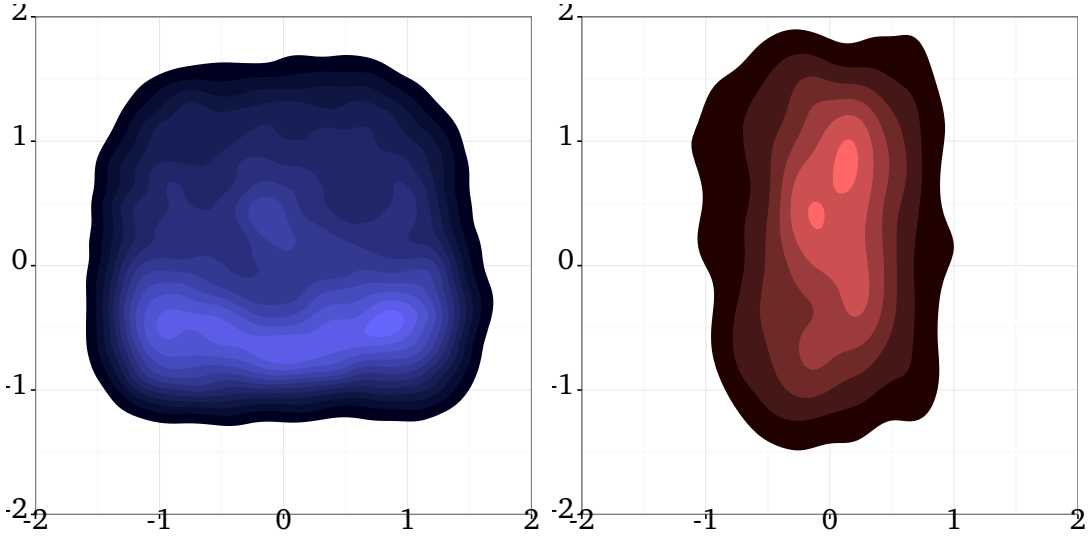
We can get some intuition about the empirical intensity functions through examining figure 7.3, which plots the empirical intensities of $X_\bullet$ as heat maps. Note in particular that the intensities of $\widetilde{A}$ and $\widetilde{M_{103}}$ are multimodal, with a horizontal line of high intensity below the centre mode. The remaining intensities are unimodal, but all of them lack the expected circular symmetry. Furthermore, the empirical intensity for $\Pi_A(\widetilde{M_{113}})$ (figure 7.3(e)) has shorter tails than the unmatched points $\widetilde{M_{103}}$ (figure 7.3(c)). This implies that the matched minutiae tend to come from the centre of the finger rather than the edges, which contrasts with our assumption T7.



(a) $\widetilde{A}$ (b) $\widetilde{B}$



(c) $\widetilde{A}$ (d) $\widetilde{B}$



(e) $\widetilde{A}$ (f) $\widetilde{B}$

Figure 7.3: Heat maps of the empirical intensity functions. These were obtained by taking the observed point sets $\widetilde{A}_\bullet, \widetilde{B}_\bullet, \widetilde{M_{103\bullet}}, \widetilde{M_{103\bullet}}, \widetilde{M_{013\bullet}}, \Pi_A(\widetilde{M_{113\bullet}}), \Pi_B(\widetilde{M_{113\bullet}})$, and computing kernel density estimates.

The heat maps do not provide much information about the empirical intensity near the peak. To gain some insight, we consider the scaled squared distance from the observed points to the origin, $2|r|^2$, in figure 7.4. We expect these to follow independent chi-squared distributions with two degrees of freedom (equivalently, exponential distributions with mean two). The empirical distribution of the distances in $\widetilde{A}$ and $\widetilde{B}$ (the top plot in figure 7.4) have shorter and broader peaks than expected.

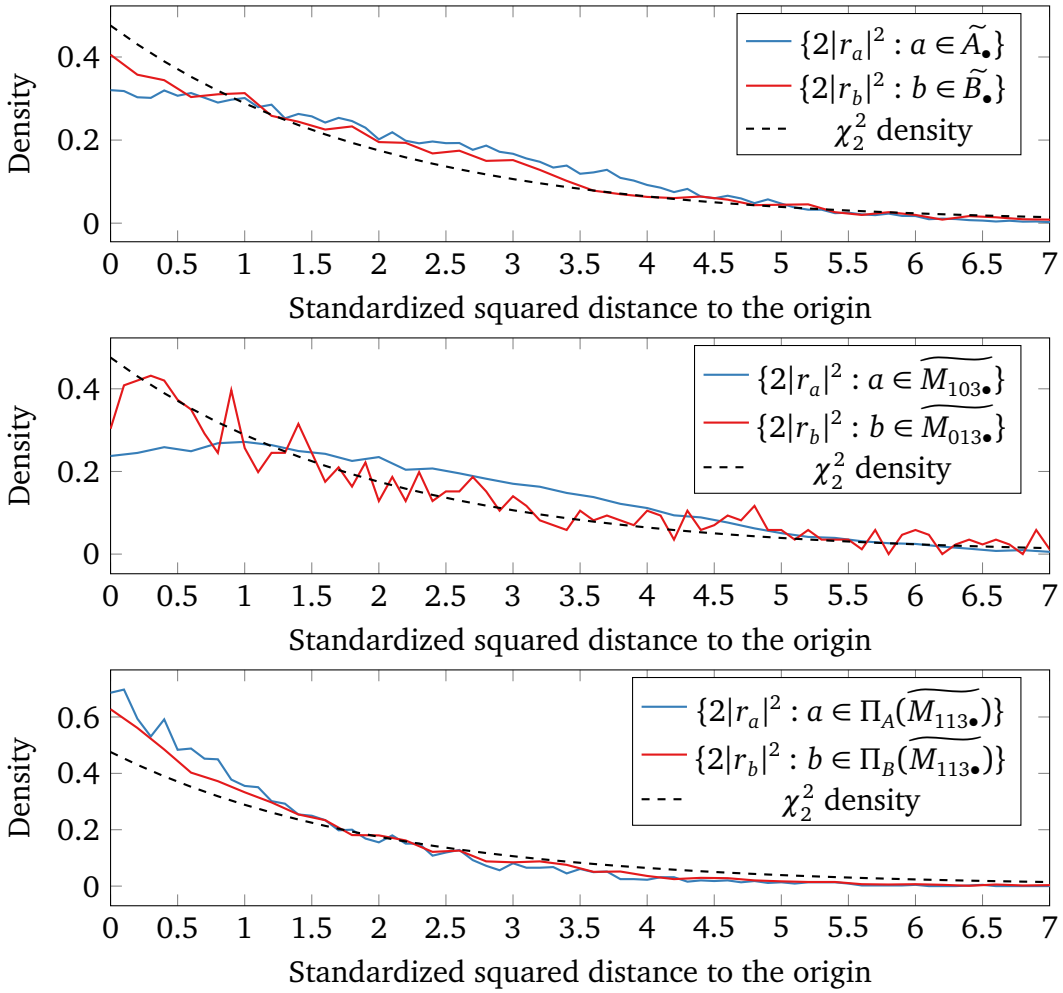


Figure 7.4: Density estimate of the transformed squared distances from the observed points to the origin.

Conversely, the bottom plot in figure 7.4 shows the minutiae which are observed in both the fingerprint and fingermark are more narrowly peaked than expected. This is likely because the minutiae which are observed in the fingermark tend to come from the centre of the finger rather than the edges. This hypothesis is consistent with the middle plot, which shows that the minutiae observed in the fingerprint alone

are broadly distributed about the origin, but the points which are observed in the fingermark alone are concentrated about the origin.

In summary, the model's intensity functions do not adequately describe the observed distributions of the minutia locations. We believe that this discrepancy has a large effect on the the value of the computed likelihood ratios. One possible solution is to allow the probability of observing a latent minutia to depend on that minutia's position. The spatially-dependent probabilities $\delta_A(r)$ and $\delta_B(r)$ could potentially be modelled by some stochastic field over $\mathbb{C}$ such as a Gaussian process.

§7.3 Test the orientation distributions

Ideally we want to test whether the latent minutiae orientations are uniformly distributed over $\mathbb{S}^1$. Since the latent minutiae are unobserved, we will resort to testing whether the fingerprint minutia orientations $\{s_a : a \in \widetilde{A}_\bullet\}$ are uniformly distributed. This test is a good proxy under the reasonable assumption that a minutia's orientation is independent of whether that minutia is observed in a fingerprint. Figure 7.5 shows a histogram of the orientations in $\widetilde{A}_\bullet$.

We notice that there are several minor peaks, and one significant valley when the phase is approximately -1.5. These peaks and valleys may be caused by the tendency for fingerprint ridges to form "loop" patterns, as shown in figure 7.6. These patterns create preferred directions which contrasts with our model's rotation invariance assumption. It appears that minutia configurations do in fact have a preferred direction for their orientations.

To further investigate this preferred direction, we compute the Rayleigh statistic (Mardia and Jupp, 1999, p. 94) for each fingerprint minutia configuration:

$$\frac{1}{n_{A,i}}\left|\sum_{a\in A_i} s_a\right|^2, \quad 1 \le i \le 258. \tag{7.6}$$

The Rayleigh statistic is large when the orientations are clustered and small when the orientations are uniformly distributed. If the orientations are uniform, then for

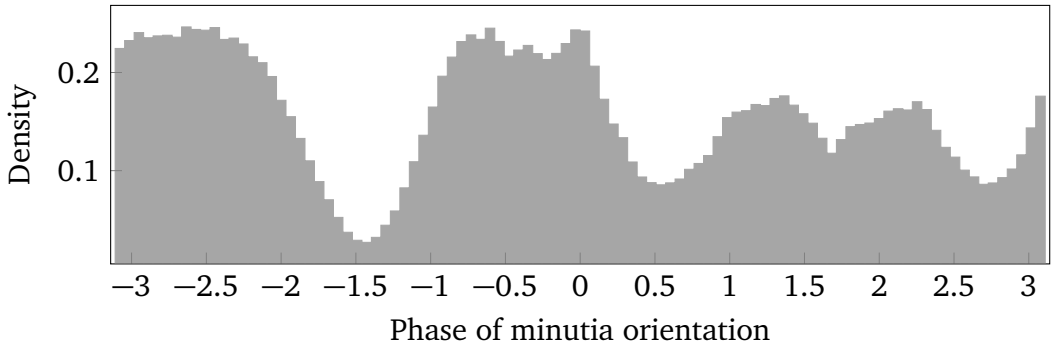
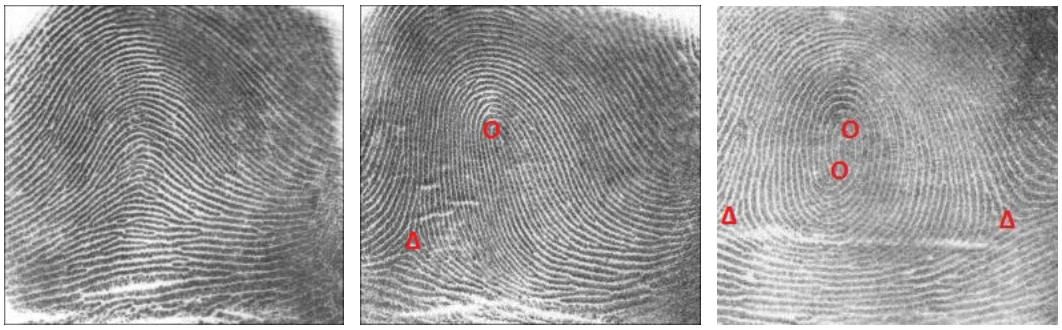


Figure 7.5: Histogram of the phase of minutia orientations over $\widetilde{A}_\bullet$.



(a) (b) (c)

Figure 7.6: An arch, loop and whorl fingerprint. So-called singular points in the ridge curvatures are marked in red (Maltoni et al., 2009). The circles are points called *cores* and the triangles are points called *deltas*; both types of singular points create preferred directions for the minutia orientations. Images from Garris and McCabe (2000).

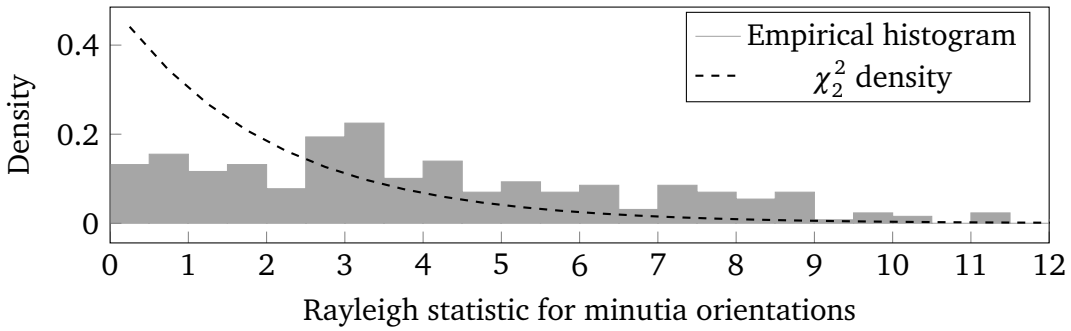


Figure 7.7: Histogram of the values of the Rayleigh statistic (7.6) for each $\widetilde{A}_i, 1 \leq i \leq 258$.

large $n_{A,i}$, the distribution of the Rayleigh statistic is approximately chi-squared with two degrees of freedom. Figure 7.7 shows a histogram of our Rayleigh statistics. The empirical statistics are much larger than expected, which gives further evidence that the orientations are not uniform, but rather have a location-dependent preferred direction. Thus the minutia orientations have less information (in the information theoretic sense, see e.g. Shannon (1948)) than our model assumes, and hence

our model may produce larger likelihood ratios than justified by the data. We will investigate alternative models for our orientations in §8.1.1.

§7.4 Test for the location distortion distributions

Given $(a,b) \in \widetilde{M_{113,i}}$ for some i, let

$$d(a,b) = \frac{r_a - r_b}{\sqrt{2-2\sqrt{1-1/\tilde{\omega}_i}}} = \left(\frac{1+\omega^{-2}}{2}\right)^{1/2}(r_a - r_b) \tag{7.7}$$

be the standardized distortion between a and b. Note that the computed values of $d(a,b)$ depend on the manually-chosen matchings $\check{\xi}_i$, and therefore the following analysis implicitly assumes those manual matchings are correct.

Under our model these distortions have independent standard complex normal distributions. Figure 7.8 shows a heat map of the empirical density: it is circularly symmetric, but it has much longer tails than a standard normal.

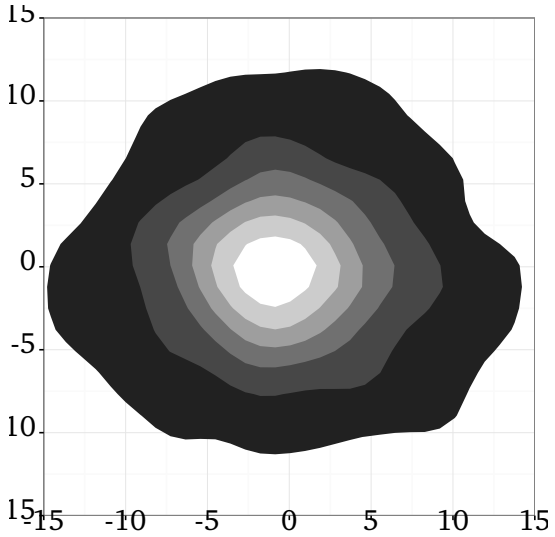


Figure 7.8: Heat map of the empirical distortion density. This was obtained by a kernel density estimate on the observed distortions (7.7), $\{d(a,b) : (a,b) \in \widetilde{M_{113\bullet}}\}$.

This is further confirmed by figure 7.9, which shows a histogram of $\{2|d(a,b)|^2 : (a,b) \in \widetilde{M_{113\bullet}}\}$. According to our model this should follow a chi-squared distribution with two degrees of freedom, and therefore have mean 2. In fact the mean is 255.

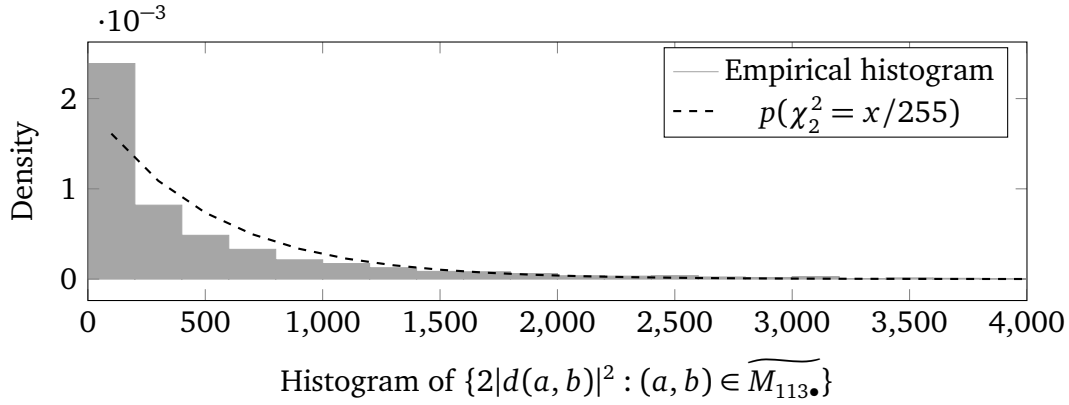


Figure 7.9: Density estimate of the standardized squared magnitude of the distortions. Under our model these should follow a χ^2_2 density, but they are off by a factor ≈ 255.

An appropriately rescaled chi-squared density matches the empirical histogram well in the tails, but has a much wider peak than the empirical histogram.

We can conclude that our distortions are circularly symmetric, but they are not Gaussian, and our model drastically underestimates their variance ω^2. This underestimation of the distortion in the minutia locations is likely a primary cause of the inflated likelihood ratios observed in chapter 6.

Our model's underestimation of the variance is likely caused by an overestimation of c_ω, which serves as a lower bound on $\tilde{\omega}$ and hence an upper bound on ω^2, in §6.2. This in turn is likely because our distortion model does not adequately describe the observed location distortions in our dataset. In particular, our model assumes that the location distortions are independent, but the observed location distortions are highly correlated. We describe one way of extending our model to allow for correlation location distortions in chapter 8.

§7.5 Test for the orientation distortion distributions

The orientation distortions $\{s_a\overline{s_b} : (a,b) \in \widetilde{M_{113,i}}\}$ are modelled with a von Mises distribution with mean zero and concentration parameter κ_i, which has density (4.4) with $\alpha_\kappa = 4.4$ and $\beta_\kappa = 0.087$. By integrating over κ we can find the distribution for $\{s_a\overline{s_b} : (a,b) \in \widetilde{M_{113\bullet}}\}$. This is plotted alongside the empirical histogram in figure 7.10.

The observed data fits the model well: we see little evidence against our model.

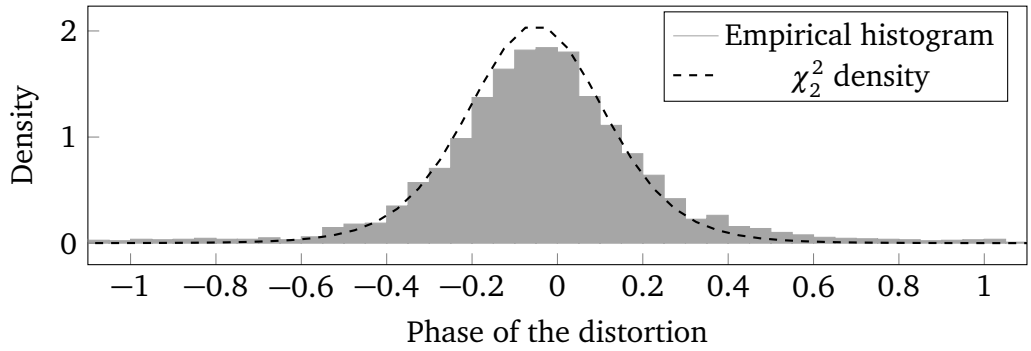


Figure 7.10: Histogram of the phase of distortions $\{s_a\overline{s_b} : (a,b) \in \widetilde{M_{113\bullet}}\}$.

§7.6 Test the Poisson point process model

We will now test whether there is evidence against the Poisson point process model for each of the observed minutia configurations. This is equivalent to testing if the minutia positions within a configuration are independent. Given a generic MPPP X_i with intensity ρ_X, we will consider the test statistic

$$K(u,X_i) = \sum_{x\in X_i} \sum_{y\in X_i\setminus\{x\}} k(u;x,y), \tag{7.8}$$

where $k(u;x,y)$ is some symmetric kernel function. One such kernel is

$$k(u;x,y) = \frac{\mathbb{1}(|r_x - r_y| \le u)\ \mathbb{1}(r_x \in V)\ \mathbb{1}(r_y \in V)}{|V|\rho_X(r_x)\rho_X(r_y)}, \tag{7.9}$$

where $V \subset \mathbb{C}$ has finite area $|V|$. In this case $K(u,X_i)$ becomes a variant of the *inhomogeneous K-function* (Baddeley et al., 2000) and, ignoring boundary terms for V, $\mathbb{E}K(u,X_i) = \pi u^2$. However, the variance of the inhomogeneous K-function blows up if ρ_X tends to zero at any point in the region V. This makes it a poor choice for our purposes, since our intensity function $\rho_X = \varphi$ tends to zero away from the origin.

We assume that $\rho_X(r) = \rho_{X0}\varphi(r)$ for some $\rho_{X0} > 0$ and use the kernel $k(u;x,y) = \pi\varphi\{(r_x - r_y)/u^2\}$. Instead of considering $K(u,X_i)$ directly we shall consider the standardized statistic

$$\widetilde{K}(u,X_i) = \frac{K(u,X_i) - \mathbb{E}_{X_i}\{K(u,X_i)\,|\,n_{X,i}\}}{\sqrt{\mathbb{V}\{K(u,X_i)\,|\,n_{X,i}\}}}, \tag{7.10}$$

where $n_{X,i} = |X_i|$. Whereas $K(u, X_i)$ depends heavily on the the number of observed points, this standardized statistic controls for the number of observed points and instead focuses on whether those points occur at locations consistent with a Poisson point process model. Using the results derived in §B.5 we have

$$\begin{aligned} \mathbb{E}_{X_i}\{K(u,X_i) \,|\, n_{X,i}\} &= \frac{n_{X,i}(n_{X,i}-1)u^2}{u^2+2}, \\ \mathbb{V}\{K(u,X_i) \,|\, n_{X,i}\} &= \frac{4n_{X,i}(n_{X,i}-1)u^2\{u^4+(n_{X,i}+2)u^2+3\}}{(u^2+1)(u^2+2)^2(u^2+3)}. \end{aligned} \tag{7.11}$$

These results were found under the assumption that ρ_X is proportional to a standard complex normal density. In §7.2 we observed evidence against this assumption, so we must view (7.11), as well as the subsequent results, with suspicion.

If $\widetilde{K}(u, X_i)$ is positive there is evidence that the points of X_i are clustered (perhaps due to an attraction between points). Conversely if $\widetilde{K}(u, X_i)$ is negative there is evidence that the points of X_i are more regularly spaced than expected for a Poisson point process (perhaps due to a repulsion between points).

In figure 7.11 we plot $\widetilde{K}(u, X_i)$ for each point set

$$\{\widetilde{A}_i, \widetilde{B}_i, \widetilde{M_{103,i}}, \widetilde{M_{013,i}}, \Pi_A(\widetilde{M_{113,i}}), \Pi_B(\widetilde{M_{113,i}}) : 1 \le i \le 258\}. \tag{7.12}$$

We notice that many of the observed point processes show evidence of repulsion at $u \approx 0.1$ and attraction at $u \approx 0.4$.

Most strikingly, figure 7.11 shows strong evidence of clustering in $\widetilde{M_{113}}$. This effect is caused by a discrepancy between our model and the actual data: the model assumes that the probability a latent minutia is observed is independent of its location, but the fingermark is often only observed on a subregion of the finger (see figure 7.12 for an illustration). As in §7.3, this clustering effect means that the minutia locations are less informative (in the information theoretic sense) than our model assumes, which in turn causes the computed likelihood ratios to be larger than justified by the data. As in §7.2, the best solution might be to allow the probability of a minutia being observed in the fingermark δ_B to depend on the minutia's location.

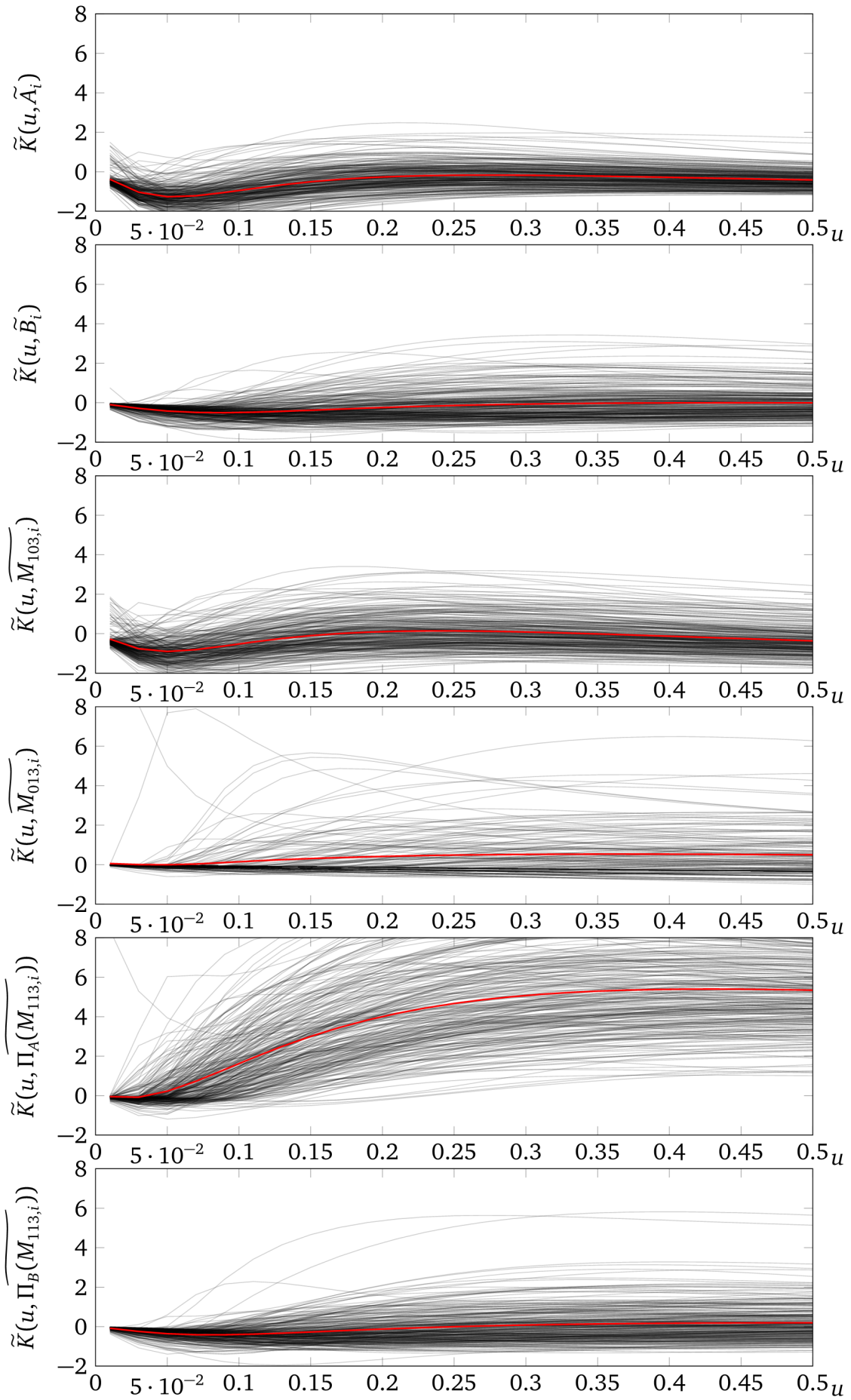


Figure 7.11: Values of $\widetilde{K}(u,X_i)$, as defined in (7.10), for each standardized observed point set. The red line shows the average value for each u.

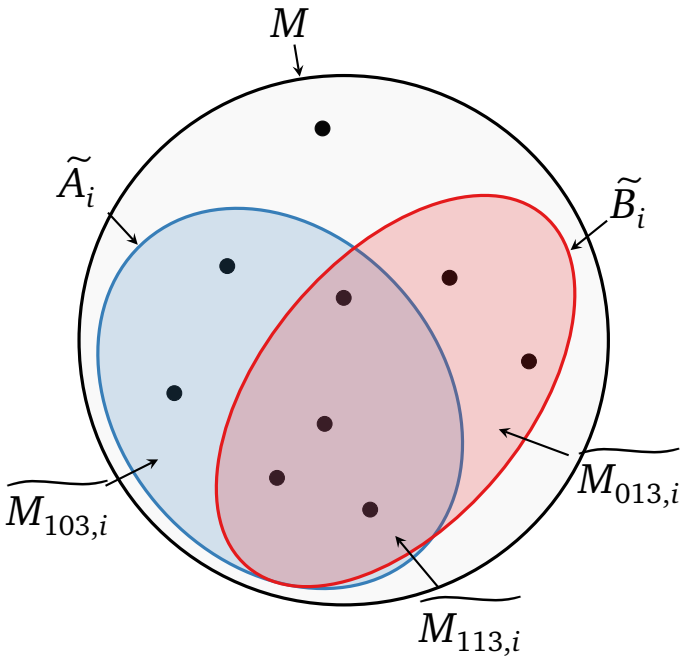


Figure 7.12: Example of regional thinning. The model expects the points of $\widetilde{M_{113,i}}$ to be uniformly distributed over the latent finger M. However, the points only occur in the intersecting purple section. Thus the statistic $\widetilde{K}(u, \widetilde{M_{113,i}})$ indicates strong clustering of these points.

