Distance Measures for Probabilistic Patterns

by

Ian Kennedy

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Paul McNicholas Professor and University Scholar, Dept. of Mathematics and Statistics, McMaster University
Supervisor(s):	Paul Fieguth Professor, Dept. of Systems Design Engineering, University of Waterloo
Internal Member:	David Clausi Professor, Dept. of Systems Design Engineering, University of Waterloo
Internal Member:	Alexander Wong Associate Professor, Dept. of Systems Design Engineering, University of Waterloo
Internal-External Member:	Shoja'eddin Chenouri Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Otman Basir Professor, Dept. of Electrical and Computer Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Numerical measures of pattern dissimilarity are at the heart of pattern recognition and classification. Applications of pattern recognition grow more sophisticated every year, and consequently we require distance measures for patterns not easily expressible as feature vectors. Examples include strings, parse trees, time series, random spatial fields, and random graphs [79] [117].

Distance measures are not arbitrary. They can only be effective when they incorporate information about the problem domain; this is a direct consequence of the Ugly Duckling theorem [37].

This thesis poses the question: how can the principles of information theory and statistics guide us in constructing distance measures? In this thesis, I examine distance functions for patterns that are maximum-likelihood model estimates for systems that have random inputs, but are observed noiselessly. In particular, I look at distance measures for histograms, stationary ARMA time series, and discrete hidden Markov models.

I show that for maximum likelihood model estimates, the L_2 distance involving the information matrix at the most likely model estimate minimizes the type II classification error, for a fixed type I error. I also derive explicit L_2 distance measures for ARMA(p, q) time series and discrete hidden Markov models, based on their respective information matrices.

Acknowledgements

I wish to thank my supervisor Paul Fieguth for accepting me as a part-time graduate student. Paul's assistance, guidance, and insight have been invaluable. I have benefitted greatly from my work with Paul, and hope that my thesis meets his exacting standards.

I would also like to thank Vicky Lawrence, Janine Blair, and Lauren Gatchene for their administrative work in regards to my graduate program. Also, I wish to acknowledge funding received through the NSERC Discovery Grants Program.

Finally, I would like to thank Elizabeth Janssen, who has been patiently and generously supportive while I pursued graduate studies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and problem statement

Technology is at heart a search for useful devices and algorithms. Its success rests upon our ability to discover and verify regularities, both in natural processes and in those processes due to human activity.

Some processes are not deterministic, and hence are not entirely predictable. They admit some random inputs. That is where pattern recognition comes in: it attempts to discover the underlying regularities in events, or in series of events, that have some random characteristics.

So we are interested in patterns of form, behavior, or both together. But what are patterns? First of all, a pattern implies an ensemble. We cannot deduce a pattern from a single event. Patterns require repetition. Thus, pattern analysis will necessarily involve the techniques of statistical analysis. I will consider patterns to be probabilistic graphical structures that have a grammar or rule that defines their possible construction. See Grenander [55] for more details about this perspective.

Pattern recognition and classification require a numerical measure of dissimilarity between graphical structures. Such a measure takes a pair of graphical structures as input, and returns a non-negative number indicating how dissimilar the structures are, returning zero if the structures are identical. I will define these distance measures more explicitly in Chapter 2.

As computers become more capable, we expect them to assist us in pattern recognition tasks of ever greater complexity, typically ones requiring experience and judgment. The more sophisticated the application, the more likely is it that the underlying patterns will be represented by structured data instead of feature vectors. Here are some examples of technology requiring structured patterns:

1. Autonomous vehicles need to interpret their surroundings and make navigational decisions at least as well as people do, which is to say very rapidly, with reaction times measured in milliseconds. This applies to obstacle recognition in driverless cars, terrain recognition in airborne drones, and environmental assessment for remotely operating robots [102].

2. As the population ages, governments and insurance companies spend increasing amounts on health care. It is of interest to all parties to have algorithms that perform routine diagnosis and classification into risk