§7.7 Test the thinning model

In §7.2 we observed that minutiae near the centre of a finger are more likely to be observed in the fingermark than minutiae near the edge of a finger. We also observed that a Gaussian intensity is not a good fit for the data. Thus it makes sense to consider more generic functions $\rho(r), \delta_A(r), \delta_B(r)$ which depend on the minutia locations. However, since the latent finger is unobserved, it is impossible to distinguish the actual form of the latent intensity function $\rho(r)$ from the form of the thinning probability functions $\delta_A(r), \delta_B(r)$.

Faced with this non-identifiability, we cannot fruitfully investigate the thinning model without an adequate model for the latent intensity. In §8.1.2 we discuss one potential solution: let the thinning probabilities be smooth functions drawn from some random field such as a Gaussian process, and investigate the form of the latent intensity function by marginalizing these random fields.

§7.8 Test for orientation independence

We have already seen evidence against a uniform distribution for minutia orientations in §7.3. There are two ways to account for this discrepancy: we can allow the distribution of a minutia's orientation to depend on that minutia's location, or we can allow minutiae with nearby locations to have correlated orientations. In the absence of a better model for the orientation distributions we cannot distinguish between these two cases.

In this section we will assume that the minutia orientations have marginal uniform distributions and that the non-uniformity of the observed orientations is due to correlations in the orientations of nearby minutiae. As in §7.3 we will restrict our attention to $\widetilde{A}_i$, which serves as a proxy for the latent minutia configuration M.

Examination of the fingerprint in figure 1.1(c) on page 5 suggests that nearby minutiae often have either the same orientation or polar opposite orientations. In either case, the squared orientations $\{s_a^2 : a \in \widetilde{A}_i\}$ will be positively correlated for nearby minutia. Under our model we assume that each orientation has a marginal uniform distribution, which implies the squared orientations also have marginal uniform distributions.

For each minutia $a \in \widetilde{A}_i$, define $S(u,a)$ to be the weighted average of the squared-orientations of its neighbours:

$$S(u,a) = \frac{\sum_{x \in \widetilde{A}_i \setminus \{a\}} \varphi\{(r_a - r_x)/u^2\} s_x^2}{\left| \sum_{x \in \widetilde{A}_i \setminus \{a\}} \varphi\{(r_a - r_x)/u^2\} s_x^2 \right|} \in \mathbb{S}^1. \tag{7.13}$$

The value of u determines the radius of the area which contributes most to the sum. If the orientations are independent then $S(u,a)$ is independent of s_a^2 for all $a \in \widetilde{A}_i$, and hence $\mathrm{Corr}_{ang}\{s_a, S(u,a)\}$, defined by (Jammalamadaka and Sarma, 1988)

$$\mathrm{Corr}_{ang}(x,y) = \frac{\mathbb{E}\{\mathrm{Im}(x)\,\mathrm{Im}(y)\}}{\sqrt{\mathbb{E}\{\mathrm{Im}(x)^2\}\,\mathbb{E}\{\mathrm{Im}(y)^2\}}}, \tag{7.14}$$

has expectation zero. In the above $\mathrm{Im}(x) = (x - \overline{x})/2$ is the imaginary part of $x \in \mathbb{C}$.

In figure 7.13 we plot the empirical covariance over $\widetilde{A}_\bullet$,

$$\widehat{\mathrm{Corr}_{ang}}(u;S) = \frac{\sum_{a\in\widetilde{A}_\bullet} \mathrm{Im}(s_a)\,\mathrm{Im}\{S(u,a)\}}{\sqrt{\left\{\sum_{a\in\widetilde{A}_\bullet} \mathrm{Im}(s_a)^2\right\}\left[\sum_{a\in\widetilde{A}_\bullet} \mathrm{Im}\{S(u,a)\}^2\right]}}. \tag{7.15}$$

Clearly there is a strong correlation in the orientations for minutiae separated by less than 0.5, with decreasing correlations for minutiae further apart. However, at least part of this correlation can be explained by the preferred orientation directions which were observed in §7.3. Thus we expect that, provided the minutia orientations themselves are accurately modelled, our current model of independent orientation distortions should provide an adequate fit for the observed data.

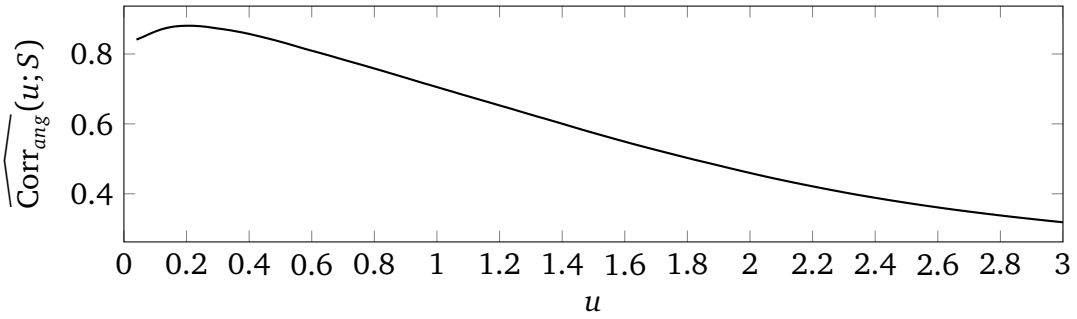


Figure 7.13: Correlation of the squared orientation with its neighbouring orientations as a function of the neighbourhood size u (7.15).

§7.9 Test for type independence

To test whether minutia types are spatially correlated, we will consider the weighted average of the proportion of nearby minutiae with the same type. For each $x \in X_i$ with an observed type $t_x \neq 0$, let

$$T(u,x) = \frac{\sum_{y\in X_i\setminus\{x\}} \varphi\{(r_x - r_y)/u^2\}\ \mathbb{1}(t_x = t_y)}{\sqrt{\chi^{|t_x|+t_x}(1-\chi)^{|t_x|-t_x}}\sum_{y\in X_i\setminus\{x\}} \varphi\{(r_x - r_y)/u^2\}\ \mathbb{1}(t_y \neq 0)}. \tag{7.16}$$

Note that $T(u,x)$ implicitly depends on which point set X_i contains x. This is defined as long as at least one minutia in $X_i \setminus \{x\}$ has an observed type. Under the assumption that the types in X_i are independent, the expectation of $T(u,x)$ is one. If the types of nearby minutia are correlated we expect $T(u,x) > 1$.

In our dataset (Garris and McCabe, 2000) most of the fingermark minutia $b \in \widetilde{B}_\bullet$ are classified, but almost all of the fingerprint minutiae $a \in \widetilde{A}_\bullet$ are unclassified. Figure 7.14 plots the average value of $T(u, b)$ over all fingermarks:

$$T(u) = \frac{\sum_{b \in \widetilde{B_\bullet}} \mathbb{1}(t_b \neq 0) T(u, b)}{\sum_{b \in \widetilde{B_\bullet}} \mathbb{1}(t_b \neq 0)}. \tag{7.17}$$

Figure 7.14 implies there is a slight tendency for nearby minutiae to be of the same type. This may be due the tendency of ridge bifurcations to occurs near regions of high ridge curvature called *deltas* (see figure 7.6). However, the overall effect seems to be fairly small, and $T(u)$ is less than 1.07 for all u. Thus the assumption that the types are independent of location seems reasonable.

We have not investigated the joint dependence structure between a minutia's location, orientation, and type. It seems likely that nearby minutiae with similar orientations will be of the same type, whereas nearby minutiae with different orientations may be of different types. In order to test this we must first find a better model for the minutia locations, orientations and types. We hope to address this in the future.

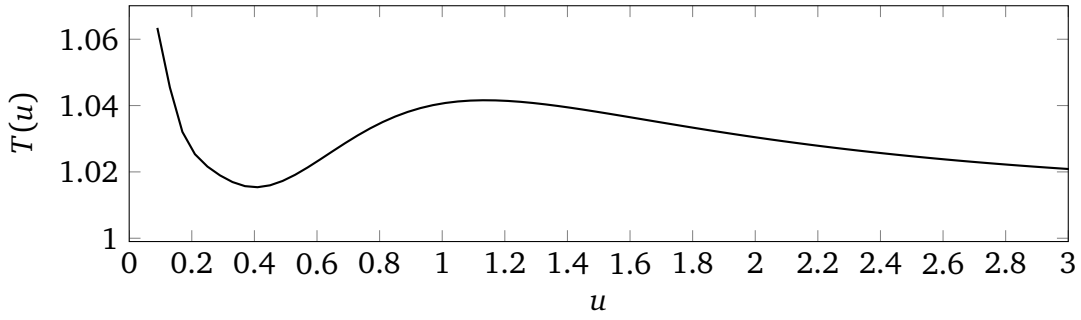


Figure 7.14: Similarity between neighbouring types for neighbourhood size u, as measured by the statistic $T(u)$ (7.17).

§7.10 Test for location distortion independence

We will now investigate the independence assumption of the distortions $d(a, b)$, as defined in (7.7). For each $(a, b) \in \widetilde{M_{113,i}}$, let

$$D(u, a, b) = \frac{\sum_{(a', b') \in \widetilde{M_{113,i}} \setminus \{(a,b)\}} \varphi\{(r_a - r_{a'})/u^2\} d(a', b')}{\sqrt{\sum_{(a', b') \in \widetilde{M_{113,i}} \setminus \{(a,b)\}} \varphi\{(r_a - r_{a'})/u^2\}^2}} \tag{7.18}$$

be the weighted average of the distortions of points in $\widetilde{M_{113,i}}$ near (a,b). We compute the empirical correlation matrix between $d(a,b)$ and $D(u,a,b)$,

$$\Sigma_D(u) = \begin{pmatrix} \Sigma_D^{11}(u) & \Sigma_D^{12}(u) \\ \Sigma_D^{12}(u) & \Sigma_D^{22}(u) \end{pmatrix} = \sum_{(a,b)\in\widetilde{M_{113\bullet}}} \begin{pmatrix} \overline{d(a,b)} - \overline{d(\bullet)} \\ \overline{D(u,a,b)} - \overline{D(u,\bullet)} \end{pmatrix}^\top \begin{pmatrix} d(a,b) - d(\bullet) \\ D(u,a,b) - D(u,\bullet) \end{pmatrix}, \tag{7.19}$$

where $d(\bullet)$ and $D(u,\bullet)$ are the average of $d(a,b)$ and $D(u,a,b)$ over $\widetilde{M_{113\bullet}}$.

The expectation of the off-diagonals of $\Sigma_D(u)$ is zero for all u. In figure 7.15 we plot the empirical absolute correlation $|\Sigma_D^{12}(u)|/\sqrt{\Sigma_D^{11}(u)\Sigma_D^{22}(u)}$. There is clearly a strong correlation in the distortions of nearby minutiae. As mentioned in §7.4, this correlation is likely a primary reason for why our model underestimates the variance of the location distortions ω^2, which in turn leads to inflated likelihood ratios.

Future work will investigate a better model for location distortion. One solution may be to use a Gaussian process rather than the current independent Gaussian distortions.

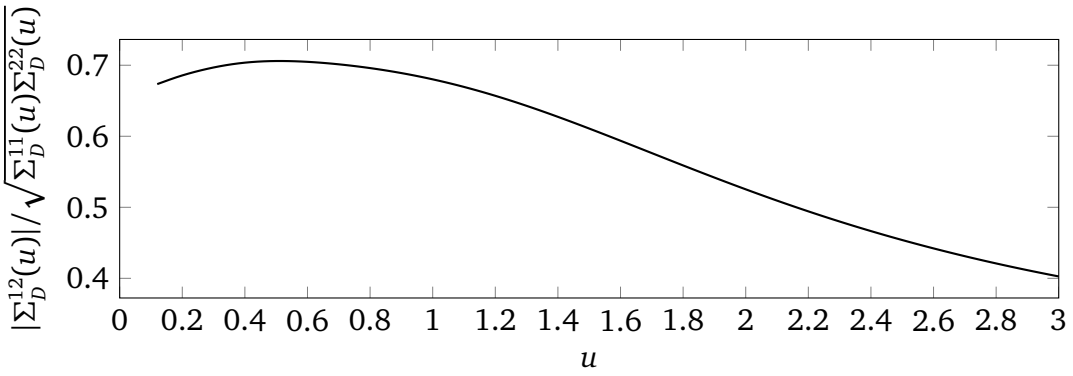


Figure 7.15: Correlation between the observed location distortions $d(a,b)$ and the average distortions in their neighbourhood $D(u,a,b)$ (7.19), as a function of the neighbourhood size u.

§7.11 Test for orientation distortion independence

For each $(a,b) \in \widetilde{M_{113,i}}$, let

$$S_D(u,a,b) = \frac{\sum_{(a',b')\in\widetilde{M_{113,i}}\setminus\{(a,b)\}} \varphi\{(r_a - r_{a'})/u^2\} s_a \overline{s_b}}{\left|\sum_{(a',b')\in\widetilde{M_{113,i}}\setminus\{(a,b)\}} \varphi\{(r_a - r_{a'})/u^2\} s_a \overline{s_b}\right|} \in \mathbb{S}^1 \tag{7.20}$$

be the average orientation distortion of nearby points. We compute the empirical angular correlation between $s_a\overline{s_b}$ and $S_D(u,a,b)$,

$$\widehat{\text{Corr}_{ang}}(u;S_D) = \frac{\sum_{(a,b)\in\widetilde{M_{113\bullet}}} \text{Im}(s_a\overline{s_b})\,\text{Im}\{S_D(u,a,b)\}}{\sqrt{\left\{\sum_{(a,b)\in\widetilde{M_{113\bullet}}} \text{Im}(s_a\overline{s_b})^2\right\}\left[\sum_{(a,b)\in\widetilde{M_{113\bullet}}} \text{Im}\{S_D(u,a,b)\}^2\right]}}, \quad (7.21)$$

and plot it in figure 7.16. There is a strong correlation between orientation distortions that does not significantly decrease with distance. We will investigate the source of this correlation in future work.

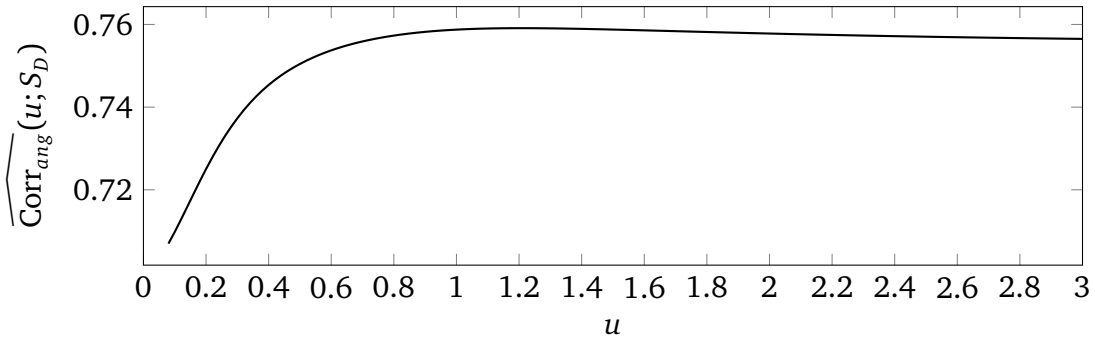


Figure 7.16: Correlation between neighbouring orientation distortions, (7.20), as a function of the neighbourhood size u.

§7.12 Conclusions

We have found substantial evidence against our model. In particular, there is strong evidence against T2, T3, T7, T8, T10, and T11. Each of these model violations could lead to an overvaluation of the amount of information in the observed data, and hence to an inflated likelihood ratio. The model would benefit from a more sophisticated model for the latent minutia locations and distortions, a location-dependent model for the probability that a latent minutia is observed in the fingerprint or fingermark, and a spatially-correlated model for the location and orientation distortions. We will briefly explore these enhancements in §8.1.

The combined effect of the modelling errors means that the likelihood ratios computed in chapter 6 are inflated and cannot be used as an honest measure of the evidence. However, despite the poor model fit, the ROC curves in figure 6.4

demonstrate that our model can often distinguish matching minutia configurations from non-matching configurations. Thus, despite our model's flaws, it seems suitable for the purpose of classifying matches as true or false. With some calibration the model's computed likelihood ratios may be useful to quantify the strength of evidence for H_p against H_d. We investigate this calibration in chapter 9.

Chapter Eight

Future work

§8.1 Model enhancements

In chapter 7 we noticed several parts of our model which conflicted with observations. In this chapter we will explore some ways to extend the model to better fit the data.

§8.1.1 Alternative latent minutia distribution

Our model assumes that the latent minutiae follow a Poisson process with the intensity function $\rho(r) = \rho_0\varphi(r)$ and a uniform distribution over minutia orientations. Figure 7.3 and figure 7.5 show that these assumptions are not supported by the data. We discuss a different model based on modelling the fingerprint ridges here.

Every finger has a small number of regions with high ridge curvature called *critical points*. The two types of critical points observed in human fingers are called *cores* (the centres of concentric ridge loops) and *deltas* (points where ridges in three different directions merge). All fingers have an equal number of cores and deltas, and the vast majority of fingers have two or fewer core-delta pairs. Examples of cores and deltas are shown in figure 7.6.

Huckerman et al. (2008) describe how the broad-scale ridge pattern of a finger can be succinctly described using quadratic differentials. Their model requires relatively few parameters: one complex number for the location of each critical point, and five real parameters to specify a conformal mapping. Conditional on these parameters

their model describes a complex field Ridge(r) whose level curves correspond to the ridge orientation at all locations r. It might be possible to add these parameters to our model and marginalize them by extending the current MCMC algorithm.

Feng et al. (2011) show that minutiae tend to cluster in regions near these critical points and in regions of high ridge curvature. Thus it makes sense to model the latent minutia intensity as

$$\rho(r) = \rho_0 f\left(\left|\frac{\mathrm{d}^2\mathrm{Ridge}(r)}{\mathrm{d}^2 r}\right|\right) + \sum_k \rho_k \varphi(r; c_k, \sigma_k^2), \tag{8.1}$$

where the sum runs over the critical points with locations c_k, and f is some increasing function. The expected number of minutiae per finger would then depend on the ridge pattern of the finger.

The latent minutia orientations could be given von Mises distributions centred at Ridge(r) with some concentration parameter $\kappa_{\text{orientations}}$. Combined, these model extensions could correct the current violations of T2, T3 and T8. As a downside, in this enhanced model the intensity functions for the observed minutiae, ρ_{A3} and ρ_{B3}, would depend explicitly on the latent minutiae. These minutiae, which are analytically marginalized in our existing model, would have to be marginalized numerically within the MCMC algorithm under the new model. This would drastically slow down the algorithm.

§8.1.2 Alternative thinning procedure

In §7.7 we found that, contrary to the existing model assumptions, the observed minutiae are not selected uniformly at random from the latent minutiae. Rather, both the fingerprint and the fingermark usually consists of an observed subregion of the latent finger. Minutiae within this subregion have a high probability of being observed, and minutiae outside of the subregion have zero probability of being observed. Modelling this subregion could correct the observed violations of T1, T6 and T7.

There are at least two ways of modelling this. The first method tries to estimate the observed subregion directly, by finding the the convex subregion which maximizes the

likelihood. This method is likely extremely sensitive to local maxima in the likelihood function. The second method introduces *quality fields* $\delta_A(r)$ and $\delta_B(r)$ over the latent finger. These quality fields, possibly transformed by the logistic function, could be modelled with Gaussian processes, which could be marginalized within the MCMC algorithm. Whether or not this model is adequate for real-world data, and whether it is computationally feasible, remains to be investigated.

§8.1.3 Alternative location distortion model

The current model assumes that the observed minutia locations are generated from the true latent minutiae using a global similarity transformation and some independent normal location distortions. Evidence against this model is observed in §7.4 and §7.10. One solution is to model the distortion globally, rather than allowing independent distortions at each point. One model for such global distortions is the thin plate spline (Bookstein, 1989). Kent and Mardia (1994) show that the thin plate spline is the maximum point estimate (also the kriging estimate) of a specific Gaussian process. We could model the distortions as a Gaussian process and marginalize them within our MCMC algorithm to account for the observed violations of T4 and T10. Once again, whether this model is adequate for our data needs to be investigated, as does its computational feasibility.

§8.2 Larger databases

So far the model has been tested against a small NIST-FBI fingerprint database of fingermarks and their corresponding fingerprints.

There are several more fingerprint databases available from the Fingerprint Verification Competition (FVC) (Biometric System Laboratory, University of Bologna, 2012). These databases each consist of 150 fingerprints, with each fingerprint contributing 12 images. To use this database we must somehow extract the minutia locations from the digital image files. Preliminary testing shows that the extraction can

be done moderately well with a free, open source *automated fingerprint identification system* (AFIS) such as Vazam (2012) or Gonzalez et al. (2012). Any test performed in this way would confound the performance of our model with the performance of the AFIS. Despite this, the FVC data could provide a useful secondary test database.

Finally, it may be possible to gain access to some government fingerprint databases to test our model, as was done in Neumann et al. (2012a).

§8.3 Use more of the fingerprint information

Our model currently uses only minutia positions, orientations and types. These are called type 2 features in the fingerprint literature (Maltoni et al., 2009). By incorporating other information we could improve the performance of our model.

Type 1 features consist of the overall ridge pattern of a fingerprint. The ridge patterns can be classified into six categories, or they can be modelled as a phase portrait (Yager and Amin, 2004a). Neumann et al. (2012b) extend the model of Neumann et al. (2012a) by setting the probability of a match between fingerprints with different ridge pattern categories to zero. We can do the same with our model. Similarly, in many cases the specific finger (e.g., left ring finger or right thumb) can be inferred from the print. This information can be used to reduce the set of possible matches.

Type 3 features include the shape of each minutia, the shape and curvature of each individual ridge, and even the position of pores if the image quality is sufficiently high. There has not been much success with using these features in AFISs so far (Maltoni et al., 2009). This is partially due to the difficulty of translating type 3 information into a mathematically convenient form. However, this extra information is used extensively by human fingerprint examiners (Peterson et al., 2009). Any good model for generating likelihood ratios for courtroom use should take this additional information into account, since it affects the fingerprint examiner's decisions.

Of course, all of these improvements must be balanced against the corresponding

reduction in algorithm speed.

§8.4 Alternative estimators for the fixed parameters

We currently estimate our fixed parameters through an empirical Bayes framework, as described in §6.2. This is a reasonable approach under the assumption that our training database is a representative sample for the population of fingerprints and fingermarks that our model will encounter in the future. This approach also enables us, in principle, to interpret the resulting likelihood ratio as the ratio of the strengths of evidence of the two hypotheses.

One alternative approach to estimating the fixed parameters would be to find the parameters which maximize the discriminatory power of the computed likelihood ratios. This approach would likely improve the model's discrimination between true and false matches, but it would result in a "likelihood ratio" which cannot be interpreted as anything more than a discrimination score.

In practice, our model does not adequately fit the observed data, and our computed likelihood ratios do not permit a direct interpretation as the ratios of strengths of evidence anyway. Thus we resort to calibrating our likelihood ratios, as described in chapter 9. Since we are already subjecting our computed likelihood ratios to calibration, there is an argument adopting the second approach and maximizing the discriminatory power of our model. It would be interesting to determine how much additional discriminatory power results from this alternative parameter estimation procedure.

§8.5 Conclusion

We have described a marked Poisson point process model for paired minutia configurations in fingerprints and fingermarks, and the corresponding matching between these minutia configurations. We can efficiently sample from the distribution of the

unknown matching and parameters in this model using a Markov chain Monte Carlo method. The resulting sample can be used to compute likelihood ratios for comparing the hypothesis that the two configurations originate from the same finger against the hypothesis that they originate from different fingers.

The method provides excellent discrimination on simulated data. Using the method on a specific NIST-FBI database indicate that the model yields good discrimination between these two hypotheses as long as the fingermark is of reasonable quality. However, model validation shows that our simple model does not describe real-world fingerprints very well.

The likelihood ratios calculated are more extreme than what can be justified. The ratios can still be used as a sensible model-based method for discrimination between true and false matches, but they would have to be calibrated against a large real dataset before they can be interpreted as an accurate measure of the strength of evidence. We perform such a calibration in chapter 9. In any case, we believe the framework developed here can be used to establish a sound and model-based foundation for the analysis of fingerprint evidence.

Part II

Proper Scoring Rules

Chapter Nine

Calibrating the likelihood ratios

The likelihood ratios found in §6.3.2 are truly astronomical, exceeding 10^{80} in many cases. These are more extreme than we can plausibly defend. Our inflated likelihood ratios are most likely the direct result of modelling inadequacies. In this chapter we will calibrate the likelihood ratios against a large real dataset so that they can be interpreted as an accurate measure of the strength of evidence.

§9.1 Preliminaries

Let $\mathscr{A} = \{A_i : 1 \leq i \leq 258\}$ be the set of fingerprint minutia configurations in our dataset (Garris and McCabe, 2000), and let $\mathscr{B} = \{B_i : 1 \leq i \leq 258\}$ be the set of fingermarks. Thus $\mathscr{A} \times \mathscr{B}$ contains all possible 258^2 pairs of fingerprints and fingermarks. Let $\mathbf{1} : \mathscr{A} \times \mathscr{B} \to \{0, 1\}$ be the indicator function for a true match, $\mathbf{1}(A_i, B_j) = \mathbb{1}(i = j)$, and let $\lambda : \mathscr{A} \times \mathscr{B} \to \mathbb{R}$ be the estimates of the log-likelihood ratios computed in chapter 6 by some method from chapter 4. It will sometimes be convenient to work with the probability of H_p rather than log-likelihood ratios. Letting $\lambda_0 = \text{logit}(1 - p_0) = \log\{(1 - p_0)/p_0\}$, we will use Z_{p_0} to convert between the two:

$$p(H_p) = Z_{p_0}\{\lambda(A_i, B_j)\} = \text{logit}^{-1}\{\lambda(A_i, B_j) + \lambda_0\} = \frac{1}{1 + \exp\{-\lambda(A_i, B_j) - \lambda_0\}}. \quad (9.1)$$

The contents of this chapter are original, but some of the material reviewed in §9.1 overlaps with Forbes (2011).

We are interested in calibrated functions $\tilde{\lambda} : \mathscr{A} \times \mathscr{B} \to \mathbb{R}$ such that the transformed log-likelihood ratios $\tilde{\lambda}(A_i, B_j)$ are reliably calibrated (in a sense to be defined below).

After reviewing the necessary background, we will give some examples of calibration techniques and apply them to the likelihood ratios computed in chapter 6.

§9.1.1 Proper scoring rules

The idea of using proper scoring rules for calibration dates back to at least Murphy (1972), and is further explored in Dawid (1982) and DeGroot and Feinberg (1983). More recently, Brümmer and Doddington (2013) investigated how to calibrate likelihood ratios for forensic evidence.

In this section we will restrict ourselves to proper scoring rules for binary outcomes $x \in \{0, 1\}$. The basis theory for the more general case of categorical outcomes is discussed in §10.1, and the case of continuous outcomes is discussed in §11.1.1.

Let $\mathscr{P} = \{(p, 1-p)^\top : 0 < p < 1\}$ be the set of Bernoulli distributions, which take values on $\{0, 1\}$. A scoring rule for binary outcomes $s(x, P) : \{0, 1\} \times \mathscr{P} \to \mathbb{R}$ is a loss function which quantifies the accuracy of the prediction P after observing the binary outcome x. We extend the domain of our scoring rules from $\{0, 1\} \times \mathscr{P}$ to $\mathscr{P}^2$ by taking the expectation over x:

$$s\{(r, 1-r)^\top, P\} = rs(1, P) + (1-r)s(0, P). \tag{9.2}$$

We can recover the original function by simply plugging in $r = 0$ or $r = 1$. A scoring rule is (strictly) *proper* if $s(R, P)$ is (uniquely) minimized over $\mathscr{P}$ at $P = R$. We will abuse notation by writing $s(r, p)$ for $s\{(r, 1-r)^\top, (p, 1-p)^\top\}$.

Theorem 10.1 on page 111 shows that there is a correspondence between differentiable concave functions on $[0, 1]$ and proper scoring rules given by

$$\begin{aligned} S(p) &= s(p, p); \\ s(r, p) &= S(p) + (r-p)\frac{\mathrm{d}S(p)}{\mathrm{d}p}. \end{aligned} \tag{9.3}$$

We call $S(p)$ the optimal expected score corresponding to $s(x, p)$.

§9.1.2 Calibration functions

In practice we do not work with the set of all possible fingerprint/fingermark pairs $\mathscr{A} \times \mathscr{B}$, but rather with a subset $\mathscr{C} \subset \mathscr{A} \times \mathscr{B}$, which consists of all true matches (i.e., υ such that $\mathbf{1}(\upsilon) = 1$) and a random 5% subset of the false matches as described in § 6.3.2. We consider functions $\tilde{\lambda} : \mathscr{C} \to \mathbb{R}$ which lie in some suitable class of functions $\mathscr{L}$. We choose our particular $\tilde{\lambda}$ to minimize some objective function $F : \{0,1\}^{|\mathscr{C}|} \times \mathscr{L} \to \mathbb{R}$ which is a weighted sum of proper scoring rules,

$$F[\{1(\upsilon) : \upsilon \in \mathscr{C}\}, \tilde{\lambda}] = \sum_{\upsilon \in \mathscr{C}} s\{\mathbf{1}(\upsilon), Z_{p_0} \circ \tilde{\lambda}(\upsilon)\}/N\{\mathbf{1}(\upsilon)\}, \tag{9.4}$$

where $N(i) = |\{\upsilon \in \mathscr{C} : \mathbf{1}(\upsilon) = i\}|$ for $i = 0, 1$. Weighting the summands in this way ensures that objective function F does not depend heavily on the fact that $\mathscr{C}$ contains only a 5% subset of the false matches in $\mathscr{A} \times \mathscr{B}$. As long as $\mathscr{C}$ is large enough to be representative of the population of fingerprints and fingermarks, we expect the below calibration results to be independent of the size of $\mathscr{C}$.

The minimizing value of F, $\tilde{\lambda}(A_i, B_j)$, clearly depends on both s and $\mathscr{L}$. It also depends on λ_0 through Z_{p_0}; we will henceforth set $\lambda_0 = 0$, which implies $Z_{p_0}(\cdot) = \text{logit}^{-1}(\cdot)$. As an aside, note that we can view F itself as a proper scoring rule for outcomes in $\{0,1\}^{|\mathscr{C}|}$, which we are minimizing over the set of probability distributions with densities

$$\left\{ \prod_{\upsilon \in \mathscr{C}} \{Z_{p_0} \circ \tilde{\lambda}(\upsilon)\}^{\mathbf{1}(\upsilon)} \{1 - Z_{p_0} \circ \tilde{\lambda}(\upsilon)\}^{1-\mathbf{1}(\upsilon)} : \tilde{\lambda} \in \mathscr{L} \right\}. \tag{9.5}$$

Ideally, a calibrated probability $Z_{p_0} \circ \tilde{\lambda}(\upsilon)$ should be equal to the proportion of fingerprint pairs similar to υ which satisfy H_p. That is, given an interval $[c_{k-1}, c_k) \subseteq [0, 1)$ and the corresponding subset of $\mathscr{C}$,

$$\mathscr{C}_k = \{\upsilon \in \mathscr{C} : Z_{p_0} \circ \tilde{\lambda}(\upsilon) \in [c_{k-1}, c_k)\}, \tag{9.6}$$

we expect that the proportion of $\upsilon \in \mathscr{C}_k$ with $\mathbf{1}(\upsilon) = 1$ should also lie in $[c_{k-1}, c_k)$. This corresponds to the *reliable calibration* criterion of Dawid (1986). Formally, given

any partition $0 = c_0 < c_1 < \ldots < c_K = 1$, we want

$$\frac{\sum_{\upsilon \in \mathscr{A}\times\mathscr{B}} I\{Z_{p_0} \circ \tilde{\lambda}(\upsilon) \in \mathscr{C}_k\}\{\mathbf{1}(\upsilon) - Z_{p_0} \circ \tilde{\lambda}(\upsilon)\}}{\sum_{\upsilon \in \mathscr{A}\times\mathscr{B}} \mathbb{1}\{Z_{p_0} \circ \tilde{\lambda}(\upsilon) \in \mathscr{C}_k\}} \to 0 \quad \text{for } k = 1, \ldots, K \tag{9.7}$$

as the number of terms in each $\mathscr{C}_k$ tends to infinity. However, since we have a finite number of terms in the sum, reliable calibration is a purely theoretical aim.

§9.2 Examples

In this section we describe some common methods for likelihood ratio calibration. We apply these calibration methods to our estimates of the log-likelihood ratios $\{\lambda(\upsilon) : \upsilon \in \mathscr{C}\}$. We will consider several different choices of λ, one for each estimation method (harmonic means, Chib, bridge, reversible jump) used in §6.3.2.

§9.2.1 Affine calibration

The simplest method of calibration simply affinely transforms the original log-likelihood ratios: $\mathscr{L} = \{\upsilon \mapsto \alpha + \beta\lambda(\upsilon) : \alpha, \beta \in \mathbb{R}\}$.

There are several common choices for the proper scoring rule in affine calibration, including:

- the log-loss score $s(x, p) = -x \log p - (1-x)\log(1-p)$,
- the probit score $s(x, p) = -x \log\{\Phi \circ \text{logit}(p)\} - (1-x)\log\{1 - \Phi \circ \text{logit}(p)\}$, where Φ is the cumulative distribution function for a standard one-dimensional normal distribution,
- the Brier score $s(x, p) = x(1-p)^2 + (1-x)p^2$.

The first two are equivalent to a generalized linear regression of $\mathbf{1}(\upsilon)$ onto $\lambda(\upsilon)$ using the logit and probit link functions respectively. One advantage of the log-loss scoring rule is that the computed values of $\tilde{\lambda}(\upsilon)$ are independent of λ_0: changing λ_0 to λ_0' simply changes α to $\alpha + \beta(\lambda_0 - \lambda_0')$, leaving $\tilde{\lambda}(\upsilon)$ unchanged.

Figure 9.1 shows the histogram of the calibrated log-likelihood ratios. Since the transformation $\lambda \mapsto \tilde{\lambda}$ is monotonic, the ROC curves (figure 6.4) are unchanged

from the uncalibrated curves in figure 6.4. Note that the calibrated log-likelihood ratios usually lie between -5 and 5, which is a far more moderate range than the uncalibrated range of -20 to 120. This moderate range is consistent with the discriminatory power of our model, as demonstrated by the overlap in the histograms in figure 6.6 and figure 9.1.

All three proper scoring rules give broadly similar results. For instance, for the log-likelihood ratios computed by Chib's method, the optimal affine calibration functions are $\tilde{\lambda}_{logloss}(v) = -2.17 + 0.0930\lambda(v)$, $\tilde{\lambda}_{probit}(v) = -1.19 + 0.0451\lambda(v)$, and $\tilde{\lambda}_{Brier}(v) = -2.74 + 0.133\lambda(v)$.

§9.2.2 Monotonic calibration

Brümmer and du Preez (2006) suggest calibrating the log-likelihood ratios by regressing onto the observed outcomes, subject only to a monotonicity constraint:

$$\mathscr{L} = \{\tilde{\lambda} : \mathscr{A} \times \mathscr{B} \to \mathbb{R}; \tilde{\lambda}(v') \geq \tilde{\lambda}(v) \text{ whenever } \lambda(v') \geq \lambda(v)\}. \tag{9.8}$$

They use the log-loss scoring rule, but Brümmer and du Preez (2013) shows the minimizing function is invariant under the substitution of any other proper scoring rule for binary outcomes.

The minimizing function can be written as $\tilde{\lambda}(v) = f \circ \lambda(v)$ where $f : \mathbb{R} \to \mathbb{R}$ is an increasing piecewise-constant function with jumps at $\{\lambda(v) : v \in \mathscr{A} \times \mathscr{B}, \mathbf{1}(v) = 1\}$. The function f can be found in linear time using the algorithm in Ahuja and Orlin (2001).

The fact that the calibration function is independent of the choice of scoring rule makes this approach initially appealing. However, in our experience it suffers from extreme overfitting. The optimal $\tilde{\lambda}$ sends all v' with $\lambda(v') < \min_{v \in \mathscr{A} \times \mathscr{B}, \mathbf{1}(v)=1}\{\lambda(v)\}$ to negative infinity, and it sends all v' with $\lambda(v') > \max_{v \in \mathscr{A} \times \mathscr{B}, \mathbf{1}(v)=0}\{\lambda(v)\}$ to positive infinity. This results in a highly concentrated bimodal distribution for the log-likelihood ratios (see figure 9.1). As in the affine calibration case, the ROC curves are unchanged from the uncalibrated curves.

The overfitting could be overcome by choosing f using only a training subset of $\mathscr{A} \times \mathscr{B}$ and evaluating the calibration $\tilde{\lambda} = f \circ \lambda$ on the remaining subset. However, in our experience the resulting calibration function $\tilde{\lambda}$ depends strongly on the precise training subset used. Thus this approach appears unsuitable for our purposes.

§9.2.3 More sophisticated calibration methods

The examples above are extreme opposites: affine regression restricts the calibration function to be linear, while monotonic regression proceeds non-parametrically with only a monotonicity restriction. The former leads to transformed likelihood ratios which are non-reliably calibrated, while the latter leads to overfitted likelihood ratios.

A middle ground approach might consider higher-order parametric calibration functions, such as polynomials or perhaps polynomial splines. By carefully choosing the number of parameters, we might find a balance between allowing the calibration function to fit the data and preventing its overfitting. Adding a shrinkage term to the objective function could also assist with preventing overfitting.

However, here we will follow an alternative approach. The above methods used only the computed log-likelihood ratios $\lambda(\upsilon)$. By including more information in our calibration procedure we can achieve a more reliable calibration. We consider regressing $\mathbf{1}(\upsilon)$ onto multiple covariates: the uncalibrated log-likelihood ratios $\lambda(\upsilon)$, the number of minutiae in the fingermark, and the fingermark quality indicator (good, bad or ugly). This is equivalent to

$$\mathscr{L} = \{\upsilon \mapsto \beta_0 \, \mathbb{1}(\upsilon \text{ good}) + \beta_1 \, \mathbb{1}(\upsilon \text{ bad}) + \beta_2 \, \mathbb{1}(\upsilon \text{ ugly}) + \beta_3 n_{B,\upsilon} + \beta_4 \lambda(\upsilon) : \beta_i \in \mathbb{R}\}. \tag{9.9}$$

When using the log-loss scoring rule this becomes logistic regression; the optimal function for the log-likelihoods computed by Chib's method has $\beta_0 = -1.62, \beta_1 = -1.05, \beta_2 = -0.70, \beta_3 = -0.068$, and $\beta_4 = 0.11$. The most important correction is the β_3 term: it corrects for the fact that the number of plausible matches increases with the number of minutiae, so that for large n_B even a false match may have a plausible

matching. When using the probit scoring rule this calibration method becomes probit regression; the regression coefficients are similar to the logistic regression case. The histograms of the calibrated log-likelihood ratios for both methods are shown in figure 9.1.

The ROC curves for the log-loss regression calibration is shown in figure 9.2. We note the calibrated log-likelihood rations provide significantly better discrimination than the uncalibrated values. Of particular interest is the naive harmonic means curve. Recall from §6.3.2 that naive harmonic means provided the least accurate estimate of the log-likelihood ratio. Yet, after calibration, it demonstrates the best discrimination between true and false matches. It has been noted previously (Raftery et al., 2007) that the naive harmonic means estimate of a marginalized likelihood is insensitive to the prior distribution of the parameters: this is a main reason why it typically provides a poor estimate of the marginalized likelihood. In effect, unless the sample size is astronomical, the harmonic means estimate doesn't approximate the marginal likelihood at all, but rather approximates some other function of the data which is insensitive to the prior distribution. In our case, where the model and the prior distributions do not fit the data well (see chapter 7), it seems that this other function actually provides better discrimination between true and false matches than the likelihood ratio.

§9.3 Comparing the calibration methods

In order to assess if the probabilities are reliably calibrated, we consider the reliable calibration test

$$G(\ell) = \frac{\sum_{v \in \mathscr{C}} \mathbb{1}\left\{\log_{10}(e) \times \tilde{\lambda}(v) \in [\ell - 0.01, \ell + 0.01)\right\}\left\{\mathbf{1}(v) - Z_{p_0} \circ \tilde{\lambda}(v)\right\}}{\sum_{v \in \mathscr{C}} \mathbb{1}\left\{\log_{10}(e) \times \tilde{\lambda}(v) \in [\ell - 0.01, \ell + 0.01)\right\}}, \quad (9.10)$$

which can be seen as a specific version of the reliable calibration definition (9.7). If our probabilities are well calibrated we expect $G(\ell)$ to be approximately zero for all ℓ. It is plotted in figure 9.3. We notice that all of the calibration methods have a slightly

negative G for values around zero, which implies that our calibrated probabilities are still too large in that region. All four of the basic calibration methods (log-loss calibration, probit calibration, monotonic calibration and Brier calibration) have a large positive peak in G when $\log_{10}(e) \times \tilde{\lambda}(\upsilon)$ is approximately -3, which corresponds to true matches with negative log-likelihood ratios. This peak vanishes for the two regression methods, which implies that these matches' negative log-likelihood ratios were due to poor quality fingermarks with a low number of minutiae. The regression methods were successful in calibrating these log-likelihood ratios.

Further insights are available from figure 9.4, which plots the calibrated probability as a function of the uncalibrated $\log_{10}$-likelihood ratio. Notice that all of the calibration methods result in a much smoother transition between $p = 0$ and $p = 1$ than the uncalibrated probabilities $p = Z_{p_0} \circ \lambda(\upsilon)$. The transformations $\lambda(\upsilon) \mapsto \tilde{\lambda}(\upsilon)$ for the regression models are not monotonic, since $\tilde{\lambda}(\upsilon)$ depends on both the fingermark quality and the number of fingermark minutiae in addition to log-likelihood ratios. These calibration curves show that the calibrated probabilities are less extreme (closer towards the midpoint of 0.5) than the uncalibrated probabilities, which is to be expected, since we anticipated that the uncalibrated probabilities were too extreme. Conversely, if our model was able to adequately explain the observed data, we would expect that the calibration procedure would not substantially change the computed likelihood ratios, and the calibration curves would be closer to the uncalibrated curve in figure 9.4.

In conclusion, we have shown that our model yields reasonable likelihood ratios and strong discrimination between true and false matches after calibration. Thus the calibrated likelihood ratios can be used as a model-based and justifiable measure of the weight of evidence for the prosecution hypothesis that a given fingerprint and fingermark originate from the same finger against the defence hypothesis that they originate from different fingers.

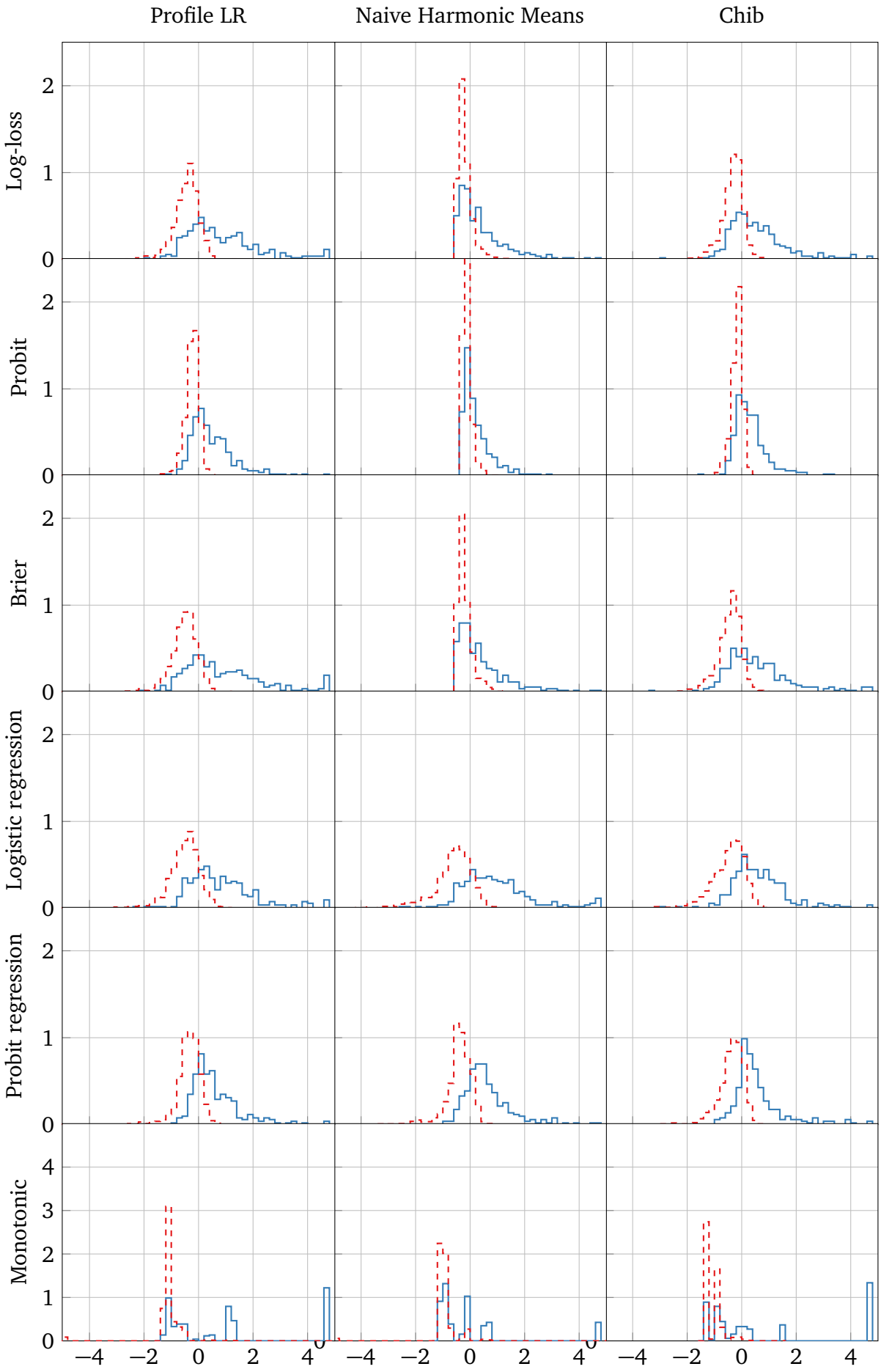


Figure 9.1: Histogram of the calibrated log-likelihood ratios on Garris and McCabe (2000) as computed by profile maximization, naive harmonic means and Chib's method. The other marginalization methods' histograms are similar to the Chib histograms. Log-likelihood ratios corresponding to false matches are dashed and red, while true matches are solid and blue.

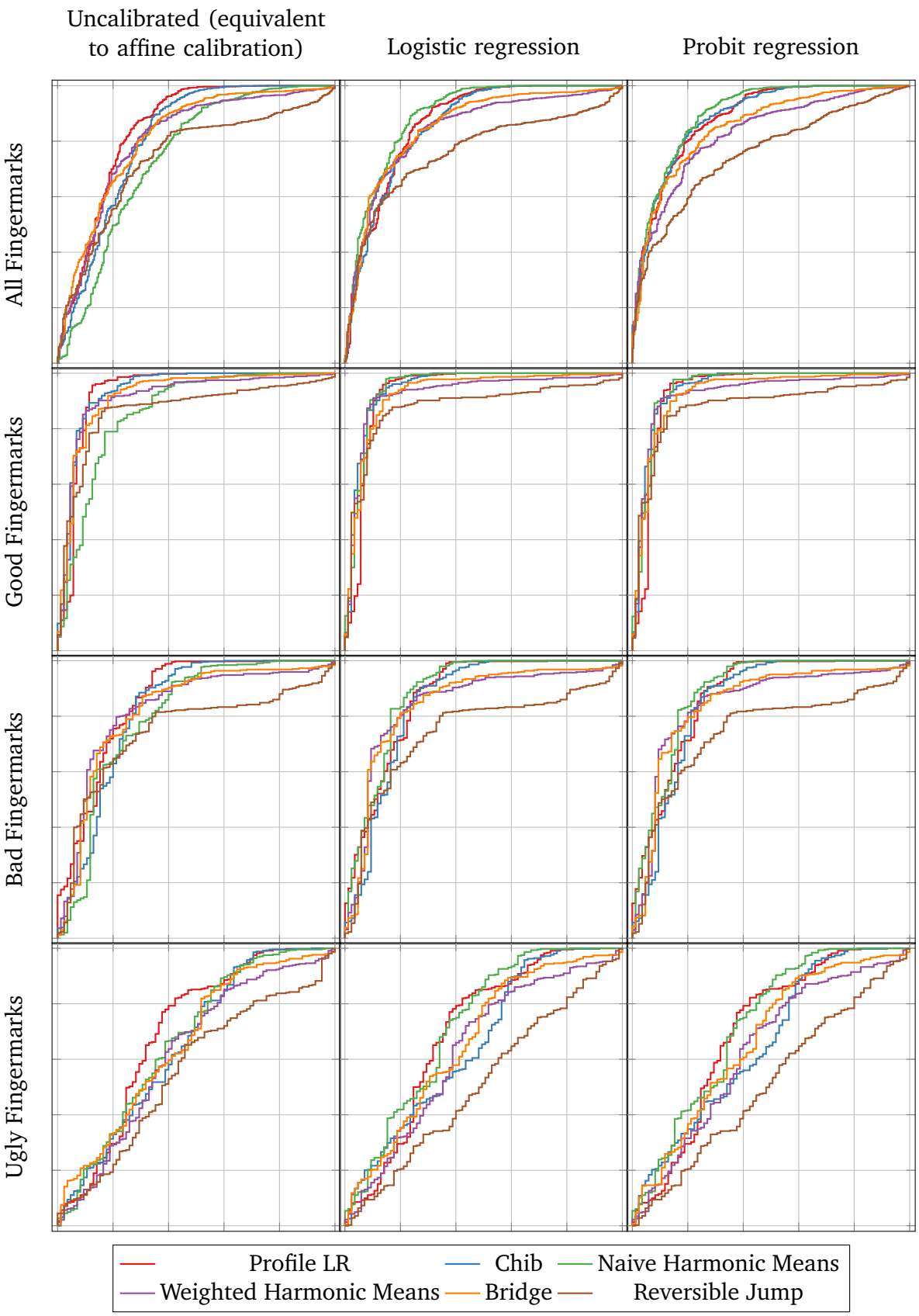


Figure 9.2: Receiver Operating Characteristic (ROC) curves for the calibrated and uncalibrated log-likelihood ratios for the dataset (Garris and McCabe, 2000) and its three subsets (good, bad and ugly). The x-axis represents the proportion of false matches classified as matches, the y-axis represents the proportion of true matches classified as matches.

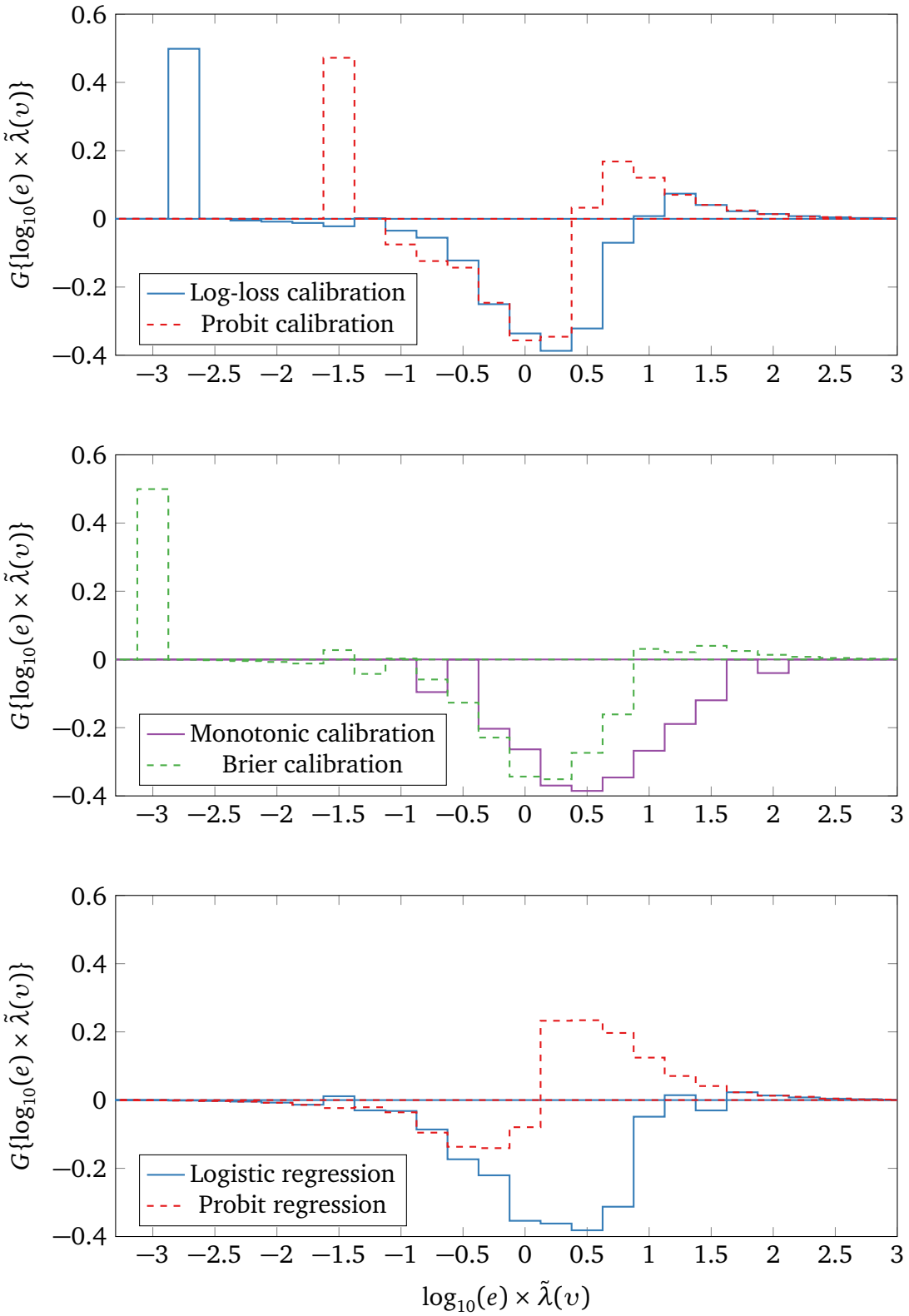


Figure 9.3: Plot of G, defined in (9.10), for the Chib-computed probabilities $p(H_p \,|A,B)$. If the probabilities are reliably calibrated we expect G to be identically 0.

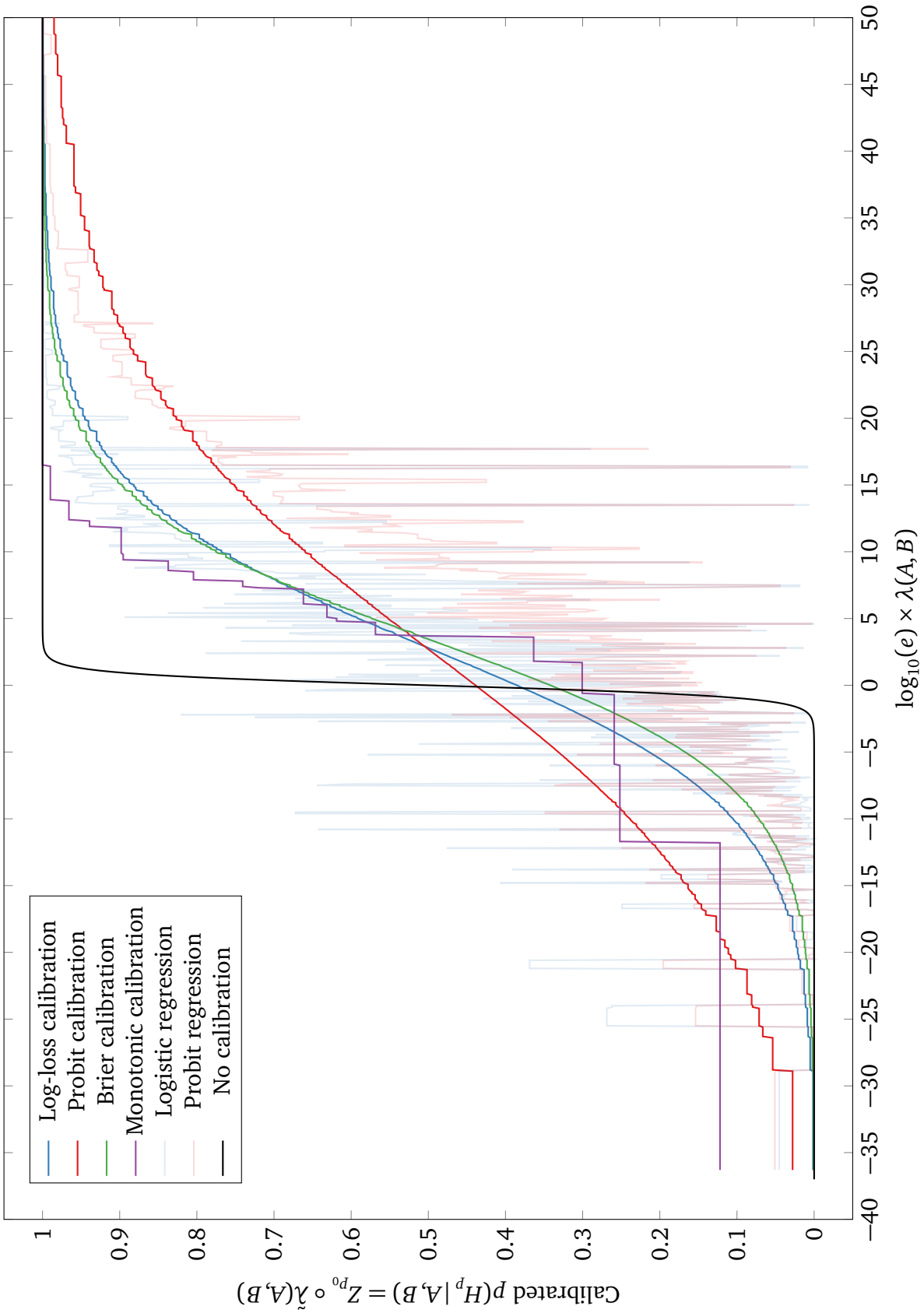


Figure 9.4: Plot of the calibration curves for the Chib-computed probabilities. Note that the regression calibrations are not monotonic, since in addition to $\lambda(A,B)$, they also depend on the fingermark quality and the number of minutiae in the fingermark.

Chapter Ten

Compatible proper scoring rules

§10.1 Introduction

Many proper scoring rules such as the Brier and log scoring rules implicitly reward a probability forecaster relative to a uniform baseline distribution. Recent work has motivated weighted proper scoring rules, which have an additional baseline parameter. To date two families of weighted proper scoring rules have been introduced, the weighted power and pseudospherical scoring families. These families are compatible with the log scoring rule: when the baseline minimizes the log scoring rule over some set of distributions, the baseline also minimizes the weighted power and pseudospherical scoring rules over the same set. In this chapter we characterize all weighted proper scoring families and prove a general property: every proper scoring rule is compatible with some weighted scoring family, and every weighted scoring family is compatible with some proper scoring rule.

Suppose X is a random variable taking values in $\mathscr{X} = \{1, \ldots, m\}$. The valid distributions for X are

$$\mathscr{P} = \{(p_1, \ldots, p_m)^\top : 0 \leq p_i \leq 1, \sum_{i=1}^{m} p_i = 1\} \subset \mathbb{R}^m. \tag{10.1}$$

This chapter is a very slight modification of Forbes (2012).

A *scoring rule* $s(x, P)$ is a real-valued loss function which quantifies the accuracy of a predictive distribution $P \in \mathscr{P}$ upon observing the realized value $x \in \mathscr{X}$. It is (strictly) *proper* if, for any $R = (r_1, \ldots, r_m)^\top \in \mathscr{P}$, the expected value $s(R, P) = \sum_{i=1}^m r_i s(i, P)$ is (uniquely) minimized over $\mathscr{P}$ at $P = R$. Note $s(R, P)$ is linear in its first argument.

We say two scoring rules equivalent if they are linearly related for all $P, R \in \mathscr{P}$:

$$s_1(R, P) = a\{s_2(R, P) + \langle b\,, R\rangle\}, \tag{10.2}$$

where $\langle \cdot\,, \cdot \rangle$ is the standard inner product on $\mathbb{R}^m$, $a > 0$, and $b \in \mathbb{R}^m$.

The main characterization theorem for proper scoring rules was stated by McCarthy (1956) and proved by Hendrickson and Buehler (1971).

Theorem 10.1. *Let $\mathscr{P}_\Lambda$ extend $\mathscr{P}$ to a convex subset of $\mathbb{R}^m$ by*

$$\mathscr{P}_\Lambda = \{\lambda P : \lambda > 0, P \in \mathscr{P}\} = \{(y_1, \ldots, y_m)^\top \in \mathbb{R}^m : y_i \geq 0 \text{ for } 1 \leq i \leq m\} \backslash \{0\}. \tag{10.3}$$

A scoring rule s is proper if and only if the function $S : \mathscr{P}_\Lambda \to \mathbb{R}$

$$S(\lambda P) = \lambda s(P, P) \tag{10.4}$$

is concave on $\mathscr{P}_\Lambda$ and satisfies $S(R) \leq s(R, P)$ for all $R, P \in \mathscr{P}$. The scoring rule is strictly proper if and only if S is strictly concave on $\mathscr{P}_\Lambda$.

The function S is called the optimal expected score. Grünwald and Dawid (2004) showed that the optimal expected score can be interpreted as a generalized entropy.

When S is differentiable on $\mathscr{P}_\Lambda$ we have (Hendrickson and Buehler, 1971)

$$s(R, P) = \langle R\,, \nabla_{\lambda P}\rangle\, S(\lambda P) = \sum_{i=1}^m r_i \frac{\mathrm{d}S\{(\lambda p_1, \ldots, \lambda p_m)^\top\}}{\mathrm{d}\lambda p_i} \tag{10.5}$$

which associates a proper scoring rule with any concave differentiable function S. For the rest of this chapter we assume that S is strictly concave and twice differentiable on $\mathscr{P}_\Lambda$. We further assume that S achieves its unique minimum over $\mathscr{P}$ in $\mathscr{P}_+ = \{p \in \mathscr{P} : p_i > 0 \text{ for } 1 \leq i \leq m\}$, the interior of $\mathscr{P}$.

We extend the domain of s to $\mathscr{P} \times \mathscr{P}_\Lambda$ by $s(R, \lambda P) = s(R, P)$. This allows us to differentiate s with respect to its second parameter. It also implies that $\langle \nabla_P s(R, P), \mathbf{1} \rangle = 0$ for any $R, P \in \mathscr{P}$, where $\mathbf{1} \in \mathbb{R}^m$ has all entries equal to one.

Consider a sequence of observations $x_1, \ldots, x_n$ with empirical distribution $R \in \mathscr{P}$. Let $P(\theta)$ be some model which takes values in $\mathscr{P}_+$ and is differentiable over some open convex set Θ. Then any scoring rule defines an optimal score estimator (Gneiting and Raftery, 2007) via

$$\tilde{\theta}(R) = \operatorname*{argmin}_{\theta \in \Theta} s\{R, P(\theta)\} = \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} s\{x_i, P(\theta)\}. \tag{10.6}$$

From (10.2), all equivalent scoring rules have the same optimal score estimator. The optimal score estimator is *well behaved* at R if $\tilde{\theta}(R)$ exists and is the unique root of $\nabla_\theta s\{R, P(\theta)\}$ in Θ. When s is the log scoring rule $s(R, P) = -\sum_{i=1}^m r_i \log p_i$, the optimal score estimator becomes the maximum likelihood estimator.

A well behaved optimal score estimate $\tilde{\theta}(R)$ yields the parameter choice that minimizes the forecaster's expected score under the assumption that the future is similar to the past. Specifically, if we constrain the forecaster to issue a prediction $P(\theta)$ for some $\theta \in \Theta$, and if she believes that the next observation's distribution is R, then $P\{\tilde{\theta}(R)\}$ minimizes her expected score.

The optimal score estimator can be generalized so that each x_i follows a different probability distribution, as long as these distributions share a common parameter $\theta \in \Theta$. Thus the optimal score estimator is applicable to regression models that depend on both θ and some additional covariates. For the sake of brevity we consider only the basic optimal score estimator here, though all the results hold in general.

§10.2 Results

We define the baseline of a strictly proper scoring rule to be the unique $Q \in \mathscr{P}_+$ that minimizes the generalized entropy $S(P)$. For example, the log scoring rule's generalized entropy is the Shannon entropy, which is minimized by the uniform distribution.

Proper scoring rules tends to give smaller losses for riskier predictions which vary significantly from the baseline. Given $Q \in \mathscr{P}_+$ and a strictly proper scoring rule $s(R,P)$, there is an equivalent rule with baseline Q given by $\tilde{s}(R,P) = s(R,P) - s(R,Q)$.

A weighted scoring family

$$s(R, P \parallel \cdot) = \{s(R, P \parallel Q) : Q \in \mathscr{P}_+\} \tag{10.7}$$

is a family of strictly proper scoring rules where each member $s(R,P \parallel Q)$ has baseline Q. Two weighted proper scoring rules families are equivalent if (10.2) is satisfied, where now a and b are functions of Q. Different members from the same family need not be equivalent. The dependence of $s(R,P \parallel Q)$ on Q is relatively arbitrary; we require only that Q minimizes the generalized entropy, $Q = \operatorname{argmin}_{P\in\mathscr{P}} s(P,P \parallel Q)$.

Weighted scoring families allow us to tailor our scoring rule to the problem at hand, as motivated in Jose et al. (2009) and Johnstone and Lin (2011). This tailoring is achieved by modifying the baseline. The baseline is easily interpretable and justifiable in many real world situations. For instance, weighted scoring families are used in Jose et al. (2008) for a optimal portfolio allocation problem, where the baseline corresponds to the market price.

Let $s(R,P)$ be a proper scoring rule and $s(R,P \parallel \cdot)$ be a weighted scoring family. We say $s(R,P \parallel \cdot)$ is compatible with $s(R,P)$ if for any $Q \in \mathscr{P}_+$ and $R \in \mathscr{P}$,

$$\nabla_P s(R,P)|_{P=Q} = a(Q)\nabla_P s(R,P \parallel Q)|_{P=Q} \tag{10.8}$$

for some function $a(Q) > 0$. In words, (10.8) says that the tangent of a weighted scoring rule at its baseline Q is parallel to the compatible scoring rule's tangent at Q. By approximating $s(R,P \parallel Q)$ with its tangent at $P = Q$ and applying (10.8), we obtain

$$s(R,P \parallel Q) \approx s(R,Q \parallel Q) + \frac{1}{a(Q)} \left\langle \nabla_P s(R,P)|_{P=Q} \, , \, P - Q \right\rangle. \tag{10.9}$$

The first term corresponds to an equivalence factor $\langle b(Q), R\rangle$. Thus, up to equivalence, every member of the weighted scoring family $s(R,P \parallel \cdot)$ is linearly approximated by the compatible proper scoring rule $s(R,P)$ in the vicinity of its baseline.

Theorem 10.2. *Any proper scoring rule is compatible with at least one weighted scoring family. Conversely, every weighted scoring family is compatible with some proper scoring rule, which is unique up to equivalence.*

Proof. Let $s(R,P)$ be a proper scoring rule. From the definition (10.8), it is compatible with the weighted scoring family where each member is equivalent to $s(R,P)$:

$$s(R,P \parallel Q) = s(R,P) - s(R,Q). \tag{10.10}$$

Conversely, consider the weighted scoring family $s(R,P \parallel \cdot)$. From (10.5) and (10.8), a proper scoring rule $s(R,P)$ is compatible with this family if and only if its optimal expected score $S(P)$ satisfies

$$\nabla_Q^2 S(Q) = a(Q)\, \nabla_P^2 S(P \parallel Q)\big|_{P=Q} \tag{10.11}$$

for some $a(Q) > 0$ and all $Q \in \mathscr{P}_+$. The right hand side is a negative definite matrix since it is the Hessian of the concave function $S(P \parallel Q)$. Thus the S satisfying (10.11) is concave and corresponds to a strictly proper scoring rule. This solution is unique up to equivalence since the solution of a second-order differential equation is unique up to a linear term. □

Having shown that a compatible proper scoring rule always exists, we now provide an alternative characterization for compatibility which has direct applications to optimal score estimation and decision theory.

Lemma 10.3. *A weighted scoring family $s(R,P \parallel \cdot)$ is compatible with the proper scoring rule $s(R,P)$ if and only if $\nabla_\theta s\{R,P(\theta)\}|_{\theta=\theta_0} = 0$ implies $\nabla_\theta s\{R,P(\theta) \parallel P(\theta_0)\}|_{\theta=\theta_0} = 0$ for all differentiable models $P(\theta)$, all $\theta_0 \in \Theta$, and all $R \in \mathscr{P}_+$.*

Proof. Choose some model $P(\theta)$ and $R \in \mathscr{P}_+$. Suppose that $s(R,P \parallel \cdot)$ is compatible with $s(R,P)$, so that (10.8) holds for all $Q \in \mathscr{P}_+$. Then (10.8) certainly holds when $Q = P(\theta_0)$ for any $\theta_0 \in \Theta$. Left multiplying both sides of (10.8) with the matrix $\nabla_\theta P^\top(\theta)\big|_{\theta=\theta_0}$ and using the chain rule,

$$\nabla_\theta s\{R,P(\theta)\}|_{\theta=\theta_0} = a\{P(\theta_0)\}\, \nabla_\theta s\{R,P(\theta) \parallel P(\theta_0)\}|_{\theta=\theta_0}. \tag{10.12}$$

Thus if $\nabla_\theta s\{R, P(\theta)\}|_{\theta=\theta_0} = 0$ then $\nabla_\theta s\{R, P(\theta) \parallel P(\theta_0)\}|_{\theta=\theta_0} = 0$.

Conversely, suppose $\nabla_\theta s\{R, P(\theta)\}|_{\theta=\theta_0} = 0$ implies $\nabla_\theta s\{R, P(\theta) \parallel p(\theta_0)\}|_{\theta=\theta_0} = 0$. When $Q = R$, both sides of (10.8) are being evaluated at their critical points and hence are zero. We will show (10.8) holds for $Q \neq R$ by showing that $v = \nabla_P s(R, P \parallel Q)|_{P=Q}$ is parallel to $w = \nabla_P s(R, P)|_{P=Q}$. Using (10.5) we can rewrite v as

$$v = \nabla_P^2 S(P \parallel Q)\Big|_{P=Q} R, \tag{10.13}$$

where $\nabla_P^2 S(P \parallel Q)$ is the negative definite Hessian of $S(P \parallel Q)$. This implies $v \neq 0$ since $R \neq 0$. Furthermore since v is a gradient of $s(R, P \parallel Q)$, $\langle v\,,\, \mathbf{1}\rangle = 0$. The same arguments show that $w \neq 0$ and $\langle w\,,\, \mathbf{1}\rangle = 0$.

Suppose v is not parallel to w. Then we can define the non-zero vector

$$b = v - \frac{\langle v\,,\, w\rangle}{\langle w\,,\, w\rangle} w. \tag{10.14}$$

By construction $\langle b\,,\, w\rangle = 0$. Consider the model $P(\theta) = Q + \theta b$ where θ takes values on Θ, an open neighbourhood of zero small enough such that $\{P(\theta) : \theta \in \Theta\} \subset \mathscr{P}_+$. It follows from $\langle v\,,\, \mathbf{1}\rangle = 0$ and $\langle w\,,\, \mathbf{1}\rangle = 0$ that $P(\theta)$ is normalized for all $\theta \in \Theta$. Thus $P(\theta)$ is a valid distribution for $\theta \in \Theta$, and, by our choice of w and $P(\theta)$,

$$\nabla_\theta s\{R, P(\theta)\}|_{\theta=0} = \left\langle \nabla_\theta P(\theta)|_{\theta=0}\,,\, w\right\rangle = \langle b\,,\, w\rangle = 0. \tag{10.15}$$

Hence by assumption, $\nabla_\theta s\{R, P(\theta) \parallel Q\}|_{\theta=0} = 0$. By definition of v we have

$$\nabla_\theta s\{R, P(\theta) \parallel Q\}|_{\theta=0} = \langle b\,,\, v\rangle \tag{10.16}$$

and thus $\langle b\,,\, v\rangle = 0$. Substituting this into (10.14), $\langle v\,,\, v\rangle \langle w\,,\, w\rangle = \langle v\,,\, w\rangle^2$ and the Cauchy–Schwarz inequality implies that v is parallel to w: $w = a(Q,R)v$. Using (10.13), we rewrite $w = a(Q,R)v$ as

$$\nabla_P^2 S(P \parallel Q)\Big|_{P=Q} R = a(R,Q)\, \nabla_P^2 S(P)\Big|_{P=Q} R. \tag{10.17}$$

Since both matrices are negative definite, $a(R,Q) > 0$. Since the left hand side is linear in R, we see $a = a(Q)$, which proves (10.8). □

Consider a forecaster motivated by a weighted scoring rule with baseline Q to issue a prediction $P(\theta)$ for X. She chooses her prediction based on some decision rule $\breve{\theta}(R)$, where R is the empirical distribution of the previous observations of X. For instance, $\breve{\theta}$ could be the optimal score estimator for her weighted scoring rule. Her expected loss is $s[P^*, P\{\breve{\theta}(P^*)\} \parallel Q]$, which depends on the unknown true distribution P^* of X. Since P^* is unknown it is approximated with the empirical distribution R.

Suppose the baseline is determined by the optimal score estimator of the compatible scoring rule, $Q = P\{\tilde{\theta}(R)\}$. Then, assuming $\breve{\theta}$ and $\tilde{\theta}$ to be well behaved at R, Lemma 1 implies that the forecaster's loss is uniquely minimized when she issues the prediction Q. The optimal score estimator of the compatible scoring rule dominates any other estimator $\breve{\theta}$ for this choice of baseline.

§10.3 Examples

Define the quasi-Bregman weighted scoring families to be the proper scoring rules with optimal expected scores

$$S(P \parallel Q) = g'(1) \sum_{i=1}^{m} \frac{p_i f(q_i)}{q_i} h' \left\{ g(1) \sum_{j=1}^{m} f(q_j) \right\} - h \left\{ \sum_{i=1}^{m} f(q_i) g\left(\frac{p_i}{q_i}\right) \right\}, \quad (10.18)$$

where g' denotes the derivative of g with respect to its parameter, and similarly for h'. We require that f is positive, g is twice differentiable and strictly convex, and that h is twice differentiable and strictly increasing. This defines a weighted scoring family for each choice of f, g, and h. The expected score $S(P \parallel Q)$ is strictly concave since g is strictly convex, f is positive and h is increasing. Hence the quasi-Bregman weighted scoring families are strictly proper. The first term of (10.18) ensures that $S(P \parallel Q)$ has baseline Q. Removing it yields a simpler, equivalent rule for optimal score estimation.

The weighted power and pseudospherical scoring families of Jose et al. (2008),

$$\begin{aligned} s^{\text{pow}}(R, P \parallel Q) &= \frac{1 - \sum_{i=1}^{m} r_i p_i^{\beta-1} q_i^{1-\beta}}{\beta - 1} - \frac{1 - \sum_{i=1}^{m} p_i^{\beta} q_i^{1-\beta}}{\beta}, \\ s^{\text{ps}}(R, P \parallel Q) &= \frac{1}{\beta - 1} \left\{ 1 - \frac{\sum_{i=1}^{m} r_i p_i / q_i}{\left(\sum_{i=1}^{m} p_i^{\beta} q_i^{1-\beta}\right)^{1/\beta}} \right\}, \end{aligned} \quad (10.19)$$

for $\beta > 1$, are quasi-Bregman weighted scoring families with $f(x) = x$ and

$$h^{\text{pow}}(x) = \frac{x-1}{\beta(\beta-1)}, \quad g^{\text{pow}}(x) = x^{\beta}, \quad h^{\text{ps}}(x) = \frac{x^{1/\beta}-1}{\beta(\beta-1)}, \quad g^{\text{ps}}(x) = x^{\beta}. \quad (10.20)$$

Johnstone and Lin (2011) proved that $\nabla_\theta s\{R, P(\theta)\}|_{\theta=\theta_0} = 0$ implies

$$\nabla_\theta s\{R, P(\theta) \parallel P(\theta_0\}|_{\theta=\theta_0} = 0 \quad (10.21)$$

when $s(R, P \parallel R)$ is a power or pseudospherical weighted scoring family and $s(R, P)$ is the log scoring rule. From Lemma 1, this is equivalent to showing that the power and pseudospherical weighted scoring families are compatible with the log scoring rule.

Corollary 10.4. *The log scoring rule is compatible with any quasi-Bregman weighted scoring family with $f(x) = x$. This holds for any twice differentiable and strictly convex g, and any twice differentiable and strictly increasing h.*

Proof. By substituting $f(x) = x$ into (10.18) and using (10.5), we obtain

$$s(R, P \parallel Q) = -h'\left\{\sum_{i=1}^{m} q_i g\left(\frac{p_i}{q_i}\right)\right\} \sum_{j=1}^{m} r_j g'\left(\frac{p_j}{q_j}\right). \quad (10.22)$$

The log scoring rule is $s(R, P) = -\sum_{i=1}^{m} r_i \log p_i$. Substituting (10.22) and the log scoring rule into (10.8) shows that the equality holds with $a = h'\{g(1)\}\, g'(1)$. The functions h and g enter only through their values and first derivatives at 1. □

We define the Bregman weighted scoring families as the quasi-Bregman weighted scoring families with $h(x) = x$. By substituting (10.18) into (10.5) and using equivalence, the Bregman weighted scoring families take the simple form

$$s(R, P \parallel Q) = -\sum_{i=1}^{m} f(q_i)\left\{g\left(\frac{p_i}{q_i}\right) + g'\left(\frac{p_i}{q_i}\right)\frac{r_i - p_i}{q_i}\right\}. \quad (10.23)$$

We recover the unweighted Bregman scoring rules of Grünwald and Dawid (2004), i.e.,

$$s(R, P) = -\sum_{i=1}^{m} \left\{\tilde{g}(p_i) + \tilde{g}'(p_i)(r_i - p_i)\right\}, \quad (10.24)$$

by using a flat baseline and rescaling g to $\tilde{g}(p_i) = f(q_i)g(p_i/q_i) = f(m^{-1})g(mp_i)$. The unweighted Bregman scoring rules are uniquely specified through the convex function $\tilde{g}$ alone.

Corollary 10.5. *The unweighted Bregman rule specified by $\tilde{g}$ is compatible with all weighted Bregman families with $f(x) = x^2\tilde{g}''(x)$. This holds for any twice differentiable and strictly convex g.*

Proof. We use (10.8) with $s(R, P \parallel Q)$ given by (10.23) with $f(x) = x^2\tilde{g}''(x)$ and $s(R, P)$ given by (10.24). □

We illustrate the use of this corollary via an example. The unweighted power scoring rule is defined by $\tilde{g}(x) = x^\beta/\{\beta(\beta-1)\}$ for $\beta > 1$. Using the above corollary with $f(x) = x^\beta$, we see that the unweighted power scoring rule is compatible with all weighted scoring families of the form

$$s(R, P \parallel Q) = -\sum_{i=1}^{m} q_i^\beta \left\{ g\left(\frac{p_i}{q_i}\right) + g'\left(\frac{p_i}{q_i}\right)\frac{r_i - p_i}{q_i} \right\}, \tag{10.25}$$

for any choice of g.

§10.4 Discussion

As an application of compatible proper scoring rules, consider a portfolio allocation problem similar to Jose et al. (2008). There is a market consisting of m assets, and a market maker who sets the prices at Q. After one time period asset X will be worth 1 unit and the other assets will be worthless. The investor purchases a portfolio, spending a proportion of his wealth $p_i(\theta)$ on each asset and thus receiving $p_i(\theta)/q_i(\theta)$ units of each asset. He chooses θ based on the current prices Q and the historical outcome distribution R. Suppose the investor's risk is given by a weighted scoring rule $s\{R, P(\theta) \parallel Q\}$. The market maker does not know the form of the investor's scoring rule, but he believes it to come from a weighted scoring family compatible with some known proper scoring rule. The market maker prices the assets using the compatible rule's optimal score estimator, $Q = P\{\tilde{\theta}(R)\}$. Then the market maker's price coincides with the investor's minimal risk portfolio $P(\theta)$: when the pricing is done by a compatible proper scoring rule, the investor is best served by buying the

same number of units of each asset. Johnstone (2011) interprets this minimal risk portfolio from an economic perspective, for the special case where the compatible rule is the log score.

Until now, the only weighted scoring families considered in the literature were the weighted power and pseudospherical scoring rules. Since both are compatible with the log scoring rule, their optimal score estimators are dominated by the maximum likelihood estimator when the baseline is given by the latter. Johnstone and Lin (2011) conjectured the existence of a characterization theorem for all weighted proper scoring families whose optimal score estimators are dominated in this way. They suggested that this theorem might reveal an unrecognized property of the log scoring rule.

We have found their conjectured characterization theorem: the optimal score estimator of any weighted proper scoring rule is dominated by the compatible proper scoring rule's optimal score estimator when the baseline is set to the compatible proper scoring rule's optimal score estimate. However, instead of revealing a special property of the log scoring rule, we have shown that every proper scoring rule is compatible with some family of weighted proper scoring rules.

Chapter Eleven

Score matching estimator

The increasing interest in analysis of high-dimensional data has necessitated the development of parsimonious multivariate models with reliable, computationally efficient estimation procedures. For example, sparse Gaussian graphical models (Dobra et al., 2004; Ma et al., 2007; Rothman et al., 2008; Bickel and Levina, 2008; Chandrasekaran et al., 2012) have drawn significant interest and several computationally efficient estimation procedures have been developed (Banerjee et al., 2006, 2008; Friedman et al., 2008). Gaussian graphical models with symmetry (Højsgaard and Lauritzen, 2008) form another example, though no efficient estimation procedures have yet been developed. Here we describe and exploit a method which provides linear estimating equations when applied to any exponential family, in particular to any Gaussian graphical model with symmetry. In contrast to the maximum likelihood estimator, which often requires iterative methods, this *score matching estimator* is computationally efficient for such families. Even when the maximum likelihood estimator is desired, it must be computed iteratively and the score matching estimator may provide a useful initial value for the iterations.

This chapter is a slight modification of Forbes and Lauritzen (2014).

§11.1 Preliminaries

Consider a random quantity taking values in an open subset $\mathscr{X}$ of $\mathbb{R}^p$, which is equipped with the standard inner product $\langle \cdot, \cdot \rangle_p$, the associated norm $\|\cdot\|_p$, and the canonical basis. Throughout the chapter, $\mathscr{P}$ denotes a class of distributions over $\mathscr{X}$ with twice continuously differentiable densities with respect to the Lebesgue measure on $\mathscr{X}$. The general developments below are equally valid for $\mathscr{X}$ being a Riemannian manifold with associated geometric measure (Dawid and Lauritzen, 2005), but as our main focus is the multivariate Gaussian distribution we shall refrain from working at this level of generality. We use ∇ for the gradient and Δ for the Laplace operator on $\mathscr{X}$ so that

$$\nabla f(x) = \left\{ \frac{\partial}{\partial x_i} f(x) \right\} \in \mathbb{R}^p, \quad \Delta f(x) = \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2} f(x). \tag{11.1}$$

§11.1.1 Scoring rules

A *scoring rule* $S(x, Q)$ is a real-valued function which quantifies the accuracy of a predictive distribution $Q \in \mathscr{P}$ upon observing the realized value $x \in \mathscr{X}$. It is (strictly) *proper* if the expected value over X with respect to P, $\mathbb{E}_{X \sim P} S(X, Q)$, is (uniquely) minimized over $\mathscr{P}$ at $Q = P$. Two scoring rules are *equivalent* if they differ by a positive scalar multiple and a function of x alone.

Every proper scoring rule induces a *divergence* (Dawid, 1998; Grünwald and Dawid, 2004):

$$d(P, Q) = \mathbb{E}_{X \sim P}\{S(X, Q) - S(X, P)\}. \tag{11.2}$$

The divergences of equivalent scoring rules are proportional. A much used scoring rule is the *log score* $S(x, Q) = -\log g(x)$, where g is the density of Q (Bernardo, 1979; Good, 1952; McCarthy, 1956); the corresponding divergence is then the Kullback–Leibler divergence.

Given an independent sample $x^1, \ldots, x^n$ with empirical distribution $\widehat{P}$ and unknown distribution Q, the *optimal score estimator* (Gneiting and Raftery, 2007) $\widehat{Q}$

of Q is determined as the minimizer of the empirical score

$$\widehat{Q} = \operatorname*{argmin}_{Q \in \mathscr{P}} \mathbb{E}_{X \sim \widehat{P}}\{S(X,Q)\} = \operatorname*{argmin}_{Q \in \mathscr{P}} \sum_{i=1}^{n} S(x^i, Q) = \operatorname*{argmin}_{Q \in \mathscr{P}} d(\widehat{P}, Q). \tag{11.3}$$

The first two expressions are well-defined even when $\widehat{P} \notin \mathscr{P}$ whereas the latter may not be. Dawid and Lauritzen (2005) show that for a parametrized family $\mathscr{P} = \{Q_\theta : \theta \in \Theta\}$ with Θ being an open subset of $\mathbb{R}^d$, this minimization gives rise to an *unbiased estimating equation* (Godambe, 1991)

$$\sum_{i=1}^{n} S'(x^i, \theta) = 0, \tag{11.4}$$

where $S'(x,\theta)$ is the vector of derivatives of $S(x, Q_\theta)$ with respect to θ. Solutions to such equations are also known as M-estimators (Huber, 1964, 1967) and these are typically consistent and asymptotically normal although not necessarily efficient. If $S(x,Q) = -\log g(x)$ is the log score, the equation (11.4) is the likelihood equation and the corresponding estimator is the maximum likelihood estimator (MLE).

§11.1.2 Score matching estimator

Suppose the density g of $Q \in \mathscr{P}$ is twice continuously differentiable and satisfies the regularity assumptions:

$$\mathbb{E}_{X \sim P} \|\nabla \log g(X)\|_p^2 < \infty \text{ for all } P, Q \in \mathscr{P}, \tag{11.5}$$

$$g(x) \text{ and } \|\nabla g(x)\|_p \text{ tend to zero as } x \text{ approaches the boundary of } \mathscr{X}. \tag{11.6}$$

Then, using integration by parts, Hyvärinen (2005, 2007) showed that the divergence function

$$d_2(P,Q) = \mathbb{E}_{X \sim P} \|\nabla \log g(x) - \nabla \log f(x)\|_p^2, \tag{11.7}$$

where f is the density of P, is induced by the scoring rule

$$S_2(x,Q) = \frac{1}{2} \|\nabla \log g(x)\|_p^2 + \Delta \log g(x). \tag{11.8}$$

This scoring rule can be shown to be proper (Dawid, 2007; Dawid and Lauritzen, 2005). The *score matching estimator* (SME) is the optimal score estimator for this

scoring rule. Note that the SME is not invariant under transformations of x, nor under change of base measure. Hence care must be taken when choosing the representation of the data to ensure the resulting estimator is suitable for its intended purpose.

§11.2 Exponential families

As indicated in (Hyvärinen, 2007), the SME is particularly simple when $\mathscr{P}$ is an exponential family. In this section, we briefly review some theory about exponential families and prove some results of the SME when it is applied to such a family. Let $g(x \mid \theta)$ be a family of densities defined by

$$\log g(x \mid \theta) = \langle \theta, t(x) \rangle_d - a(\theta) + b(x), \quad \theta \in \Theta. \tag{11.9}$$

Here $t(x) \in L$ is the canonical sufficient statistic, L is a d-dimensional vector space, $\langle \cdot, \cdot \rangle_d$ an inner product on L, and $\Theta \subseteq L$ is the (convex) canonical parameter space

$$\Theta = \operatorname{int} \left\{ \theta \in L : a(\theta) = \log \int_{\mathscr{X}} e^{\langle \theta, t(x) \rangle_d + b(x)} \, \mathrm{d}x < \infty \right\}. \tag{11.10}$$

Let $D(x)$ be the linear map from L to $\mathbb{R}^p$ determined by the equation $D(x)\eta = \nabla \langle \eta, t(x) \rangle_d$ for all $\eta \in L$. Then we have

$$\nabla \log g(x \mid \theta) = D(x)\theta + \nabla b(x). \tag{11.11}$$

Further we get

$$\Delta \log g(x \mid \theta) = \langle \theta, \Delta t(x) \rangle_d + \Delta b(x), \tag{11.12}$$

where, similarly, Δt is given by $\langle \eta, \Delta t(x) \rangle_d = \Delta \langle \eta, t(x) \rangle_d$. We assume that the regularity conditions (11.5) and (11.6) are satisfied; thus in particular both $\mathbb{E}_\theta \|D(X)\eta\|_p^2$ and $\mathbb{E}_\theta \|\nabla b(X)\|_p^2$ are finite, where we write $\mathbb{E}_\theta$ for the expectation of X with respect to the density $g(x \mid \theta)$.

Let $D(x)^*$ be the transpose of $D(x)$ given by $\langle y, D(x)\eta \rangle_p = \langle D(x)^* y, \eta \rangle_d$ for all $\eta \in L$ and $y \in \mathbb{R}^p$. We furthermore assume that the linear map $\Psi(\theta) = \mathbb{E}_\theta\{D(X)^* D(X)\}$ from L to L is invertible so the corresponding quadratic form

$$\langle \eta, \Psi(\theta)\eta \rangle_d = \mathbb{E}_\theta \|D(X)\eta\|_p^2 \tag{11.13}$$

is positive definite, i.e. it is non-zero unless $\eta = 0$. The objective function $J_2(\theta) = \sum_{i=1}^n S_2(x^i, Q_\theta)$ becomes

$$\sum_{i=1}^{n} \frac{1}{2} \left\| D(x^i)\theta \right\|_p^2 + \left\langle \theta, D(x^i)^* \nabla b(x^i) \right\rangle_d + \left\langle \theta, \Delta t(x^i) \right\rangle_d + \frac{1}{2} \left\| \nabla b(x^i) \right\|_p^2 + \Delta b(x^i). \tag{11.14}$$

The last two terms depend on x only and will henceforth be ignored: this yields an equivalent scoring rule and does not alter the SME.

The objective function J_2 depends quadratically on θ and the minimizer of J_2 is unique if and only if the quadratic form on L

$$D_2(\eta) = \sum_{i=1}^{n} \left\| D(x^i)\eta \right\|_p^2 \tag{11.15}$$

is positive definite, i.e. if $D(x^i)\eta = 0$ for $i = 1, \ldots, n$ implies $\eta = 0$. The SME is the minimizer of $J_2(\theta)$ over Θ, leading to the linear estimating equation for θ

$$J_2'(\theta) = \sum_{i=1}^{n} D(x^i)^* \left\{ D(x^i)\theta + \nabla b(x^i) \right\} + \Delta t(x^i) = 0. \tag{11.16}$$

Thus, provided $\sum_{i=1}^n D(x^i)^* D(x^i)$ is invertible, the score estimation equation has the unique solution

$$\check{\theta}_n = -\left\{ \sum_{i=1}^{n} D(x^i)^* D(x^i) \right\}^{-1} \sum_{i=1}^{n} \left\{ D(x^i)^* \nabla b(x^i) + \Delta t(x^i) \right\}. \tag{11.17}$$

In the case $b(x) = 0$ this is equivalent to the *variational estimator* (eq. 1.9) of Almeida and Gidas (1993) and to eq. 34 of (Hyvärinen, 2007). By taking the inner product of (11.16) with the minimizer $\check{\theta}_n$ we further obtain

$$\sum_{i=1}^{n} \left\| D(x^i)\check{\theta}_n \right\|_p^2 = -\sum_{i=1}^{n} \left\{ \left\langle \check{\theta}_n, D(x^i)^* \nabla b(x^i) \right\rangle_d + \left\langle \check{\theta}_n, \Delta t(x^i) \right\rangle_d \right\}. \tag{11.18}$$

Inserting this relation into the expression for $J_2(\check{\theta}_n)$ yields a linear dependence of the minimal value of J_2 on $\check{\theta}_n$ with the simple expression

$$J_2(\check{\theta}_n) = \sum_{i=1}^{n} \left\langle \check{\theta}_n, D(x^i)^* \nabla b(x^i) + \Delta t(x^i) \right\rangle_d / 2, \tag{11.19}$$

where we have ignored the terms depending on x alone.

As mentioned earlier, SMEs are M-estimators and thus typically consistent, as also claimed in Corollary 3 of (Hyvärinen, 2005). However, the argument in (Hyvärinen, 2005) is incomplete since the convergence of J_2/n to its expectation does not imply that the minimizer $\operatorname{argmin}_\phi J_2(\phi)$ converges to θ without additional conditions on J_2. General consistency results are established under suitable regularity conditions in (Huber, 1967), but for an exponential family we can establish consistency directly.

Proposition 11.1. *Assume $\mathscr{P} = P_\theta, \theta \in \Theta$ is an exponential family as above and $X^1, \ldots, X^n$ is a sample of size n from P_θ. Then the SME $\check{\theta}_n$ is asymptotically consistent for θ.*

Proof. Each term in the last sum of (11.17) has expectation

$$\mathbb{E}_\theta\{D(X)^*\nabla b(X) + \Delta t(X)\} = \int g(x \mid \theta) D(x)^* \nabla b(x)\,\mathrm{d}x + \int g(x \mid \theta)\Delta t(x)\,\mathrm{d}x \tag{11.20}$$

and both integrals on the right-hand side are finite by our assumptions. Using integration by parts on the second term yields

$$\int g(x \mid \theta)\Delta t(x)\,\mathrm{d}x = -\int D(x)^* \nabla g(x \mid \theta)\,\mathrm{d}x, \tag{11.21}$$

since the boundary terms vanish by (11.6). Using that $\nabla g(x) = g(x)\nabla \log g(x)$ and substituting the expression (11.11) for $\log g(x \mid \theta)$ on the right-hand side yields

$$\int g(x \mid \theta)\Delta t(x)\,\mathrm{d}x = -\int g(x \mid \theta) D(x)^* \{D(x)\theta + \nabla b(x)\}\,\mathrm{d}x, \tag{11.22}$$

which then gives

$$\mathbb{E}_\theta\{D(X)^*\nabla b(X) + \Delta t(X)\} = -\mathbb{E}_\theta\{D(X)^* D(X)\}\theta = -\Psi(\theta)\theta. \tag{11.23}$$

Thus, by the law of large numbers

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left\{D(X^i)^* \nabla b(X^i) + \Delta t(X^i)\right\} = -\Psi(\theta)\theta \tag{11.24}$$

with probability one. Similarly, for the first sum in (11.17) we get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left\{D(X^i)^* D(X^i)\right\} = \Psi(\theta) \tag{11.25}$$

which is invertible by assumption. Hence the estimate $\breve{\theta}_n$ converges to θ and is thus consistent for θ, as desired. □

Under a few additional assumptions, we can also show that the SME is asymptotically normally distributed.

Proposition 11.2. *Assume $\mathscr{P} = P_\theta, \theta \in \Theta$ is an exponential family as above and $X^1, \ldots, X^n$ is a sample of size n from P_θ. Assume further that all of $\mathbb{E}_\theta \|D(X)^* \nabla b(X)\|_d^2$, $\mathbb{E}_\theta \|\Delta t(X)\|_d^2$, and $\mathbb{E}_\theta \|D(X)^* D(X)\theta\|_d^2$ are finite. Then $\sqrt{n}(\breve{\theta}_n - \theta)$ converges in distribution to a normal distribution on L with mean zero.*

Proof. From (11.17) we get

$$\sqrt{n}(\breve{\theta}_n - \theta) = -\Psi_n^{-1} \sum_{i=1}^{n} \left\{ D(X^i)^* \nabla b(X^i) + \Delta t(X^i) + D(X^i)^* D(X^i)\theta \right\} / \sqrt{n}, \quad (11.26)$$

where $\Psi_n = \frac{1}{n} \sum_{i=1}^n D(X^i)^* D(X^i)$.

As in the proof of proposition 11.1, we conclude that Ψ_n converges in probability to its expectation $\Psi(\theta)$, which is invertible by assumption. The i^{th} term of the sum in (11.26) has expectation zero and finite variance and hence, by the Central Limit Theorem, the normalized sum converges in distribution to a normal distribution on L with mean zero. By Slutsky's theorem, so does $\sqrt{n}(\breve{\theta}_n - \theta)$, as desired. □

The asymptotic inverse covariance of the SME is $nG(\theta)$ where $G(\theta)$ is the *Godambe information* (Godambe, 1991), $G(\theta) = \Psi(\theta) H(\theta)^{-1} \Psi(\theta)$, and

$$H(\theta) = \mathbb{V}_\theta \left\{ D(X)^* D(X)\theta + D(X)^* \nabla b(X) + \Delta t(X) \right\}, \quad (11.27)$$

where $\mathbb{V}_\theta$ is the variance with respect to the density $g(x \mid \theta)$; see for example (Barndorff-Nielsen and Cox, 1994, Section 9.2) or (van der Vaart, 1998, Theorem 5.21). Note that, as we have shown above, Ψ_n is a consistent estimator of $\Psi(\theta)$ and $H(\theta)$ can be estimated consistently by the corresponding empirical covariance.

For finite n it is possible that $\breve{\theta} \notin \Theta$. Even in this case $\breve{\theta}$ itself may be useful: the value of $J_2(\breve{\theta})$ can be very quickly computed and used for model screening, or $\breve{\theta}$ can

be used as a starting value for iterative estimation methods. Consistency ensures that $\check{\theta} \in \Theta$ for n sufficiently large.

As noted earlier, the score matching equation is not invariant under data reparametrization, nor under change of base measure. We could in principle use the estimating equation after reducing to the sufficient statistic $t = \sum_i t(x^i)$, which has density $\tilde{g}(t \mid \theta)$ where

$$\log \tilde{g}(t \mid \theta) = \langle \theta, t \rangle_d - na(\theta) + h_n(t), \tag{11.28}$$

and where $\exp\{h_n(t)\}$ is the density of the product of measures $\exp\{b(x^i)\}\,\mathrm{d}x^i$ transformed by the sufficient statistic. Then the SME becomes $\check{\theta} = -\nabla h_n(t)/n$, which coincides with both Martin-Löf's exact estimator (Martin-Löf, 1977) and the maximum plausibility estimator (Barndorff-Nielsen, 1976) for this case. The exact estimator is known to be consistent and efficient (Höglund, 1974). Unfortunately, the form of $h_n(t)$ is often intractable and the exact estimator is often more difficult to calculate than the MLE.

We have chosen not to reduce by sufficiency: our SME may lose statistical efficiency, but it gains computational speed from the simplicity of its estimating equations.

§11.3 Gaussian linear concentration models

We now consider Gaussian models with linear structure in the concentration matrix (Anderson, 1970), exploiting that they are special instances of the exponential families discussed above.

Let L be a d-dimensional linear subspace of $\mathscr{S}^p$, the symmetric $p \times p$ matrices equipped with the trace inner product $\langle A, B \rangle_d = \mathrm{tr}(AB)$ and associated *Frobenius* norm $\|A\|_d^2 = \mathrm{tr}(A^2)$. This is the subspace which will contain the set of candidate models for

the concentration matrix K. Consider the family of Gaussian densities

$$\begin{aligned}\log p(x \mid K) &= \{\log\det(K) - p\log(2\pi) - \langle x, Kx\rangle_p\}/2 \\ &= -\left\langle K, xx^\top\right\rangle_d/2 + \{\log\det(K) - p\log(2\pi)\}/2 \\ &= \left\langle K, -\Pi_L(xx^\top)/2\right\rangle_d + \{\log\det(K) - p\log(2\pi)\}/2\end{aligned} \tag{11.29}$$

which clearly has the form (11.9) with $\mathscr{X} = \mathbb{R}^p$, $\theta = K$, $a(K) = p\log(2\pi)/2 - \log\det(K)/2$, $b(x) = 0$, and $t(x) = -\Pi_L(xx^\top)/2$, where Π_L is the orthogonal projection onto L in $\mathscr{S}^p$. The canonical parameter space is $\Theta = L \cap \mathscr{S}^p_+$, where $\mathscr{S}^p_+$ is the set of positive definite symmetric $p \times p$ matrices (Barndorff-Nielsen, 1978, p. 116). The maps discussed above become

$$D(x)K = -Kx, \quad D(x)^* y = -\Pi_L(xy^\top + yx^\top)/2, \quad \Delta t(x) = -\Pi_L(I_p), \tag{11.30}$$

where I_p is the $p \times p$ identity matrix. For simplicity we assume in the following that $I_p \in \Theta$; this can always be achieved for non-empty Θ by choosing a suitable basis for $\mathbb{R}^p$. Then we have $\Pi_L(I_p) = I_p$ so the Laplacian becomes $\Delta t(x) = -I_p$.

To see that (11.30) holds, note that for any $K \in L$ we have

$$\nabla\left\langle K, \Pi_L(xx^\top)\right\rangle_d = \nabla\left\langle K, xx^\top\right\rangle_d = \nabla\operatorname{tr}(Kxx^\top) = 2Kx \tag{11.31}$$

so $D(x)K = -Kx$. Furthermore we have

$$\begin{aligned}\langle D(x)^* y, K\rangle_d &= \langle y, D(x)K\rangle_p = -\langle y, Kx\rangle_p = -\operatorname{tr}(xy^\top K) \\ &= -\left\langle xy^\top + yx^\top, K\right\rangle_d/2 = \left\langle -\Pi_L(xy^\top + yx^\top)/2, K\right\rangle_d\end{aligned} \tag{11.32}$$

and finally

$$\Delta\left\langle K, \Pi_L(xx^\top)\right\rangle_d = \Delta\left\langle K, xx^\top\right\rangle_d = 2\operatorname{tr}(K) = 2\left\langle K, I_p\right\rangle_d. \tag{11.33}$$

In particular we get

$$D(x)^* D(x)K = -D(x)^* Kx = \Pi_L(xx^\top K + Kxx^\top)/2 = \Pi_L(K \circ xx^\top), \tag{11.34}$$

where $A \circ B = (AB^\top + BA^\top)/2$ is the *Jordan product* (Albert, 1946) of the symmetric matrices A and B. The estimating equation (11.16) now specializes to

$$\Pi_L(K \circ W) = I_p, \tag{11.35}$$

where we have let $W = n^{-1}\sum_{i=1}^{n} x^i x^{i\top}$ be the scaled Wishart matrix of sums of squares and products. The expression (11.19) for J_2 becomes simply $-n\,\mathrm{tr}\,\check{K}/2$ which can be evaluated very quickly.

We next verify that $p(x\,|\,K)$ satisfies the assumptions for consistency and asymptotic normality. For consistency we must satisfy the regularity conditions (11.5) and (11.6). It is obvious that both $p(x\,|\,K)$ and $\|\nabla p(x\,|\,K)\|_p = |p(x\,|\,K)|\,\|Kx\|_p$ tend to zero as $\|x\|_p \to \infty$, and furthermore, for any $K_0 \in L$,

$$\mathbb{E}_K \|\nabla \log p(X\,|\,K_0)\|_p^2 = \mathbb{E}_K\{\mathrm{tr}(K_0 X X^\top K_0)\} = \mathrm{tr}(K_0 K^{-1} K_0) < \infty. \tag{11.36}$$

For asymptotic normality we note that $\mathbb{E}_K \|D(X)\nabla b(X)\|_d^2 = 0$ since $b(x) = 0$. Furthermore $\mathbb{E}_K \|\Delta t(X)\|_d^2 = \mathbb{E}_K\{\mathrm{tr}(I_p)\} = p < \infty$. Finally,

$$\begin{aligned}\mathbb{E}_K\{\|D(X)^* D(X) K\|_d^2\} &= \mathbb{E}_K[\{\mathrm{tr}\,\Pi_L(K \circ XX^\top)\}^2] \\ &\le \mathbb{E}_K[\{\mathrm{tr}(K \circ XX^\top)\}^2] \\ &= \mathbb{E}_K\{(X^\top K X)^2\},\end{aligned} \tag{11.37}$$

which is the fourth moment of a normal distribution and hence finite. Thus all the conditions for consistency and asymptotic normality are satisfied.

Having established asymptotic normality, we next derive the asymptotic covariance matrix of the SME and find its efficiency relative to the MLE. We introduce the covariance form on L defined as

$$\mathrm{Cov}_L(\check{K})[\eta_1, \eta_2] = \mathrm{Cov}\left\langle \eta_1, \check{K}\right\rangle_d, \left\langle \eta_2, \check{K}\right\rangle_d \tag{11.38}$$

for any $\eta_1, \eta_2 \in L$. Given a basis $e^1, \ldots, e^d$, this can be identified with a covariance matrix with entries $\mathrm{Cov}_L(\check{K})[e^u, e^v]$. Equation (11.26) implies

$$\lim_{n\to\infty} n\,\mathrm{Cov}_L(\check{K})[\eta_1, \eta_2] = C[\eta_1, \eta_2] = \mathrm{Cov}_L\{D(X)^* D(X) K\}\left[\Psi^{-1}(\eta_1), \Psi^{-1}(\eta_2)\right]. \tag{11.39}$$

From (11.34) we know that $D(X)^* D(X) K = \Pi_L(K \circ XX^\top)$ and thus

$$\begin{aligned} C[\eta_1, \eta_2] &= \mathrm{Cov}_L\{\Pi_L(K \circ XX^\top)\}\left[\Psi^{-1}(\eta_1), \Psi^{-1}(\eta_2)\right] \\ &= \mathrm{Cov}_L(K \circ XX^\top)\left[\Psi^{-1}(\eta_1), \Psi^{-1}(\eta_2)\right] \\ &= \mathrm{Cov}\left\langle K \circ XX^\top, \Psi^{-1}(\eta_1)\right\rangle_d, \left\langle K \circ XX^\top, \Psi^{-1}(\eta_2)\right\rangle_d \\ &= \mathrm{Cov}\left\langle XX^\top, K\Psi^{-1}(\eta_1)\right\rangle_d, \left\langle XX^\top, K\Psi^{-1}(\eta_2)\right\rangle_d \\ &= \mathrm{Cov}_{p\times p}(XX^\top)\left[K\Psi^{-1}(\eta_1), K\Psi^{-1}(\eta_2)\right], \end{aligned} \tag{11.40}$$

where $\mathrm{Cov}_{p\times p}(XX^\top)$ is the covariance form of $XX^\top$ over the $p \times p$ matrices. Now $XX^\top$ has a Wishart distribution with mean $\Sigma = K^{-1}$ and one degree of freedom, for which it is well known that $\mathrm{Cov}_{p\times p}(XX^\top)[\eta_1, \eta_2] = 2\,\mathrm{tr}(\eta_1 \Sigma \eta_2 \Sigma)$ (see, e.g. Lauritzen (1996, p.258)). Hence we have

$$C[\eta_1, \eta_2] = 2\,\mathrm{tr}\{\Psi^{-1}(\eta_2)\Psi^{-1}(\eta_1)\}. \tag{11.41}$$

Taking the expectation of (11.34) we get $\Psi(K) = I_p$, and by linearity $\Psi(K \circ \eta_1) = \eta_1$. Since Ψ is invertible by assumption, this yields $\Psi^{-1}(\eta_1) = K \circ \eta_1$ and thus

$$n\,\mathrm{Cov}_L(\check{K})[\eta_1, \eta_2] \to C[\eta_1, \eta_2] = 2\,\mathrm{tr}\{(K \circ \eta_1)(K \circ \eta_2)\}. \tag{11.42}$$

For comparison, the covariance form of the the maximum likelihood estimate can be obtained from the inverse-covariance form, which is the Fisher information $\mathscr{I}(\widehat{K})[\eta_1, \eta_2] = n\,\mathrm{tr}(\Sigma\eta_1\Sigma\eta_2)/2$ (Højsgaard and Lauritzen, 2008). Specifically, given a specific basis $e^1, \ldots, e^d$ for L, the covariance matrix (and hence the covariance form) can be found by inverting the matrix with entries $\mathscr{I}(\widehat{K})[e^u, e^v]$.

The relative efficiency of the score matching estimator for a specific model L and $\eta \in L$ is simply

$$\lim_{n\to\infty} \frac{\mathrm{Cov}_L(\widehat{K})[\eta, \eta]}{\mathrm{Cov}_L(\check{K})[\eta, \eta]}. \tag{11.43}$$

For instance, for the four cycle (figure 11.1) with $d = 2$ and

$$K = \sigma e^1 + \mu e^2 = \frac{\sigma}{2} I_4 + \frac{\mu}{2\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \tag{11.44}$$

the asymptotic covariance matrix of the SME is

$$\mathrm{Cov}[\check{K}] \to \frac{1}{2n}\begin{pmatrix} \sigma^2+\mu^2 & 2\sigma\mu \\ 2\sigma\mu & \sigma^2+2\mu^2 \end{pmatrix} \tag{11.45}$$

and the asymptotic covariance matrix of the MLE is

$$\mathrm{Cov}[\widehat{K}] \to \frac{\sigma^2}{2n(\sigma^2+\mu^2)}\begin{pmatrix} \sigma^2+2\mu^2 & 2\sigma\mu \\ 2\sigma\mu & \sigma^2-\mu^2+2\sigma^{-2}\mu^4 \end{pmatrix}. \tag{11.46}$$

Thus the relative efficiency for σ is $1-\mu^4(\sigma^2+\mu^2)^{-2}$ and the relative efficiency for μ is $1-4\sigma^2\mu^2/(\sigma^4+3\sigma^2\mu^2+2\mu^4)$.

§11.3.1 Jordan linear concentration models

Consider the special case where L is closed under the Jordan product, i.e. it forms a Jordan subalgebra of $\mathscr{S}^p$, or equivalently $\Theta = L \cap \mathscr{S}^p_+$ is closed under inversion. Such hypotheses are exactly those which are linear in both the covariance and inverse covariance (Jensen, 1988). In particular, L contains all models which are determined by invariance under a subgroup of the general linear group (Andersson, 1975). There are many such models, some examples are given in Jensen (1988).

We shall show that the MLE and the SME coincide for these models. First we need a lemma.

Lemma 11.3. *If L is a Jordan subalgebra of $\mathscr{S}^p$ then for any $A \in L$ and $B \in \mathscr{S}^p$ we have $\Pi_L(A \circ B) = A \circ \Pi_L(B)$.*

Proof. Let $B_0 = \Pi_L(B)$. Clearly we have $A \circ B_0 \in L$ since L is closed under the Jordan product. Further, for any $K \in L$ we have

$$\langle A \circ B - A \circ B_0, K\rangle_d = \mathrm{tr}(ABK) - \mathrm{tr}(AB_0K) = \langle B - B_0, K \circ A\rangle_d = 0, \tag{11.47}$$

where the last equality follows because $K \circ A \in L$. Thus $A \circ B - A \circ B_0$ is orthogonal to L and the lemma follows. □

We now obtain the desired result.

Theorem 11.4. *If the subspace L is a Jordan subalgebra then the SME is equal to the MLE. Furthermore, if $\Pi_L(W)$ is invertible we have*

$$\widehat{K} = \check{K} = \{\Pi_L(W)\}^{-1}. \tag{11.48}$$

Proof. The family is a full and canonical exponential family with $\Pi_L(W)$ as the canonical sufficient statistic, and hence the likelihood equation becomes

$$\Pi_L(W) = \mathbb{E}_K\{\Pi_L(W)\} = \Pi_L(K^{-1}) = K^{-1}. \tag{11.49}$$

This implies $\widehat{K} = \{\Pi_L(W)\}^{-1}$.

As L is a Jordan subalgebra we have $I_p \in L$. Using lemma 11.3, the score matching equation (11.35) reduces to

$$\Pi_L(K \circ W) = K \circ \Pi_L(W) = I_p, \tag{11.50}$$

whence we get $\check{K} = \{\Pi_L(W)\}^{-1}$. This completes the proof. □

§11.3.2 Existence and uniqueness

Having observed a sample $x = (x^1, \ldots, x^n)$, the score matching equation (11.35) has a unique solution if and only if any of the following equivalent conditions hold:

1. The quadratic form $D_2(K) = \sum_{i=1}^n \|Kx^i\|^2$ is positive definite on L
2. $K = 0$ is the only element of L which maps all x^i to zero
3. The kernel of the linear map $K \to \Pi_L(K \circ W)$ is trivial.

We say *the SME exists* if (11.35) has a unique solution, ignoring the fact that $\check{K}$ may not be positive definite. We have the following relation between existence of the SME and the MLE.

Proposition 11.5. *Consider a Gaussian linear concentration model and let W be an empirical covariance matrix as above. If the SME for K given W exists, then the MLE for W also exists.*

Proof. We proceed by assuming that the MLE for W does not exist and showing that the SME does not exist either. If the MLE does not exist, the convex level set of the likelihood function

$$C = \{K \in \Theta = L \cap \mathcal{S}_+^p : \ell(K) \geq \ell(I_p)\} \tag{11.51}$$

must be unbounded. Here $\ell(K)$ is proportional to the logarithm of the likelihood function

$$\ell(K) = \log\det(K) - \operatorname{tr}(KW), \tag{11.52}$$

and we have assumed without loss of generality that $I_p \in L$. This implies (Barndorff-Nielsen, 1978, Theorem 5.5) that there is an $A \in \Theta$ such that $I_p + tA \in C$ for all $t \geq 0$:

$$\ell(I_p + tA) \geq \ell(I_p) \text{ for all } t \geq 0, \tag{11.53}$$

or equivalently that

$$\log\det(I_p + tA) \geq t \operatorname{tr}(AW) \text{ for all } t \geq 0. \tag{11.54}$$

But if a_i are the positive eigenvalues of A we have

$$\log\det(I_p + tA) = \sum_{i=1}^{p} \log(1 + ta_i) \tag{11.55}$$

and for large t this grows slower with t than any line through the origin with positive slope. Thus we must have $\operatorname{tr}(AW) = \operatorname{tr}(A \circ W) \leq 0$. Hence, since A, W, and thus $A \circ W$ are positive semidefinite, we have $A \circ W = 0$ and thus $\Pi_L(A \circ W) = 0$. Thus the third condition for the existence of the SME does not hold. □

By choosing an orthogonal basis $e^1, \ldots, e^d$ for L, we can write the matrix for the quadratic form D_2 as $M(x) = \{m_{uv}(x)\}$ where

$$m_{uv}(x) = \sum_{i=1}^{n} \left\langle e^u x^i, e^v x^i \right\rangle_p = n \operatorname{tr}(e^u W e^v). \tag{11.56}$$

Hence D_2 is positive definite if and only if $\det M(x) > 0$. This determinant is a polynomial in x and hence it either holds that $\det M(x) = 0$ for all x or $\det M(x) > 0$ except for a set of Lebesgue measure zero (Okamoto, 1973). In other words, either

the SME exists with probability one, or else it never exists. This is in contrast to the MLE, which can exist with some probability strictly between zero and one (Buhl, 1993; Uhler, 2012).

We shall say that the linear space L is *n-estimable* if the SME exists with probability one, or equivalently, if there is an $x = (x^1, \dots, x^n) \in \mathbb{R}^{p\times n}$ such that $\det M(x) > 0$.

For $n \geq p$ it is well-known that W is positive definite with probability one and hence $M(x)$ is positive definite and any L is n-estimable. We may thus without loss of generality assume $n < p$ in the following. As many high-dimensional data sets have n much less than p, this case is highly relevant. Let $r = p - n$ and $T_k = k(k+1)/2$ denote the k^{th} triangular number.

Proposition 11.6. *Let L be a linear subspace of $\mathcal{S}^p$. If $\dim L > T_p - T_r$ then L is not n-estimable.*

Proof. Let $\mathbb{X} = \operatorname{span}\{x^1, \dots, x^n\}$ and let $\mathcal{S}_0^p(\mathbb{X}) = \{K \in \mathcal{S}^p : \mathbb{X} \subseteq \ker(K)\}$ be the space of symmetric matrices that send all vectors in $\mathbb{X}$ to zero. Since $\dim(\mathbb{X}) = n$ with probability one, we have $\dim\{\mathcal{S}_0^p(\mathbb{X})\} = T_r$. Noticing that the quadratic form D_2 is positive definite over L if and only if $L \cap \mathcal{S}_0^p(\mathbb{X}) = \{0\}$, and that

$$\dim\{L \cap \mathcal{S}_0^p(\mathbb{X})\} = \dim L + \dim\{\mathcal{S}_0^p(\mathbb{X})\} - \dim\{\mathcal{S}_0^p(\mathbb{X}) + L\} \geq d + T_r - T_p, \quad (11.57)$$

we see that D_2 is not positive definite if $T_r + d > T_p$. □

Unfortunately the converse is not always true and it may happen that L is not n-estimable even if $\dim L \leq T_p - T_r$. In particular we note that the subspace $L = \operatorname{span}\{e_1, e_2, e_3, e_4\}$, defined by

$$L = \left\{ ae^1 + be^2 + ce^3 + fe^4 = \begin{pmatrix} a & c & 0 & f \\ c & b & -f & 0 \\ 0 & -f & a & c \\ f & 0 & c & b \end{pmatrix} : a, b, c, f \in \mathbb{R} \right\}, \quad (11.58)$$

yields a counterexample. We have $p = 4$ and $\dim L = 4$ so if we consider a single observation, we have $T_p - T_r = T_4 - T_3 = 4 = \dim L$. Letting $x = (x_1, x_2, x_3, x_4)$ be

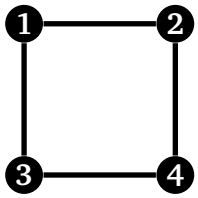


Figure 11.1: A four-cycle Gaussian graphical model.

our single observation, the corresponding quadratic form $m_{uv}(x)$ is

$$\begin{pmatrix} x_1^2+x_3^2 & 0 & x_1x_2+x_3x_4 & x_1x_4-x_2x_3 \\ 0 & x_2^2+x_4^2 & x_1x_2+x_3x_4 & x_1x_4-x_2x_3 \\ x_1x_2+x_3x_4 & x_1x_2+x_3x_4 & x_1^2+x_2^2+x_3^2+x_4^2 & 0 \\ x_1x_4-x_2x_3 & x_1x_4-x_2x_3 & 0 & x_1^2+x_2^2+x_3^2+x_4^2 \end{pmatrix}. \tag{11.59}$$

Direct computation shows that $m_{uv}(x)$ is singular since it has the zero eigenvector

$$\left(x_2^2+x_4^2,\quad x_1^2+x_3^2,\quad -x_1x_2-x_3x_4,\quad x_2x_3-x_1x_4\right). \tag{11.60}$$

This particular L, investigated by Jensen (1988), is an example of a Jordan subalgebra and we conclude — as Jensen — that the MLE also fails to exist.

Gaussian graphical models (Dempster, 1972) are special instances of linear concentration models. For example, in the model given by the four-cycle (figure 11.1), the MLE can only be calculated with iterative methods. If $n = 2$, the MLE may or may not exist, whereas for $n = 3$ the MLE always exists (Buhl, 1993). For $n = 2$ we have $T_p - T_r = 10 - 3 = 7$, and since $\dim L = 8$, proposition 11.6 yields that the SME does not exist. For $n = 3$ and above both the SME and MLE exist with probability one, the SME being a solution to a system of eight linear equations.

In the following we list a number of facts about n-estimability which may assist in determining whether a given subspace L is n-estimable. In particular, we show that a subspace of an n-estimable space is n-estimable, and that a change of coordinate system does not affect n-estimability.

Lemma 11.7. *If L is n-estimable and $L_0 \subseteq L$, then L_0 is n-estimable. If L is n-estimable with $n' > n$, then L is n'-estimable.*

Proof. The first statement follows since $L_0 \cap \mathscr{S}_0^p(\mathbb{X}) \subseteq L \cap \mathscr{S}_0^p(\mathbb{X}) = \{0\}$. If we let $\mathbb{Y} = \operatorname{span}\{x^1, \dots, x^{n'}\}$ we have $\mathscr{S}_0^p(\mathbb{Y}) \subseteq \mathscr{S}_0^p(\mathbb{X})$ and the second statement follows since $L \cap \mathscr{S}_0^p(\mathbb{Y}) \subseteq L \cap \mathscr{S}_0^p(\mathbb{X}) = \{0\}$. □

Lemma 11.8. *If $A \in \mathbb{GL}^p$ is invertible, then L is n-estimable if and only if $L_1 = ALA^\top$ is n-estimable.*

Proof. This follows as $L_1 \cap \mathscr{S}_0^p(\mathbb{X}) = L \cap \mathscr{S}_0^p(A^\top \mathbb{X})$. □

We next identify n-estimability with the ability to transform L into what we call *standard form*. This condition may be easier to check in some situations. We first identify $K \in \mathscr{S}^p$ with $A \in \mathscr{S}^n, B \in \mathbb{R}^{r \times n}$ and $C \in \mathscr{S}^r$ via

$$K = \begin{pmatrix} A & B^\top \\ B & C \end{pmatrix}. \tag{11.61}$$

Denote by $\mathscr{S}_r^p$ the subspace of $\mathscr{S}^p$ with $C = 0$. We then have $\dim \mathscr{S}^p = T_p$, $\dim \mathscr{S}^r = T_r$, and $\dim \mathscr{S}_r^p = T_p - T_r$. We say that L has *n-standard form* if

$$L = \left\{ \begin{pmatrix} A & B^\top \\ B & F(A,B) \end{pmatrix} : A \in \mathscr{S}^n, B \in \mathbb{R}^{r \times n}, \right\} \tag{11.62}$$

for some linear function $F : \mathscr{S}^n \times \mathbb{R}^{r \times n} \to \mathscr{S}^r$. Note that if L has n-standard form then we have $d = \dim L = \dim \mathscr{S}_r^p = T_p - T_r$.

Lemma 11.9. *If L has n-standard form then L is n-estimable.*

Proof. If L has n-standard form we can choose $x = (x^1, \dots, x^n)$ as the first n standard basis vectors $e^1, \dots, e^n$ of $\mathbb{R}^p$. Then for any K of the form (11.62) we have $Ke^i = 0$ for all $i = 1, \dots, n$ if and only if $A = 0$ and $B = 0$. Hence we must have $K = 0$, so the quadratic form $D_2(K)$ is positive definite. □

Corollary 11.10. *If $L \subseteq L_0$ and for some $A \in \mathbb{GL}^p$, $L_1 = AL_0A^\top$ has n-standard form, then L is n-estimable.*

Proof. This follows by combining lemma 11.7, lemma 11.8, and lemma 11.9. □

We also note that the converse to corollary 11.10 holds.

Lemma 11.11. *If L is n-estimable then there exists $A \in \mathbb{GL}^p$ and $L_0 \supseteq L$ such that $L_1 = AL_0A^\top$ has n-standard form.*

Proof. Let n be the smallest integer m such that L is m-estimable. Then there exist orthogonal vectors $x^1 \dots, x^n$ such that $W : K \to K \circ \sum_1^n x^i x^{i\top}/n$ has full rank. Let A denote the transformation to a coordinate system where $x^1, \dots, x^n$ are the first n basis vectors. In this coordinate system we have

$$W(K) = K \circ \sum_1^n x^i x^{i\top}/n = \begin{pmatrix} K_{11} & K_{12}/2 \\ K_{21}/2 & 0 \end{pmatrix}/n, \tag{11.63}$$

where

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \tag{11.64}$$

with $K_{11} \in \mathscr{S}^n, K_{12} \in \mathbb{R}^{n\times r}$ and $K_{22} \in \mathscr{S}^r$. Since L is n-estimable, this map is injective and thus we must have $K_{22} = F(K_{11}, K_{12})$ for some linear function F. Thus in this basis $L \subseteq L_0$ where L_0 has standard form. □

Finally we show that all non-trivial L with $d \leq T_p - T_r$ and $r = p - n = 1$ are n-estimable.

Proposition 11.12. *Suppose that L is a linear subspace of $\mathscr{S}^p$ with $L \cap \mathscr{S}^p_+ \neq \emptyset$ and that $\dim L \leq T_p - T_r$ with $r = 1$. Then L is n-estimable.*

Proof. For contradiction, assume $d = \dim L = T_p - 1$ and that L is not n-estimable for $n = p - 1$. Thus for *any* $x^1, \dots, x^n$, the map

$$W : K \to K \circ \sum_1^n x^i x^{i\top}/n \tag{11.65}$$

has rank less than d. Assume now that $x^i = e^i, i = 1, \dots, n$ are orthonormal so that $e^1, \dots, e^p$ form an orthonormal basis for $\mathbb{R}^p$. In this basis we have

$$\sum_1^n e^i e^{i\top} = \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix}. \tag{11.66}$$

and thus

$$W(K) = K \circ \sum_1^n e^i e^{i\top}/n = \begin{pmatrix} K_{11} & k/2 \\ k^\top/2 & 0 \end{pmatrix}/n, \tag{11.67}$$

where

$$K = \begin{pmatrix} K_{11} & k \\ k^\top & k_{22} \end{pmatrix}, \tag{11.68}$$

with $K_{11} \in \mathcal{S}^n, k \in \mathbb{R}^{p-1}$ and $k_{22} \in \mathbb{R}$.

Since L is not n-estimable, there must be an $A \in L$ with $W(A) = 0$, which implies that $A_{ij} = 0$ unless $i = j = p$. We may thus assume that

$$A = \begin{pmatrix} 0_n & 0 \\ 0 & 1 \end{pmatrix}. \tag{11.69}$$

In the original basis, we have $A = e^i e^{i\top}$. Since $e^1, \ldots, e^p$ were arbitrary, we have shown that for any vector u of length one, $uu^\top \in L$. Since L is a linear subspace we conclude that any matrix K of the form

$$K = \sum_{i=1}^{p} \lambda_i e^i e^{i\top} \tag{11.70}$$

is in L, and hence $\mathcal{S}^p \subseteq L$. This implies $d = T_p$, which is a contradiction. We conclude that L is n-estimable. □

Notice that even when L is n-estimable, the estimated concentration matrix may not be positive definite if L is not a Jordan subalgebra. All we can say is that the estimate will be positive definite for sufficiently large n by the consistency result in proposition 11.1.

If $\check{K}$ is not positive definite and the estimate of K itself is of interest, it may be necessary to calculate the MLE $\widehat{K}$: the latter exists and is positive definite whenever $\check{K}$ exists. In any case, lack of positive definiteness of $\check{K}$ indicates that the estimate may not be reliable and that there could be too few observations to justify the use of a model of such complexity.

§11.4 Gaussian graphical models with symmetries

Gaussian graphical models with symmetries (Højsgaard and Lauritzen, 2008) are linear concentration models generated by a coloured graph. More precisely, we let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a *coloured graph* where $\mathcal{V}$ is a partition of a finite vertex set V

into *vertex colour classes* and $\mathscr{E}$ a partition of an edge set E into *edge colour classes*. Such a graph determines a linear concentration model with $L = \mathscr{S}(\mathscr{V}, \mathscr{E})$ being the set of symmetric $p \times p$ matrices K with entries $k_{\alpha\beta} = 0$ whenever α and β are not neighbours in $\mathscr{G}$, any two off-diagonal elements being identical if the corresponding edges are in the same colour class, and any two diagonal elements identical if the corresponding vertices are in the same colour class. These constraints on K allow for parsimonious models where the number of parameters can be much smaller than the number of vertices or edges in the coloured graph. Such parsimonious models are particularly useful when the dimension of the observations d is large but the number of observations n is relatively small.

Let e^u for $u \in \mathscr{V}$ denote the $|V| \times |V|$ diagonal matrix with $e^u_{\alpha\alpha} = 1$ if $\alpha \in u$ and 0 otherwise. Similarly, for each edge colour class $u \in \mathscr{E}$ we let e^u be the $|V| \times |V|$ symmetric matrix with $e^u_{\alpha\beta} = 1$ if $\langle \alpha, \beta \rangle \in u$ and 0 otherwise. Then $\{e^u, u \in \mathscr{V} \cup \mathscr{E}\}$ form an orthogonal basis for L. The likelihood equations (Højsgaard and Lauritzen, 2008) become

$$\operatorname{tr}(e^u W) = \operatorname{tr}(e^u K^{-1}), \quad u \in \mathscr{V} \cup \mathscr{E}, \tag{11.71}$$

which are non-linear in K and must be solved by iterative methods in most cases.

One motivation for introducing these models was the potential reduction in the number of parameters of the corresponding uncoloured graphical model. This increases the stability of estimates and allows estimators to exist for a smaller number of observations. The last issue was considered in detail by Uhler (2012) for specific examples. As an aside we note that the models determined by the coloured graphs 11, 14 and 17 in (Uhler, 2012, Table 2) are supermodels of the Jordan linear concentration model (11.58) and hence we can confirm Uhler's conjecture that in these cases, the MLE does not exist for $n = 1$.

We note that our proposition 11.6 implies that the SME does not exist if $|\mathscr{V}|+|\mathscr{E}| > n(2|V| - n + 1)/2$ and believe that the condition $|\mathscr{V}| + |\mathscr{E}| \leq n(2|V| - n + 1)/2$ is sufficient to ensure n-estimability for this particular class of linear concentration

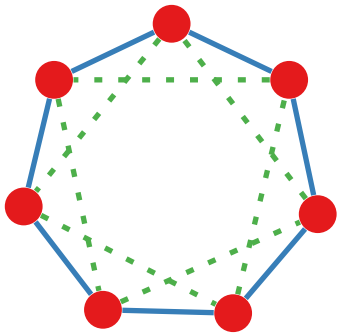


Figure 11.2: The circular autoregressive process of order 2 with $n = 7$ as a coloured Gaussian graphical model.

models, but have not been able to show this except for the case of $n = p - 1$, cf. proposition 11.12. However, none of the examples in (Jensen, 1988) or (Uhler, 2012) provide counterexamples to this conjecture. Note that the MLE may well exist even if the SME does not exist; see for example the earlier discussion of the four-cycle. If our conjecture is correct, proposition 11.5 implies that $|\mathcal{V}| + |\mathcal{E}| \leq n(2|V| - n + 1)/2$ is also sufficient for the existence of the MLE and hence provides a simple method of checking for this.

The linear score matching equations (11.16) for graphical Gaussian models with symmetries are

$$\operatorname{tr}(e^u WK) = \operatorname{tr}(e^u), \quad u \in \mathcal{V} \cup \mathcal{E}, \tag{11.72}$$

which should be compared to (11.71); they have a strong similarity with the Yule–Walker equations for estimating the parameters of autoregressive processes in a time series, as also noted by Almeida and Gidas (1993).

Indeed, a circular autoregressive process of order q is an example of a coloured graphical model with symmetry determined by the cyclic permutation group, as displayed in figure 11.2. In this case the Yule–Walker equations are exactly equivalent to the score matching equations.

§11.4.1 Model selection

Model selection in Gaussian graphical models with symmetry is problematic as the number of possible models is enormous. This affects both stepwise methods, as used in

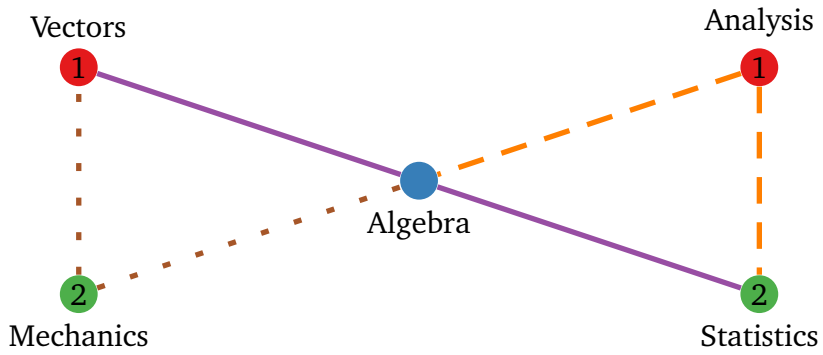


Figure 11.3: A Gaussian graphical model with symmetries for the **mathmarks** dataset.

(Højsgaard and Lauritzen, 2007), and lattice based methods (Edwards and Havránek, 1987), as used in (Gehrmann, 2011). The computational efficiency of the SME allows rapid screening of a large number of potential models as the minimized objective function indicating the model fit $J_2(\check{\theta}) = -n\,\mathrm{tr}(\check{K})/2$ is particularly simple to calculate. Note that this minimum can be calculated even though $\check{K}$ may not be positive definite; in particular a time consuming check of positive definiteness can then be avoided. In this case the minimum may overestimate the model fit as it corresponds to the minimum of J_2 over the entire space L rather than over $\Theta \subseteq L$.

To prevent overfitting a penalty for the number of parameters $d = \dim L$ should be added to J_2 to give the objective function

$$J_\lambda(L) = J_2(\check{K})/n + \lambda d = -\frac{1}{2}\,\mathrm{tr}(\check{K}) + \lambda d. \tag{11.73}$$

The scalar multiple λ for the penalty can, for example, be determined by a method such as cross-validation. Such rapid model screening may be useful to identify a small set of plausible models to be considered by more sophisticated search procedures.

§11.4.2 Examples

We briefly describe some numerical experiments with the SME for Gaussian graphical models with and without symmetries. These indicate that the SME provides an extremely fast estimate which is reasonably accurate for large n.

First we consider the **mathmarks** dataset (Mardia et al., 1979) included in **gRc** (Højsgaard and Lauritzen, 2007). Following the analysis in (Højsgaard and Lauritzen,

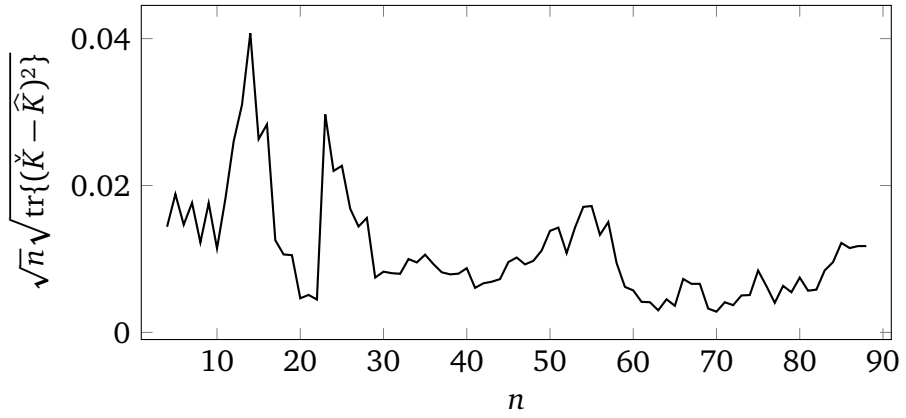


Figure 11.4: The scaled Frobenius distance between the SME and the MLE for the **mathmarks** database as n increases.

2008), we use three vertex colour classes and three edge colour classes as shown in figure 11.3. We vary the number of observations n from 4 to 88 and compute both the SME and the MLE for each n. figure 11.4 shows how the SME approximates the MLE as n grows. The SME appears to provide a computationally efficient estimator with good accuracy for large n.

Next we shall give an example of the SME identifying a non-decomposable graphical Gaussian model. We first simulated data from a square lattice model with $p = s^2$ vertices: the concentration matrix for our model is symmetric with upper-triangular entries

$$K_{ij} = \begin{cases} 1 & \text{if } j = i, \\ 0.2 & \text{if } j = i + s, \\ 0.2 & \text{if } j = i + 1 \text{ and } i \not\equiv 0 \bmod s, \\ 0 & \text{otherwise,} \end{cases} \tag{11.74}$$

for $1 \leq i \leq j \leq s^2$. The previously mentioned four-cycle (figure 11.1) is such a model with $s = 2$. We then used the SME to conduct a rapid model search over uncoloured graphs G on p vertices.

We now describe our model search method. We began by initializing the graph to the best-fitting tree via Kruskal's algorithm (Kruskal, 1956) using squared correlations as weights. This initialization is very fast, and if we were only searching over trees, it corresponds to the maximum likelihood estimate (Chow and Liu, 1968; Edwards et al., 2010). In our case we are searching among all undirected graphs, but the tree

step provides a computationally efficient starting position for the search. Next we conducted a greedy "forward search" and successively add edges to G. To ensure our method is scalable for large p, we considered adding edges in order of decreasing squared correlations and we terminated the forward search after attempting to add an edge that fails to improve the objective function. Finally, we conducted a greedy "backward search" by successively removing existing edges from G.

The algorithm is described in more detail below. In the following we use $G + \langle i, j \rangle$ to denote the graph resulting from adding the undirected edge $\langle i, j \rangle$ to G, $G \setminus \langle i, j \rangle$ to denote the graph after removing the edge $\langle i, j \rangle$, $L(G)$ to denote the subspace $\{K \in \mathcal{S}^p : K_{ij} = 0 \text{ whenever edge } \langle i, j \rangle \notin G\}$, and $J_\lambda(L) = \text{argmin}_{K \in L} J_\lambda(K)$.

Algorithm 11.1 Rudimentary model search algorithm utilizing the SME.

1: Construct the $p \times p$ sample covariance matrix W for the observed data.
2: Construct a complete graph G_C on p vertices with edge weights $W_{ij}^2 / \sqrt{W_{ii} W_{jj}}$.
3: Initialization of G: let G be the maximum spanning tree for G_C.
4: Forward search: create a list A of the edges in G_C, ordered by decreasing edge weight. Let a be the first edge in A.
5: If $a \in G$ then let a be the next edge in A and go to line 5.
6: If $J_\lambda[L\{G + a\}] < J_\lambda\{L(G)\}$, then let $G \leftarrow G + a$ and go to line 5.
7: Backward search: create a list B of the edges in G ordered by decreasing $J_\lambda(G) - J_\lambda(G \setminus \langle i, j \rangle)$. Let b be the first edge in B.
8: If $J_\lambda[L\{G \setminus b\}] < J_\lambda\{L(G)\}$ then let $G \leftarrow G \setminus b$ and b be the next edge in B. Go to line 8.
9: Repeat: If the number of repetitions so far is fewer than 3, then go to line 4.

Before running this algorithm we first identified a suitable penalty λ. We considered the change in the objective function J_2 after adding the single extra edge which most improved the objective to the true model. By simulating several thousand samples over the grid with $s = 2, \ldots, 8$ and $n = s^2, 2s^2, 3s^2, \ldots, 10s^2$, we found that the expected change in J_2 was approximately proportional to $\sqrt{p} \log\log(np)$. We chose to proceed with $\lambda = \sqrt{p} \log\log(np)/(2n)$.

We quantified the accuracy of the fitted model $\check{K}$ by considering the number of missing edges ($\check{K}_{ij} = 0$ but the true $K_{ij} \neq 0$) and extra edges ($\check{K}_{ij} \neq 0$ but the true $K_{ij} = 0$) in figure 11.5.

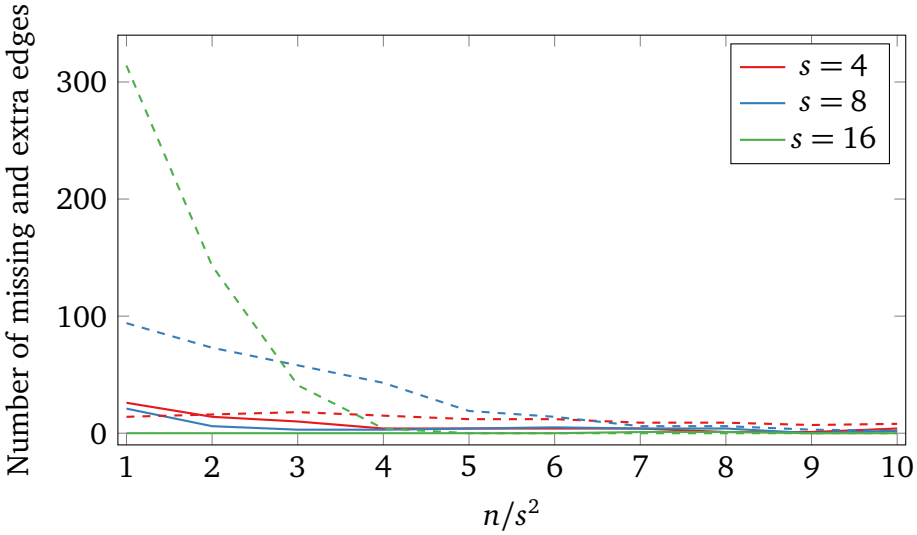
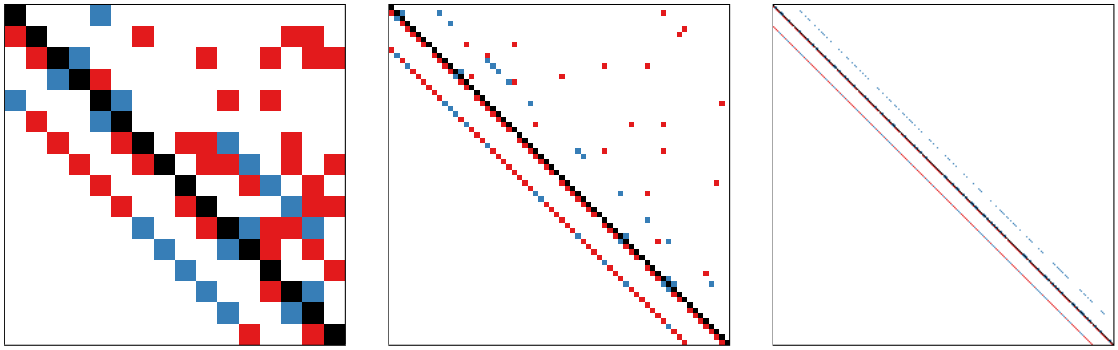


Figure 11.5: The number of missing (dashed line) and extra (solid line) edges in the SME for the lattice with $s = 4, 8, 16$ and n from $n = s^2$ to $n = 10s^2$.

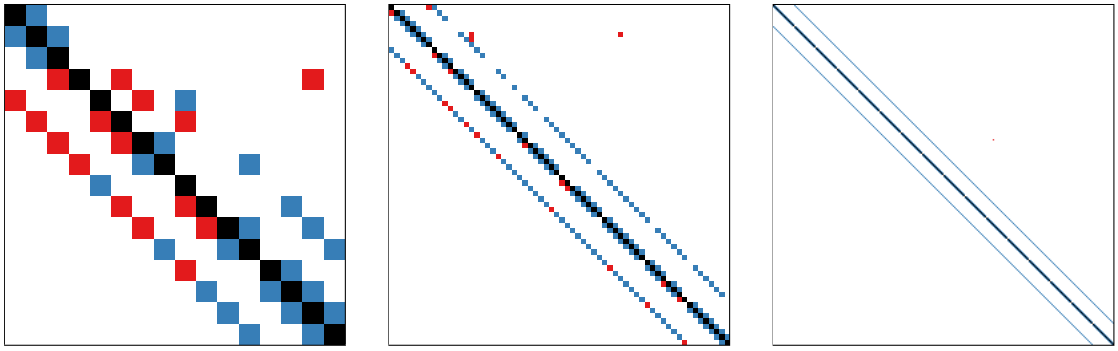
Finally, we visually inspect how the score matching estimate becomes more accurate with increasing n in figure 11.6. We see that the estimate is unstable at $n = p$, but for $n = 10p$ the SME correctly identifies the majority of the true non-zero entries with few extra edges. Note that for $p = 256$ and $n = 10p$ the model is correctly identified.

In the case of searching for uncoloured graphical models as above, the SME may be seen as alternative to the graphical lasso algorithm (Friedman et al., 2008). It is difficult to directly compare the accuracy of the SME to the graphical lasso due to the unspecified regularization parameter in the latter algorithm. The graphical lasso estimate is extremely sensitive to the precise value of this regularization parameter: by fine-tuning the parameter for each n and p we were able to achieve results equal to or better than those of the SME, however the range of values which yields accurate estimates is narrow and highly dependent on n and p. By contrast the SME seems relatively robust against small changes in λ.

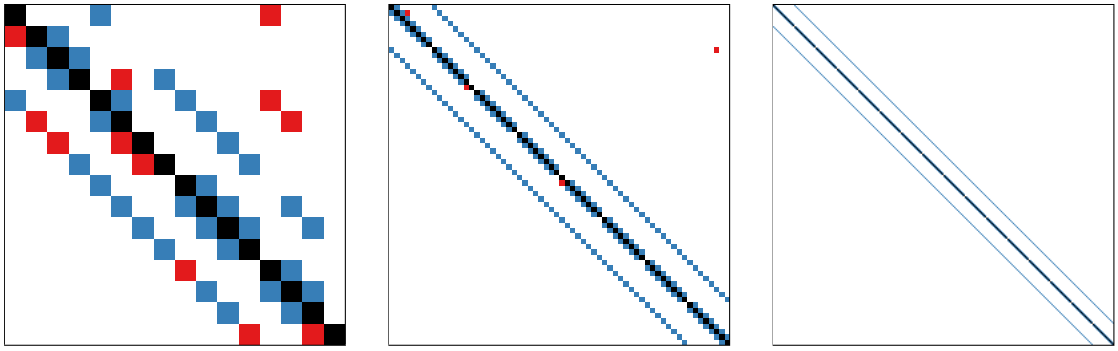
We also suspect that the SME scales better for large p than the graphical lasso. Using an implementation of the SME written in *C#* we were able to consider models up to $s = 100$ ($p = 10^4$) before running out of computer memory: even at this large p each estimate of the SME could be completed within ten seconds, and with $n = 10p$ our rudimentary search procedure correctly identified the model. We attempted to



(a) $p = 16, n = 1 \times p$ (b) $p = 64, n = 1 \times p$ (c) $p = 256, n = 1 \times p$



(d) $p = 16, n = 5 \times p$ (e) $p = 64, n = 5 \times p$ (f) $p = 256, n = 5 \times p$



(g) $p = 16, n = 10 \times p$ (h) $p = 64, n = 10 \times p$ (i) $p = 256, n = 10 \times p$

Figure 11.6: Sparsity patterns of the estimated concentration matrices. Above the diagonal is the SME $\check{K}$; below the diagonal is the true concentration matrix K. Entries that are present in both K and $\check{K}$ are blue and entries which are present in only one matrix are red. Thus red entries below the diagonal correspond to true edges that were missed by the SME and red entries above the diagonal correspond to extra edges in the SME.

test the graphical lasso at such p using the R package **glasso** (Friedman et al., 2011), however the R environment ran out of memory while attempting to load the sample covariance matrix.

Finally, we should emphasize that, in contrast to the graphical lasso, the SME can be used for graphical model with symmetries. Thus the SME may be useful for model screening over such models, though this would require the development of computationally efficient model search strategies. It is outside the scope of the present article to study such strategies in any detail.

§11.5 Discussion

Score matching is an efficient method of parameter estimation for distributions with intractable normalization constants. It is particularly suitable to parameter estimation within an exponential family, where the score estimating equations are linear and yield a consistent estimate. The ready availability of highly optimized algorithms for linear equations means that the SME can be computed quickly and with a small memory footprint, even when the number of parameters is very large.

The method seems particularly promising for rapid model screening and it would be well worthwhile to further investigate the optimal form of the penalty for model complexity, i.e. the coefficient λ in (11.73), in particular how it should depend on n to ensure consistent identification of the model along the lines in Hannan and Quinn (1979), see also the discussion in van Erven et al. (2012). It would be an advantage to have a simple sufficient condition for the existence of the SME — and hence also the MLE — in Gaussian graphical models (with or without symmetries) so that a model search could be automatically restricted to models of sufficiently limited complexity. We believe the condition in proposition 11.6 is necessary and sufficient for the SME to exist in this case and thus sufficient for existence of the MLE. Unfortunately we only been able to show this in general for $n \geq p - 1$. We hope to return to these and other questions in the future.

Part III

Appendices

Appendix A

Sampling from the posterior of a von Mises distribution

§A.1 Introduction

There has been renewed interest in directional Bayesian analysis in view of its fundamental applications to molecular biology (Boomsma et al., 2008; Frellsen et al., 2009; Mardia, 2013). Due to chemical constraints on the bonds of biomolecules, the geometry of these molecules can be described by a set of angles. Other applications include locating and tracking an electric signal (Guttorp and Lockhart, 1988) and the analysis of forensic fingerprint evidence (Forbes and Lauritzen, 2013). All these applications involve circular data which is naturally modelled by the von Mises distribution.

The probability density function of the von Mises distribution with mean $\nu \in \mathbb{S}^1$ and concentration parameter $\kappa \geq 0$ is (Mardia and Jupp, 1999)

$$p(\psi) = \frac{1}{I_0(\kappa)} \exp\{\kappa \cos(\psi - \nu)\} \tag{A.1}$$

with respect to the uniform probability measure $\mu_{\mathbb{S}^1}$, where $I_m(\cdot)$ is the modified Bessel function of the first kind and order m. The circular variance can be described by $1 - r(\kappa)$ where $r(\kappa) = I_1(\kappa)/I_0(\kappa)$. Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ be a vector of observations from a von Mises distribution. When a conjugate prior is used, the posterior distribu-

This chapter is a slight modification of Forbes and Mardia (2014).

tion of the mean $\nu|\boldsymbol{\psi},\kappa$ is itself von Mises, and can be easily sampled via Best and Fisher (1979).

Let $\pi(\kappa) \propto I_0(\kappa)^{-a}\exp(-b\kappa)$ be the conjugate prior for the concentration. The posterior is

$$p(\kappa) = \frac{A}{I_0(\kappa)^{\eta}}\exp(-\eta\beta_0\kappa)\ \mathbb{1}(\kappa \geq 0) \tag{A.2}$$

with respect to the measure $\mu_{\mathbb{R}}$, where $\eta > 0$ and $\beta_0 \in (-1, \infty)$ are observed constants; in this case $\eta = a + n$ and $\beta_0 = b/(a+n) - n^{-1}\sum_{i=1}^{n}\cos(\psi_i - \nu)$. For the case $\eta = 1$ and $\beta_0 > 1$ the normalization constant is $A = \sqrt{\beta_0^2 - 1}$, however in general the normalization constant is intractable (Mardia, 2007). In this paper we shall call (A.2) the *Bessel exponential distribution*.

Existing algorithms to sample (A.2) tend to generate from approximate distributions (Guttorp and Lockhart, 1988) or have a large overhead of sampled auxiliary variables (Damien and Walker, 1999). We present a new, extremely fast algorithm to sample from the Bessel exponential distribution.

For large κ, $I_0(\kappa)$ is approximately $\exp(\kappa)/\sqrt{2\pi\kappa}$ (Olver et al., 2010, §10.30). Plugging this approximation into (A.2) yields a gamma density with shape $\eta/2 + 1$ and rate $\eta(\beta_0 - 1)$.

This insight motivates us to use a gamma-based acceptance-rejection sampler for κ. However, the above approximation for $I_0(\kappa)$ breaks down for small κ, and thus great care is needed to ensure our rejection sampler is efficient for all values of κ. We derive the optimal gamma-based proposal distribution and show that the resulting sampler has an acceptance probability of at least 0.7 for all η and β_0. The minimum acceptance probability of ≈ 0.7 occurs when the distribution is concentrated around $\kappa = 0$.

The algorithm is described in § A.2 and derived in § A.3. Enhancements are considered in §A.4 and the algorithm's efficiency is explored in §A.5.

§A.2 The algorithm

As discussed above, we can approximate the Bessel exponential distribution with a gamma distribution. However, because the ratio of these densities diverges as $\kappa \to 0$, we cannot directly use a gamma proposal for our rejection sampler. Instead we propose values $\kappa = x - \varepsilon$ where $\varepsilon > 0$ and x has a gamma distribution. This is an application of Marsaglia's exact approximation procedure, see Marsaglia (1984) for more details.

Using a shifted gamma proposal with shape $\eta\alpha + 1 \geq 1$ and scale $\eta\beta > 0$ leads to the envelope function

$$q(\kappa; \alpha, \beta, \varepsilon) = M(\alpha, \beta, \varepsilon)(\kappa + \varepsilon)^{\eta\alpha} \exp(-\eta\beta\kappa); \quad \kappa \geq 0 \tag{A.3}$$

for $p(\kappa)$, where the amplitude $M(\alpha, \beta, \varepsilon)$ is chosen to ensure the ratio p/q is bounded below one. We can generate a sample from the Bessel exponential distribution by generating a sample κ from q and accepting it with probability

$$\frac{p(\kappa)}{q(\kappa; \alpha, \beta, \varepsilon)} = \frac{A}{M(\alpha, \beta, \varepsilon)} \exp\{\eta g(\kappa; \alpha, \beta, \varepsilon)\}, \tag{A.4}$$

where

$$\begin{aligned} g(\kappa; \alpha, \beta, \varepsilon) &= (\beta - \beta_0)\kappa - \alpha \log(\kappa + \varepsilon) - \log I_0(\kappa), \\ M(\alpha, \beta, \varepsilon) &= A \exp\{\eta g(\kappa_0; \alpha, \beta, \varepsilon)\}, \end{aligned} \tag{A.5}$$

and $\kappa_0 = \operatorname{argmax}_{\kappa \geq 0} g(\kappa; \alpha, \beta, \varepsilon)$. This procedure is valid for any choice of the proposal parameters $\alpha, \beta, \varepsilon$, though the values of these parameters will affect the algorithm's efficiency.

In §A.3 we show that the approximate optimal choices for the proposal parameters are

$$\begin{aligned} \beta &= \begin{cases} \beta_0 + 1 & \text{if } \beta_0 \leq 1/(4\eta) - 2/\left(3\sqrt{\eta}\right), \\ \beta_0 + r(\kappa_0) + \dfrac{1 - r(\kappa_0)}{1 + 40\eta\{\beta_0 - 1/(4\eta) + 2/\left(3\sqrt{\eta}\right)\}^2} & \text{otherwise,} \end{cases} \\ \varepsilon &= \frac{\kappa_0 \mathcal{W}_0\{c_3 \exp(c_3)\}}{c_3 - \mathcal{W}_0\{c_3 \exp(c_3)\}}, \quad \alpha = \{\beta - \beta_0 - r(\kappa_0)\}(\kappa_0 + \varepsilon), \end{aligned} \tag{A.6}$$

where the terms κ_0 and c_3 are

$$\begin{gathered}\kappa_0 = \frac{1-1/\eta+1/(2\eta^2)}{\eta\beta_0+\sqrt{2\eta+\eta^2\beta_0^2}} + \frac{1+3/(2\eta)-1/(4\eta^3)}{(\eta+1)\beta_0+\sqrt{2\eta+1+\eta^2\beta_0^2}}, \\ c_3 = -\frac{\beta-\beta_0-\log\{I_0(\kappa_0)\}/\kappa_0}{\beta-\beta_0-r(\kappa_0)},\end{gathered} \tag{A.7}$$

and $\mathscr{W}_0(\cdot)$ is the principal branch of the Lambert W function defined as $t = \mathscr{W}_0(t)\exp\{\mathscr{W}_0(t)\}$ for $\mathscr{W}_0(t) > -1$ (Olver et al., 2010, §4.13).

The acceptance-rejection algorithm to generate a sample from the Bessel exponential distribution proceeds as follows:

1. Find efficient proposal parameters $\alpha, \beta, \varepsilon$, which will depend η and β_0.
2. Draw x from a gamma distribution with shape $\eta\alpha+1$ and rate $\eta\beta$.
3. Draw u from a Uniform distribution on $[0,1]$.
4. Accept $\kappa = x-\varepsilon$ if $\log u < \eta g(\kappa;\alpha,\beta,\varepsilon) - \eta g(\kappa_0;\alpha,\beta,\varepsilon)$, else go to 2.

The detailed procedure is described in algorithm A.1. When implementing the algorithm, both the Bessel functions and the Lambert W function can be computed using software such as the General Scientific Library (Galassi et al., 2009) or its R wrapper, the CRAN package **gsl**. In practice it is often possible to avoid computing these functions, as we show in §A.4.

§A.3 Derivation of the algorithm

We will now derive the optimal parameters $\alpha, \beta, \varepsilon$ for the proposal distribution of $x = \kappa - \varepsilon$ which follows a gamma distribution with shape $\eta\alpha+1$ and rate $\eta\beta$. We do so by maximizing the expected probability of acceptance,

$$\begin{aligned}&E_{x\,|\,\alpha,\beta}\big[\,\mathbb{1}(x \geq \varepsilon)\exp\{\eta g(x-\varepsilon;\alpha,\beta,\varepsilon) - \eta g(\kappa_0;\alpha,\beta,\varepsilon)\}\big] \\ &\qquad = \frac{(\eta\beta)^{\eta\alpha+1}}{\Gamma(\eta\alpha+1)}\exp\{-\eta\beta\varepsilon - \eta g(\kappa_0;\alpha,\beta,\varepsilon)\}\int_0^\infty \frac{\exp(-\eta\beta_0\kappa)}{I_0(\kappa)^\eta}\,\mathrm{d}\kappa,\end{aligned} \tag{A.8}$$

over $\alpha, \beta, \varepsilon, \kappa_0$ subject to the constraint $\kappa_0 = \operatorname{argmax}_{\kappa\geq 0} g(\kappa;\alpha,\beta,\varepsilon)$. In order for the maximum to be finite as $\kappa \to \infty$ we require $\beta \leq \beta_0 + 1$.

Algorithm A.1 Rejection sampler for the Bessel exponential distribution

Initialization: find parameters for proposal distribution
$\kappa_L \leftarrow 2/\left(\eta\beta_0 + \sqrt{2\eta + \eta^2\beta_0^2}\right)$
$\kappa_U \leftarrow (2+1/\eta)/\left\{(\eta+1)\beta_0 + \sqrt{2\eta+1+\eta^2\beta_0^2}\right\}$
$c_1 = 1/2 + \{1 - 1/(2\eta)\}/2\eta$
$\kappa_0 \leftarrow (1-c_1)\kappa_L + c_1\kappa_U$
$i_0 \leftarrow I_0(\kappa_0)$
$r \leftarrow I_1(\kappa_0)/i_0$
$c_2 \leftarrow 1/(4\eta) - 2/\left(3\sqrt{\eta}\right)$
if $\beta_0 \le c_2$ **then**
    $\beta \leftarrow \beta_0 + 1$
else
    $\beta \leftarrow \beta_0 + r + (1-r)/\{1 + 40\eta(\beta_0 - c_2)^2\}$
end if
$c_3 \leftarrow \{\log(i_0)/\kappa_0 - \beta + \beta_0\}/(\beta - \beta_0 - r)$
$c_4 \leftarrow \mathscr{W}_0\{c_3\exp(c_3)\}$ # Lambert's W function, see §A.4
$\varepsilon \leftarrow c_4\kappa_0/(c_3 - c_4)$
$\alpha \leftarrow (\beta - \beta_0 - r)(\kappa_0 + \varepsilon)$
# Perform rejection sampling
repeat
    $x \leftarrow$ sample from a Gamma$(\eta\alpha + 1, \eta\beta)$ left-truncated at ε
    $\kappa \leftarrow x - \varepsilon$
    $u \leftarrow$ sample from a Uniform$(0,1)$
until $\log(u)/\eta < (\beta - \beta_0)(\kappa - \kappa_0) - \alpha\log\{(\kappa+\varepsilon)/(\kappa_0+\varepsilon)\} - \log\{I_0(\kappa)/i_0\}$
return κ

By taking logs we see that maximizing (A.8) with respect to κ is equivalent to maximizing

$$h(\kappa_0; \alpha, \beta, \varepsilon) = (\alpha + \eta^{-1})\log(\eta\beta) - \eta^{-1}\log\Gamma(\eta\alpha + 1) - \beta\varepsilon - g(\kappa_0; \alpha, \beta, \varepsilon). \quad \text{(A.9)}$$

The constraint $\kappa_0 = \operatorname{argmax}_{\kappa \ge 0} g(\kappa; \alpha, \beta, \varepsilon)$ implies either $\kappa_0 = 0$ or $\frac{\mathrm{d}}{\mathrm{d}\kappa} g(\kappa = \kappa_0; \alpha, \beta, \varepsilon) = 0$. The Lagrangians for constrained optimization corresponding to these conditions are $h + \lambda\kappa_0$ and $h + \lambda\frac{\mathrm{d}}{\mathrm{d}\kappa}g$, neither of which have interior critical points over $(\alpha, \beta, \varepsilon, \lambda)$ because the α and ε derivatives have no common root. Thus the optimal parameters must lie on the boundary of the parameter space. An examination of the boundaries show that the maximum satisfies $\frac{\mathrm{d}}{\mathrm{d}\kappa} g(\kappa = \kappa_0; \alpha, \beta, \varepsilon) = 0$ and $g(\kappa_0; \alpha, \beta, \varepsilon) = g(0; \alpha, \beta, \varepsilon)$. Intuitively, this says that for any κ_0 we should pick ε as small as possible while still having κ_0 be the maximizer of g. Thus our Lagrangian

is

$$L(\kappa_0;\alpha,\beta,\varepsilon,\lambda_1,\lambda_2) = h(\kappa_0;\alpha,\beta,\varepsilon) + \lambda_1 \frac{\mathrm{d}g(\kappa=\kappa_0;\alpha,\beta,\varepsilon)}{\mathrm{d}\kappa} + \lambda_2\{g(\kappa_0;\alpha,\beta,\varepsilon) - g(0;\alpha,\beta,\varepsilon)\}. \quad \text{(A.10)}$$

Our optimal parameters are either a critical point of L or lie on one or more of the boundaries $\alpha = 0, \varepsilon = 0$ or $\beta = \beta_0 + 1$. If either $\alpha = 0$ or $\varepsilon = 0$ then direct differentiation shows the maximum occurs when $\alpha = 0, \varepsilon = 0, \beta = \beta_0 + r(\kappa)$, and κ_0 is the unique positive root of $\beta - 1/(\eta\kappa)$. We shall see this is a limiting case of the critical point solution. The only other boundary is $\beta = \beta_0 + 1$; we shall see this is the solution when β_0 is close to -1.

To find the critical points of L, we start by setting the derivatives with respect to $\alpha, \varepsilon, \lambda_1$, and λ_2 to zero and rearranging yields the optimal parameters as functions of κ_0 and β. This yields

$$\begin{aligned}
\alpha &= \{\beta - \beta_0 - r(\kappa_0)\}(\kappa_0 + \varepsilon), \\
\varepsilon &= \frac{\kappa_0 \mathscr{W}_0\{c_3 \exp(c_3)\}}{c_3 - \mathscr{W}_0\{c_3 \exp(c_3)\}}, \\
c_3 &= -\frac{\beta - \beta_0 - \log\{I_0(\kappa_0)\}/\kappa_0}{\beta - \beta_0 - r(\kappa_0)}, \\
\lambda_2 &= \frac{\Psi(\eta\alpha + 1) - \log\{\eta\beta(\kappa_0 + \varepsilon)\} - 1 + \beta(\kappa_0 + \varepsilon)/\alpha}{\log(1 + \kappa_0/\varepsilon) - \kappa_0/\varepsilon}, \\
\lambda_1 &= \frac{\kappa_0 + \varepsilon}{\alpha\varepsilon}\{\beta\varepsilon^2 + (\kappa_0\beta - \alpha)\varepsilon + \alpha\lambda_2\kappa_0\},
\end{aligned} \quad \text{(A.11)}$$

where $\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$ is the digamma function (Olver et al., 2010, §5.2). We must have $\beta > \beta_0 + r(\kappa_0)$ so that α and ε are positive. Notice that the limit $\beta \to \beta_0 + r(\kappa_0)$ corresponds to the boundary case $\alpha = \varepsilon = 0$ discussed above.

The above equations give all optimal parameters in terms of β and κ_0. Note that since constraint $\frac{\mathrm{d}}{\mathrm{d}\kappa}g(\kappa;\alpha,\beta,\varepsilon) = 0$ at $\kappa = \kappa_0$ is satisfied whenever $\alpha = \{\beta - \beta_0 - r(\kappa_0)\}(\kappa_0 + \varepsilon)$, we are free to choose alternative, sub-optimal values for the other parameters if the true optimal values are too difficult to compute. We shall explore this in §A.4 when we use an approximation for the Lambert W function.

Finally, we find the optimal β as follows. This value must either lie on the boundary $\beta = \beta_0 + 1$ or else satisfy

$$\frac{\partial L(\kappa_0; \alpha, \beta, \varepsilon, \lambda_1, \lambda_2)}{\partial \beta} = (\alpha + 1/\eta)/\beta - (\kappa_0 + \varepsilon) + \lambda_1 - \lambda_2 \kappa_0 = 0. \tag{A.12}$$

Unfortunately plugging (A.11) into (A.12) and solving for β as a function of κ_0 alone is analytically intractable. However, one can check that $\partial L/\partial \beta$ decreases from positive infinity at $\beta = \beta_0 + r(\kappa_0)$ to negative infinity as $\beta \to \infty$. Since all admissible β lie in the finite interval

$$\max\{0, \beta_0 + r(\kappa_0)\} < \beta < \beta_0 + 1, \tag{A.13}$$

we can easily find the optimal β through any standard one-dimensional root-finding algorithm. If the root lies to the right of $\beta_0 + 1$, the optimal value is $\beta = \beta_0 + 1$.

We plug the optimal β into (A.11) to find all of our optimal parameters in terms of κ_0. Doing this for each κ_0 and plugging the resulting parameters $(\alpha, \beta, \varepsilon)$ into $h(\kappa_0; \alpha, \beta, \varepsilon)$ yields a function which we numerically maximize over κ_0. Let κ^* be the optimal value of κ_0 and let $(\alpha^*, \beta^*, \varepsilon^*)$ be the optimal parameters corresponding to κ^*. These are the desired parameters that maximize the expected acceptance probability (A.8).

The above numeric maximizations for β and κ_0 may be acceptable when η and β_0 are known a priori. However, they are computationally prohibitive in the standard Monte Carlo case where we wish to generate many samples from the Bessel exponential distribution with different values of η and β_0 for each sample. Thus our next task is to approximate the optimal parameters with easily computable functions of η and β_0.

For all η and β_0, κ^* is well approximated by κ_a, the positive root of $\beta_0 + r(\kappa) - 1/(\eta\kappa)$; indeed κ_a is the exact optimum in the boundary case $\alpha = \varepsilon = 0$. To approximate κ_a we use the bounds (Amos, 1974, eq. 11),

$$\frac{\kappa}{1 + \sqrt{1 + \kappa^2}} \leq r(\kappa) \leq \frac{\kappa}{\sqrt{4 + \kappa^2}}. \tag{A.14}$$

Rearranging these bounds shows that $\kappa_L \leq \kappa_a \leq \kappa_U$ where

$$\kappa_L = 2\left(\eta\beta_0 + \sqrt{2\eta + \eta^2\beta_0^2}\right)^{-1}, \quad \kappa_U = (2 + 1/\eta)\left\{(\eta+1)\beta_0 + \sqrt{2\eta + 1 + \eta^2\beta_0^2}\right\}^{-1}. \tag{A.15}$$

These bounds are relatively tight, we found that the convex combination $(1-c_1)\kappa_L + c_1\kappa_U$ with $c_1 = 1/2 + \{1 - 1/(2\eta)\}/2\eta$ provides a good approximation to κ_a and hence to κ^*.

The parameter β^* is exactly equal to $\beta_0 + 1$ when β_0 is close to its lower bound of negative one. For β_0 sufficiently large, β^* drops from its upper limit $\beta_0 + 1$ towards its lower limit $\beta_0 + r(\kappa^*)$. The transition between the two limits is very rapid for $\eta > 10$. We achieve good accuracy with the approximation

$$\beta^* \approx \begin{cases} \beta_0 + 1 & \text{if } \beta_0 < c_2, \\ \beta_0 + r(\kappa^*) + \dfrac{1 - r(\kappa^*)}{1 + 40\eta(\beta_0 - c_2)^2} & \text{otherwise,} \end{cases} \tag{A.16}$$

where $c_2 = 1/(4\eta) - 2/\left(3\sqrt{\eta}\right)$. This approximation is very good when η is large or when $|\beta_0| > 0.1$. A slower, more precise approximation for β^* may lead to parameters which provide better efficiency for small η and $\beta_0 \approx 0$; we address this in §A.5.

Given these approximations of κ^* and β^*, the parameters α^* and ε^* are given by (A.11).

§A.4 Further speed enhancements

The truncated gamma on line 20 can be sampled using Dagpunar's algorithm (Dagpunar, 1978). Alternatively, one can use a standard gamma sampling algorithm such as Marsaglia–Tsang (Marsaglia and Tsang, 2000a) and reject when $x < \varepsilon$. Indeed, the Marsaglia–Tsang algorithm is itself a rejection sampler with a Gaussian proposal, and its rejection step can be combined with the rejection step on line 23 for an additional speed-up.

The function $\mathscr{W}_0$ on line 15 can be approximated by (Winitzki, 2003)

$$\mathscr{W}_0(t) = \frac{et}{1 + \{(2et+2)^{-1/2} + (e-1)^{-1} - 2^{-1/2}\}^{-1}} \tag{A.17}$$

with no noticeable drop in the expected probability of acceptance.

Finally, we can implement the simple squeezes

$$\begin{aligned} I_0(\kappa) &< \{1 + 1/(2\kappa)\}\exp(\kappa)/\sqrt{2\pi\kappa} : && \kappa > 0 \\ I_0(\kappa) &> \exp(\kappa)/\sqrt{2\pi\kappa} : && \kappa > 0.259, \end{aligned} \tag{A.18}$$

to avoid computing the Bessel function within the rejection loop on line 23. Specifically, the loop on lines 19 to 24 can be replaced with algorithm A.2.

Algorithm A.2 Optimized loop replacing lines 19–24 of algorithm A.1

$c_5 = \log(i_0)$
loop
  $x \leftarrow$ sample from a Gamma$(\eta\alpha + 1, \eta\beta)$ left-truncated at ε
  $\kappa \leftarrow x - \varepsilon$
  $c_6 \leftarrow \frac{1}{2}\log(2\pi\kappa) - \kappa$
  $u \leftarrow$ sample from a Uniform$(0, 1)$
  $v \leftarrow \log(u)/\eta - (\beta - \beta_0)(\kappa - \kappa_0) + \alpha\log\{(\kappa + \varepsilon)/(\kappa_0 + \varepsilon)\} - c_5$
  if $\kappa < 0.258$ **or** $v < c_6$ **then**
    if $v < c_6 - \log\{1 + 1/(2\kappa)\}$ **or** $v < -\log\{I_0(\kappa)\}$ **then**
      return κ
    end if
  end if
end loop

§A.5 Efficiency analysis

We now analyse the efficiency of algorithm A.1. When using the Winitzki approximation, the initial setup (lines 2–17) involve arithmetic operations, four square roots, two Bessel function evaluations, one logarithm and one exponentiation. In total this setup requires approximately 70 microseconds on a 2.4GHz Intel i5 computer when using the R package **gsl**. Implementation in a lower-level language would increase the speed significantly.

Each iteration of the rejection loop requires a gamma sample, a uniform sample, between three and five logarithms, and in the worst case a Bessel function evaluation. The squeeze in algorithm A.2 does a good job of avoiding the Bessel computation most of the time, and each iteration of the loop requires approximately 10 microseconds.

Most of these iterations are accepted, and the algorithm, implemented in R, yields approximately 80,000 von Mises samples per second when $\eta = 10$ and β_0 is drawn uniformly over $(-1, 1)$. When using a compiled language such as C++, the algorithm yields over one million samples per second.

In figure A.1 we plot the expected probabilities of acceptance as functions of η and β_0. The figures were generated by numerically integrating the expected probability of acceptance (A.8) for each $\eta = 1, 5, 10, 100$ and for each of 2,000 equally spaced values of $\beta_0 \in (-1, 1)$.

There is a noticeable dip in efficiency near $\beta_0 = 0$. Recalling that

$$\beta_0 = -n^{-1} \sum_{i=1}^{n} \cos(\psi_i - \nu), \tag{A.19}$$

we see that this region corresponds to diffused ψ_i, i.e. the true κ is near zero. This is precisely the region where our Bessel function approximation fails, so this drop is to be expected. Fortunately the drop in efficiency is not severe and our efficiency remains above 0.7 for all η and β_0.

From figure A.1 we see that our algorithm with the approximate optimal parameters does noticeably worse than the numerically computed true optimal parameters when $\beta_0 \approx 0$. This corresponds to the transition region where the optimal β rapidly drops from its upper limit of $\beta_0 + 1$ to its lower limit of $\beta_0 + r(\kappa_0)$. Our approximation of the optimal β is inaccurate in this transition region. A more sophisticated approximation of the optimal β would increase the algorithm's efficiency, however, the region $\kappa \approx 0$ is not usually an area of primary interest and we prefer to use the faster approximation.

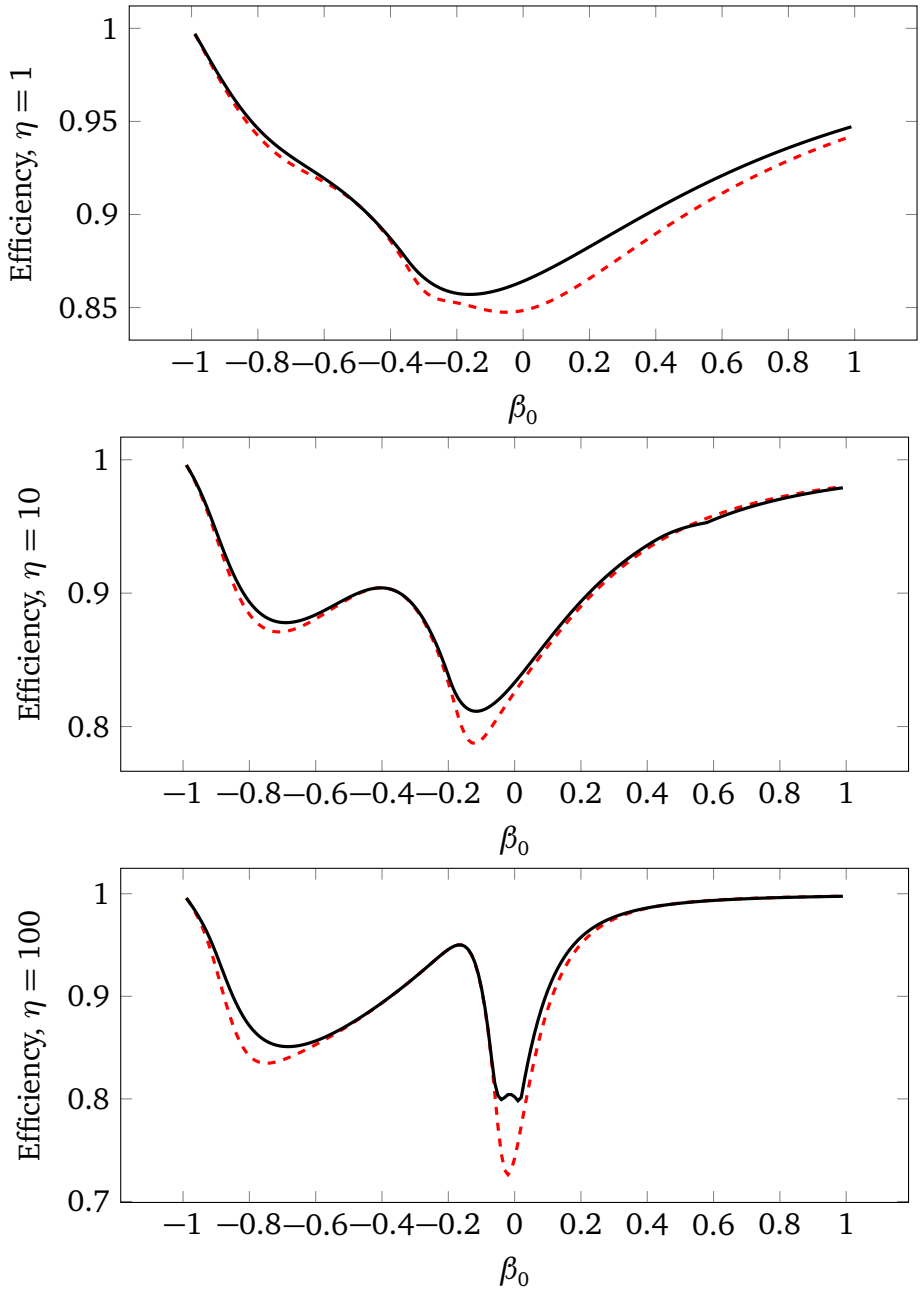


Figure A.1: Expected probability of accepting a proposed κ (A.8), for $\eta = 1, 10, 100$. Red dashed lines correspond to algorithm A.1 with the Winitzki approximation, black lines correspond to the numerically computed true optimal values for β and κ_0.

Appendix B

Miscellaneous technical results

§B.1 Derivation of the Radon-Nikodym derivative for $\tilde{\zeta}$

To find the Radon-Nikodym derivative for the change of measure from $\zeta \times \zeta \times \zeta_2$ to $\tilde{\zeta}$ in §2.2.3, we first define Ω_1 and Ω_2 be the normalized distributions

$$\begin{aligned} \mathrm{d}\Omega_1(A,B,\xi) &= \frac{\mathrm{d}\tilde{\zeta}(A,B,\xi)}{|\Xi(A,B)|}, \\ \mathrm{d}\Omega_2(A,B,\xi) &= \mathrm{d}\zeta\{A \setminus \Pi_A(\xi)\}\,\mathrm{d}\zeta\{B \setminus \Pi_B(\xi)\}\,\mathrm{d}\zeta_2[\{(a,b) : \langle a,b\rangle \in \xi\}]. \end{aligned} \tag{B.1}$$

That is, Ω_1 is the distribution induced by $\tilde{\zeta}$ and Ω_2 is the distribution induced by $\zeta \times \zeta \times \zeta_2$ after applying the bijection (2.23). Note that, conditional on n_A and n_B, the marked points $m \in A \cup B$ are i.i.d. with density $\varphi(r_m)/3$ under both Ω_1 and Ω_2:

$$\mathrm{d}\Omega_1(A,B \,|\, n_A,n_B) = \prod_{m\in A\cup B} \frac{\varphi(r_m)}{3} = \mathrm{d}\Omega_2(A,B \,|\, n_A,n_B). \tag{B.2}$$

Furthermore, the marked points are independent of the matching ξ, which implies that $\xi\,|A,B$ must be uniform over its support, which is the discrete space $\Xi(A,B)$. Recalling that $|\Xi(A,B)|$ depends on A and B only through n_A and n_B, we have

$$\Omega_1(\xi\,|A,B) = \Omega_1(\xi\,|\,n_A,n_B) = \frac{1}{|\Xi(A,B)|} = \Omega_2(\xi\,|\,n_A,n_B) = \Omega_2(\xi\,|A,B). \tag{B.3}$$

Now, since $|A| = n_A$ and $|B| = n_B$ have unit Poisson distributions under Ω_1, we have $\Omega_1(n_A,n_B) = \exp(-2)/(n_A!n_B!)$. Furthermore, since $|A \setminus \Pi_A(\xi)| = n_A - n_\xi$, $|B \setminus \Pi_A(\xi)| = n_B - n_\xi$, and $|\xi| = n_\xi$ have unit Poisson distributions under Ω_2, we

have

$$\Omega_2(n_A - n_\xi, n_B - n_\xi, n_\xi) = \frac{\exp(-3)}{(n_A - n_\xi)!(n_B - n_\xi)! n_\xi!}. \tag{B.4}$$

By summing this over n_ξ, we see that

$$\begin{aligned}\Omega_2(n_A, n_B) &= \sum_{n_\xi=0}^{\min(n_A,n_B)} \frac{\exp(-3)}{(n_A - n_\xi)!(n_B - n_\xi)! n_\xi!} = \frac{\exp(-3)}{n_A! n_B!} \sum_{n_\xi=0}^{\min(n_A,n_B)} |\Xi(A, B, n_\xi)| \\ &= \frac{\exp(-3)}{n_A! n_B!} |\Xi(A, B)| = \exp(-1)|\Xi(A, B)|\Omega_1(n_A, n_B),\end{aligned} \tag{B.5}$$

where we used (2.25), the cardinality of $\Xi(A, B, n_\xi)$, in the second equality. Combining the above equations, we have

$$\begin{aligned}\mathrm{d}\Omega_2(A, B, \xi) &= \Omega_2(\xi \,|\, A, B, n_A, n_B)\Omega_2(n_A, n_B)\,\mathrm{d}\Omega_2(A, B \,|\, n_A, n_B) \\ &= \Omega_2(\xi \,|\, n_A, n_B)\Omega_2(n_A, n_B)\,\mathrm{d}\Omega_2(A, B \,|\, n_A, n_B) \\ &= \Omega_1(\xi \,|\, n_A, n_B)\Omega_2(n_A, n_B)\,\mathrm{d}\Omega_1(A, B \,|\, n_A, n_B) \\ &= \exp(-1)|\Xi(A, B)|\Omega_1(\xi \,|\, n_A, n_B)\Omega_1(n_A, n_B)\,\mathrm{d}\Omega_1(A, B \,|\, n_A, n_B) \\ &= \exp(-1)|\Xi(A, B)|\,\mathrm{d}\Omega_1(A, B, \xi)\end{aligned} \tag{B.6}$$

and thus the ratio of Ω_2 to Ω_1 is $\exp(-1)|\Xi(A, B)|$. This implies the Radon–Nikodym derivative for the change of measure from $\zeta \times \zeta \times \zeta_2$ to $\tilde{\zeta}$ is $\exp(-1)$.

§B.2 Change of variables from σ_A, σ_B to σ_P, σ_Q

The full conditional for σ_A and σ_B is proportional to

$$\begin{aligned}&\sigma_A^{-2(\alpha_\sigma + n_A)-3} \sigma_B^{-2(\alpha_\sigma + n_B)-3} \\ &\quad \times \exp\{-(R_1 + \beta_\sigma + k_\tau|\tau_A|^2)\sigma_A^{-2} - (R_2 + \beta_\sigma + k_\tau|\tau_B|^2)\sigma_B^{-2} - 2R_3/(\sigma_A\sigma_B)\},\end{aligned} \tag{B.7}$$

where R_1, R_2, R_3 are given in (3.14). This distribution has an intractable normalization constant. That is not a problem for sampling, as shown in §5.3. However, it prevents us from using directly using Chib's method in § 4.3.2. In this section we find a reparameterization where the normalization constant can be easily computed.

In the following we assume $n_A \geq n_B$ and define $\sigma_Q = \sigma_B/\sigma_A$ and $\sigma_P = 1/(\sigma_A\sigma_B)$. In those rare circumstances where the number of fingermark minutiae is greater than

the number of fingerprint minutiae (i.e., $n_B > n_A$), we redefine σ_Q as σ_A/σ_B and change the following equations as appropriate.

By changing variables to σ_P, σ_Q, we have

$$p(\sigma_P, \sigma_Q) \propto \sigma_P^{2\alpha_\sigma + n_A + n_B + 1} \sigma_Q^{n_A - n_B - 1} \times \exp[-\sigma_P\{(R_1 + \beta_\sigma + k_\tau|\tau_A|^2)\sigma_Q + (R_2 + \beta_\sigma + k_\tau|\tau_B|^2)/\sigma_Q + 2R_3\}]. \quad \text{(B.8)}$$

Thus the full conditional of σ_P is a gamma density with shape $2\alpha_\sigma + n_A + n_B + 2$ and rate $(R_1 + \beta_\sigma + k_\tau|\tau_A|^2)\sigma_Q + (R_2 + \beta_\sigma + k_\tau|\tau_B|^2)/\sigma_Q + 2R_3$. The full conditional of σ_Q is a generalized inverse Gaussian density (Barndorff-Nielsen, 1997) with density

$$p(\sigma_Q) = \frac{(R_1 + \beta_\sigma + k_\tau|\tau_A|^2)^{(n_A - n_B)/2}(R_2 + \beta_\sigma + k_\tau|\tau_B|^2)^{(n_B - n_A)/2}}{2K_{n_A - n_B}(2\sigma_P\sqrt{R_1 + \beta_\sigma + k_\tau|\tau_A|^2}\sqrt{R_2 + \beta_\sigma + k_\tau|\tau_B|^2})} \times \sigma_Q^{n_A - n_B - 1} \exp\left[-\sigma_P\{(R_1 + \beta_\sigma + k_\tau|\tau_A|^2)\sigma_Q + (R_2 + \beta_\sigma + k_\tau|\tau_B|^2)/\sigma_Q\}\right] \quad \text{(B.9)}$$

on $\sigma_Q \geq 0$, where $K_{n_A - n_B}$ is the modified Bessel function of the second kind and order $n_A - n_B$ (see Olver et al. 2010, §10 for definitions and properties).

Hence both σ_P and σ_Q have known normalization constants. In algorithm 4.1 we need to simulate from the distribution $\sigma_P \,|\, \sigma_Q$, which is simply gamma. At no point do we need to simulate from the Generalized Inverse Gaussian distribution $\sigma_Q \,|\, \sigma_P$, though sampling algorithms do exist (Dagpunar, 1989; Hörmann and Leydold, 2013).

§B.3 Approximation for the normalized posterior of κ

In order to use Chib's method we must find the normalized full conditional density of κ. It can be written as $\lambda_\kappa^{-1} \exp\{f(\kappa)\}$ where

$$f(\kappa) = -\eta\beta\kappa - \eta \log I_0(\kappa). \quad \text{(B.10)}$$

The normalization constant $\lambda_\kappa = \int_0^\infty \exp\{f(\kappa)\}\, d\kappa$ is intractable. We will estimate it with Laplace's method. First we must find the mode of the distribution, κ_0.

Define $R(\kappa)$ to be the ratio of the first- and zeroth- order modified Bessel functions of the first kind,

$$R(\kappa) = I_1(\kappa)/I_0(\kappa), \quad \text{(B.11)}$$

which is a strictly increasing function on $[0, \infty)$. Then κ_0 satisfies $R(\kappa_0) = -\beta$. Note if $\beta \geq 0$ the mode occurs at the boundary $\kappa = 0$ and the Laplace approximation will be inaccurate. Luckily for our application, β is almost always negative and the approximation works well.

By rearranging the bounds in Amos (1974), we have

$$\max\left(\frac{\kappa}{1+\sqrt{\kappa^2+1}}, \frac{\kappa}{0.5+\sqrt{\kappa^2+2.25}}\right) \leq R(\kappa) \leq \frac{\kappa}{\sqrt{\kappa^2+4}}, \tag{B.12}$$

which yields the inequality for κ_0

$$\max\left(\frac{-2\beta}{\sqrt{1-\beta^2}}, \frac{-\beta}{1-\beta^2}\right) \leq \kappa_0 \leq \frac{-2\beta}{1-\beta^2}. \tag{B.13}$$

Given β we can use the Newton-Raphson method restricted to this interval to find $\kappa_0 = r^{-1}(-\beta)$. The details are described in algorithm B.1.

Algorithm B.1 Newton-Raphson algorithm to find $\kappa = R^{-1}(x)$, for $x > 0$.

$L \leftarrow \max\left\{2x/\sqrt{1-x^2}, x/(1-x^2)\right\}$
$U \leftarrow 2x/(1-x^2)$
$\kappa \leftarrow (L+U)/2$
repeat
  $\kappa_0 \leftarrow \kappa$
  $R \leftarrow I_1(\kappa_0)/I_0(\kappa_0)$
  $g \leftarrow R - x$
  $g' \leftarrow 1 - R/\kappa_0 - r^2$
  $\kappa \leftarrow \max(L, \min(U, \kappa_0 - g/g'))$
until $|\kappa - \kappa_0| < 0.001$
return κ # κ solves $g(\kappa) = I_1(\kappa)/I_0(\kappa) - x = 0$

After computing κ_0, Laplace's method gives

$$\lambda_\kappa \approx \left\{\frac{2\pi}{\eta|\beta^2 - \beta/\kappa_0 - 1|}\right\}^{1/2} I_0(\kappa_0)^{-\eta} \exp(-\eta\beta\kappa_0). \tag{B.14}$$

This approximation is accurate provided κ_0 is bounded away from the boundary at zero. This is almost always the case in our application. On those rare instances when $\kappa_0 < 1$ the Laplace approximation becomes inaccurate and we resort to numerical integration to find λ_κ.

§B.4 Approximation for the normalized posterior of $\tilde{\omega}$

Chib's method requires us to compute the normalized full conditional of $\tilde{\omega}$. It can be written as $\lambda_\omega^{-1} \exp\{f(\tilde{\omega})\}\ \mathbb{1}(\tilde{\omega} \geq c_\omega)$ where

$$f(\tilde{\omega}) = (\alpha - 1)\log\tilde{\omega} - \beta\tilde{\omega} - \gamma\sqrt{\tilde{\omega}^2 - \tilde{\omega}}. \tag{B.15}$$

The normalization constant $\lambda_\omega = \int_{c_\omega}^{\infty} \exp\{f(\tilde{\omega})\}\,\mathrm{d}\tilde{\omega}$ is intractable. Unlike §B.3, we cannot use a Laplace approximation of λ_ω here because the mode often occurs at the boundary $\tilde{\omega} = c_\omega$.

We notice that for c_ω sufficiently large, $\sqrt{\tilde{\omega}^2 - \tilde{\omega}} \approx \tilde{\omega}$ and the full conditional is well approximated by a gamma distribution with shape α, rate $\beta + \gamma$, and left-truncation at c_ω. In our case $c_\omega = 65$ and the approximation is very good indeed. Integrating the gamma density, we see $\lambda_\omega \approx \Gamma\{\alpha, (\beta+\gamma)c_\omega\}(\beta+\gamma)^{-\alpha}$, where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function (Olver et al., 2010, §8.2).

§B.5 Expectation and variance of PPP test statistics

We wish to find the mean and variance of functions like

$$F(X) = \sum_{x\in X} \sum_{y\in X\setminus\{x\}} f(x,y), \tag{B.16}$$

where X is a Poisson point process with some intensity ρ and without loss of generality $f(x,y) = f(y,x)$. We shall use the *Slivnyak–Mecke* theorem (Mecke, 1967), which states that

$$\mathbb{E}\sum_{x\in X} h(x, X\setminus\{x\}) = \int \rho(x)\,\mathbb{E}\{h(x,X)\}\,\mathrm{d}x \tag{B.17}$$

for any real-valued function h. By directly applying this theorem to $h(x,X) = \sum_{y\in X} f(x,y)$, we have $\mathbb{E}\,F(X) = \int \rho(x)\,\mathbb{E}\{\sum_{y\in X} f(x,y)\}\,\mathrm{d}x$. By fixing x, we can applying it a second time with $h(y,X) = f(x,y)$ to find

$$\mathbb{E}\,F(X) = \int \rho(x)\rho(y) f(x,y)\,\mathrm{d}(x,y). \tag{B.18}$$

We are also interested in the variance of $F(X)$, for which we need to compute

$$\mathbb{E}\{F(X)^2\} = \mathbb{E} \sum_{x,y \in X} \sum_{u \in X \setminus \{x\}} \sum_{v \in X \setminus \{y\}} f(x,u) f(y,v). \tag{B.19}$$

We can partition the sums over $y \in X$ and $v \in X \setminus y$ into a term involving x and sum over $X \setminus \{x\}$, which allows us to apply (B.17). Repeating this procedure for the sum over u, after some algebra we find

$$\begin{aligned}\mathbb{E}\{F(X)^2\} = &\left\{ \int \rho(x)\rho(y) f(x,y) \,\mathrm{d}(x,y) \right\}^2 + 2 \int \rho(x)\rho(y) f(x,y)^2 \,\mathrm{d}(x,y) \\ &+ 4 \int \rho(x)\rho(y)\rho(z) f(x,y) f(x,z) \,\mathrm{d}(x,y,z)\end{aligned} \tag{B.20}$$

and

$$\begin{aligned}\mathbb{V}\{F(X)\} = &2 \int \rho(x)\rho(y) f(x,y)^2 \,\mathrm{d}(x,y) \\ &+ 4 \int \rho(x)\rho(y)\rho(z) f(x,y) f(x,z) \,\mathrm{d}(x,y,z).\end{aligned} \tag{B.21}$$

A discrepancy between the observed $F(X)$ and $\mathbb{E}\,F(X)$ provides evidence against the hypothesis that X follows a Poisson point process with intensity ρ. This cause of this discrepancy could be the number of points in X, or the position of those points, or a combination of both. In order to separate these two effects, we will consider the standardized statistic

$$\frac{F(X) - \mathbb{E}\{F(X) \,|\, n_X\}}{\sqrt{\mathbb{V}\{F(X) \,|\, n_X\}}}, \tag{B.22}$$

where $n_X = |X|$. It is well known that for any Poisson point process, the density of $X \,|\, n_X$ is $\rho_0^{-n_X} \prod_{x \in X} \rho(x)$, where $\rho_0 = \int \rho(x) \,\mathrm{d}x$ (see, e.g., Møller and Waagepetersen 2004, Proposition 3.8). By expanding out the expectation and variances, we see the expectation is

$$\mathbb{E}\{F(X) \,|\, n_X\} = n(n-1)\rho_0^{-2} \int \rho(x)\rho(y) f(x,y) \,\mathrm{d}(x,y), \tag{B.23}$$

and the variance is

$$\begin{aligned}\mathbb{V}\{F(X)\,|\,n_X\} = \frac{n(n-1)(n-2)(n-3)}{\rho_0^4}\int \rho(x)\rho(y)\rho(w)\rho(z)f(x,y)f(w,z)\,\mathrm{d}(w,x,y,z)\\ + \frac{4n(n-1)(n-2)}{\rho_0^3}\int \rho(x)\rho(y)\rho(z)f(x,y)f(x,z)\,\mathrm{d}(x,y,z)\\ + \frac{2n(n-1)}{\rho_0^2}\int \rho(x)\rho(y)f(x,y)^2\,\mathrm{d}(x,y) - \{\mathbb{E}_{X\,|\,n_X}F(X)\}^2.\end{aligned} \tag{B.24}$$

These integrals are analytically tractable for certain choices of ρ, including the complex normal densities used in this dissertation. Thus the standardized statistic (B.22) can be computed directly.

Bibliography

R. K. Ahuja and J. B. Orlin. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. *Operations Research*, 49:784–789, 2001.

C. C. G. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. Statistics in Practice. Wiley, Chichester, UK, 2nd edition, 2004.

A. A. Albert. On Jordan algebras of linear transformations. *Transactions of the American Mathematical Society*, 59:524–555, 1946.

M. P. Almeida and B. Gidas. A variational method for estimating the parameters of MRF from complete or incomplete data. *Annals of Applied Probability*, 3(1):103–136, 1993.

D. Amos. Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28(125), 1974.

T. W. Anderson. Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith, editors, *Essays in Probability and Statistics*, pages 1–24. University of North Carolina Press, Chapel Hill, N.C., 1970.

S. A. Andersson. Invariant normal models. *Annals of Statistics*, 3:132–154, 1975.

J. B. Anjos, Miguel F. Lasserre. *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer, 2012.

A. J. Baddeley, J. Møller, and R. Waagepetersen. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350, 2000.

D. J. Balding. When can a DNA profile be regarded as unique? *Science and Justice*, 39 (4):257–260, 1999.

D. J. Balding. *Weight-of-evidence for Forensic DNA Profiles*. Statistics in Practice. Wiley, Chichester, UK, 2005.

O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 89–96, New York, NY, USA, 2006. ACM.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

O. E. Barndorff-Nielsen. Plausibility inference. *Journal of the Royal Statistical Society Series B*, 38:103–131, 1976.

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, New York, 1978.

O. E. Barndorff-Nielsen. Normal inverse Gaussian distributions and stochastic volatil-

ity modelling. *Scandinavian Journal of Statistics*, 24(1):1–13, 1997.

O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Chapman and Hall, London, UK, 1994.

J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7:686–690, 1979.

D. J. Best and N. I. Fisher. Efficient simulation of the von Mises distribution. *Journal of the Royal Statistical Society Series C*, 28(2):152–157, 1979.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227, 02 2008.

Biometric System Laboratory, University of Bologna. FVC onGoing, 2012. URL `https://biolab.csr.unibo.it/FVCOnGoing/`.

F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105:8932–8937, 2008.

N. Brümmer and G. Doddington. Likelihood-ratio calibration using prior-weighted proper scoring rules. *arXiv preprint*, 2013. URL `http://arxiv.org/abs/1307.7981`.

N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, pages 230–275, 2006.

N. Brümmer and J. du Preez. The PAV algorithm optimizes binary proper scoring rules. *arXiv preprint*, 2013.

S. Buhl. On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, 20:263–270, 1993.

B. P. Carlin and S. Chib. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society Series B*, 1995.

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40:1935–1967, 08 2012.

R. C. H. Cheng. Generating beta variates with nonintegral shape parameters. *Communications of the ACM*, 21(4):317–322, 1978.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

S. A. Cole. More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law and Criminology*, 95(3):1–94, 2005.

R. G. Cowell, T. Graversen, S. Lauritzen, and J. Mortera. Analysis of forensic DNA mixtures with artefacts. *Journal of the Royal Statistical Society Series C*, 2014. To appear.

J. Dagpunar. Sampling of variates from a truncated gamma distribution. *Journal of Statistical Computation and Simulation*, 8:59–64, 1978.

J. Dagpunar. An easily implemented generalised inverse Gaussian generator. *Communications in Statistics - Simulation and Computation*, 18(2):703–710, 1989.

P. Damien and S. Walker. A full Bayesian analysis of circular data using the von Mises distribution. *The Canadian Journal of Statistics*, 27(2):291–298, 1999.

A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

A. P. Dawid. Probability forecasting. In *Encyclopedia of Statistical Sciences Volume 7*, pages 210–218. Wiley Interscience, 1986.

A. P. Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London, 1998.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

A. P. Dawid and S. Lauritzen. Compatible prior distributions. In *Bayesian methods with applications to science, policy and official statistics*, pages 109–118. International Society for Bayesian Analysis, 2000.

A. P. Dawid and S. Lauritzen. The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pages 22–28. University of Tokyo, 2005.

M. H. DeGroot and S. E. Feinberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1):12–22, 1983.

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196 – 212, 2004.

D. Edwards and T. Havránek. A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, 82:205–213, 1987.

D. Edwards, G. de Abreu, and R. Labouriau. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics*, 11:18, 2010.

H. Faulds. On the skin-furrows of the hand. *Nature*, 22(574):605, 1880.

J. Feng, A. K. Jain, and J. Zhou. Statistical modeling of fingerprint minutiae. Technical Report THU-IVG-TR-2011-1, Intelligent Vision Group at Tsinghua University, 2011.

P. G. M. Forbes. Assessing Probability Distributions. *Cambridge Masters of Advanced Studies in Mathematics Essay*, 2011. URL `http://www.stats.ox.ac.uk/~forbes/research/Part3Essay.pdf`.

P. G. M. Forbes. Compatible weighted proper scoring rules. *Biometrika*, 99(4):989–994, 2012.

P. G. M. Forbes and S. Lauritzen. Fingerprint analysis using Bayesian alignment. In *Proceedings of the Leeds Annual Statistics Research Workshop*, 2013. URL `http://www1.maths.leeds.ac.uk/statistics/workshop/lasr2013/proceedings/Forbes.pdf`.

P. G. M. Forbes and S. Lauritzen. Linear Estimating Equations for Exponential Families with Application to Gaussian Linear Concentration Models. *Linear Algebra and its Applications*, 2014. doi: 10.1016/j.laa.2014.08.015.

P. G. M. Forbes and K. V. Mardia. A Fast Algorithm for Sampling from the Posterior of a von Mises distribution. *Journal of Statistical Computation and Simulation*, 2014. doi: 10.1080/00949655.2014.928711.

P. G. M. Forbes, S. Lauritzen, and J. Møller. Fingerprint analysis with marked point

processes. *arXiv preprint*, 2014. URL http://arxiv.org/abs/1407.5809.

J. Frellsen, I. Moltke, M. Thiim, K. V. Mardia, J. Ferkinghoff-Borg, and T. Hamelryck. A probabilistic model of local RNA 3-D structure. *Public Library of Science Computational Biology*, 5:1–11, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. *glasso: Graphical lasso- estimation of Gaussian graphical models*. Comprehensive R Archive Network, 2011. URL http://CRAN.R-project.org/package=glasso. R package version 1.7.

M. Galassi, J. Davies, J. Theiler, B. Gough, and G. Jungman. *GNU Scientific Library Reference Manual, Third Edition*. Network Theory Ltd, 2009.

S. F. Galton. *Finger Prints*. Macmillan, London, 1892. URL http://galton.org/books/finger-prints/galton-1892-fingerprints-1up-lowres.pdf.

M. Garris and R. McCabe. NIST special database 27: Fingerprint minutiae from latent and matching tenprint images. Technical report, NIST, Gaithersburg, MD, USA, 2000.

H. Gehrmann. Lattices of graphical Gaussian models with symmetries. *Symmetry*, 3: 653–679, 2011.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

V. P. Godambe. *Estimating Functions*. Clarendon Press, 1991.

O. L. Gonzalez, M. A. M. Perez, A. E. G. Rodriguez, and M. G. Borroto. A framework in C# for fingerprint verification, 2012. URL http://www.codeproject.com/Articles/97590/A-Framework-in-C-for-Fingerprint-Verification.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society Series B*, 14: 107–114, 1952.

N. R. Goodman. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *Annals of Mathematical Statistics*, 34(1):pp. 152–177, 1963.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. J. Green and K. V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254, 2006.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

P. Guttorp and R. A. Lockhart. Finding the location of a signal: A Bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988.

E. J. Hannan and B. G. Quinn. The determination of an order of an autoregression. *Journal of the Royal Statistical Society Series B*, 41:190–195, 1979.

A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

S. W. J. Herschel. *The Origin of Finger-Printing*. Oxford University Press, 1916. URL http://www.gutenberg.org/files/34859/34859-h/34859-h.htm.

T. Höglund. The exact estimate — a method of statistical estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 29:257–271, 1974.

S. Højsgaard and S. Lauritzen. Inference in graphical Gaussian models with edge and vertex symmetries with the gRc package for R. *Journal of Statistical Software*, 23

(6):1–26, 2007.

S. Højsgaard and S. Lauritzen. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society Series B*, 70:1005–1027, 2008.

J. E. Hoover. *The Science of Fingerprints: Classification and Uses*. Federal Bureau of Investigation, U.S. Department of Justice, 1963. URL `http://www.gutenberg.org/files/19022/19022-h/19022-h.htm`.

W. Hörmann and J. Leydold. Generating generalized inverse Gaussian random variates. *Statistics and Computing*, pages 1–11, 2013.

P. J. Huber. Robust estimation of a location parameter. *Annals of Applied Statistics*, 35 (1):73–101, 1964.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In L. M. L. Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–223, Berkeley, CA, 1967. University of California Press.

S. Huckerman, T. Hotz, and A. Munk. Global models for the orientation field of fingerprints: an approach based on quadratic differentials. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 30(9):1507–1519, 2008.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51:2499–2512, 2007.

IAI, 2010. *IAI Resolution 2010-18*, 2010. International Association for Identification. URL `http://www.onin.com/fp/IAI_resolution_2010-18.pdf`.

S. Jammalamadaka and Y. Sarma. A correlation coefficient for angular variables. In *Statistical Theory and Data Analysis II: Proceedings of the Second Pacific Area Statistical Conference*. North Holland: New York, 1988.

S. T. Jensen. Covariance hypotheses which are linear in both the covariance and the inverse covariance. *Annals of Statistics*, 16:302–322, 1988.

D. J. Johnstone. Economic interpretation of probabilities estimated by MLE or score. *Management Science*, 57(2):308–314, 2011.

D. J. Johnstone and Y.-X. Lin. Fitting probability forecasting models by scoring rules and maximum likelihood. *Journal of Statistical Planning and Inference*, 141(5): 1832–1837, 2011.

V. R. R. Jose, R. F. Nau, and R. L. Winkler. Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56(5):1146–1157, 2008.

V. R. R. Jose, R. F. Nau, and R. L. Winkler. Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, 55(4):582–590, 2009.

J. T. Kent. The infinite divisibility of the von Mises–Fisher distribution for all values of the parameter in all dimensions. *Proceedings of the London Mathematical Society*, 35(3):359–384, 1977.

J. T. Kent and K. V. Mardia. The link between kriging and thin-plate splines. In F. P. Kelly, editor, *Probability, Statistics and Optimisation*. John Wiley and Sons, 1994.

J. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

S. Lauritzen, R. G. Cowell, and T. Graversen. Discussion on the paper by Neumann et al. (2012). *Journal of the Royal Statistical Society Series A*, 175(2):405–406, 2012.

D. V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977.

S. Ma, Q. Gong, and H. J. Bohnert. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research*, 17:1614–1625, 2007.

D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag, New York, 2nd edition, 2009.

K. V. Mardia. On some recent advancements in applied shape analysis and directional statistics. In *Proceedings of the Leeds Annual Statistics Research Workshop*, 2007. URL `https://www1.maths.leeds.ac.uk/statistics/workshop/lasr2007/proceedings/mardia.pdf`.

K. V. Mardia. Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society Series C*, 62:487–514, 2013.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, Chichester, UK, 2nd edition, 1999.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

G. Marsaglia. The exact-approximation method for generating random variables in a computer. *Journal of the American Statistical Association*, 79(385):218–221, 1984.

G. Marsaglia. Xorshift RNGs. *Journal of Statistical Software*, 8(14):1–6, 2003.

G. Marsaglia and W. W. Tsang. A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, 26(3):363–372, 2000a.

G. Marsaglia and W. W. Tsang. The ziggurat method for generating random variables. *Journal of Statistical Software*, 5(8):1–7, 2000b.

P. Martin-Löf. Exact tests, confidence regions and estimates. *Synthese*, 36:195–206, 1977.

J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42:654–655, 1956.

J. Mecke. Stationäire zufällige masse auf localcompakten abelischen gruppen. *Z. Wahrscheinlichkeitsth*, 9:36–58, 1967.

X. Meng and W. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

A. Mikalyan and J. Bigun. Ground truth and evaluation for latent fingerprint matching. In *CVPR Workshop on Biometrics*, 2012.

A. Mira and G. Nicholls. Bridge estimation of the probability density at a point. *Statistica Sinica*, 14(2):603–612, 2004.

J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton, 2004.

A. H. Murphy. Scalar and vector partitions of the probability score, part II. *Journal of Applied Meteorology*, 11(5):1183–1192, 1972.

C. Neumann, I. W. Evett, and J. E. Skerrett. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm (with discussion). *Journal of the Royal Statistical Society Series A*, 175(2):371–415, 2012a.

C. Neumann, I. W. Evett, J. E. Skerrett, and I. Mateos-Garcia. Quantitative assessment of evidential weight for a fingerprint comparison. Part II: A generalisation to take account of the general pattern. *Forensic Science International*, 214:195–199, 2012b.

M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted

likelihood bootstrap. *Journal of the Royal Statistical Society Series B*, 56(1):3–48, 1994.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Annals of Statistics*, 1:763–765, 1973.

F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010.

P. E. Peterson, C. B. Dreyfus, M. R. Gische, M. Hollars, M. A. Roberts, R. M. Ruth, H. M. Webster, and G. L. Soltis. Latent prints: A perspective on the state of the science. *Forensic Science Communications*, 11(4), 2009. URL `http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/oct2009/review`.

A. E. Raftery, M. A. Newton, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics 8*, pages 1–45, 2007.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

W. L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society A*, 232 (1188):6–31, 1955.

M. Specter. Do fingerprints lie? *Annals of Crime, The New Yorker*, May 2002. URL `http://www.newyorker.com/archive/2002/05/27/020527fa_FACT`.

D. Stoney. Measurement of fingerprint individuality. In *Advances in Fingerprint Technology, Second Edition*, pages 327–388. CRC Press, 2001.

C. E. Troup. Identification of habitual criminals. In *House of Commons Parliamentary Papers, Great Britain*, 1894. URL `http://parlipapers.chadwyck.co.uk/fullrec/fullrec.do?id=1893-070826`.

C. Uhler. Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics*, 40(1):238–261, 2012.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998.

T. van Erven, P. Grünwald, and S. de Rooij. Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma (with discussion). *Journal of the Royal Statistical Society Series B*, 74(3): 361–417, 2012. ISSN 1467-9868.

R. Vazam. SourceAFIS, 2012. URL `http://sourceforge.net/projects/sourceafis/`.

S. Winitzki. Uniform approximations for transcendental functions. In *Computational Science and Its Applications*, volume 2667 of *Lecture Notes in Computer Science*, pages 780–789. Springer, 2003.

C. F. J. Wu. On the convergence properties of the em algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

N. Yager and A. Amin. Fingerprint classification: a review. *Pattern Analysis and Applications*, 7(1):77–93, 2004a.

N. Yager and A. Amin. Fingerprint verification based on minutiae features: a review. *Pattern Analysis and Applications*, 7(1):94–113, 2004b.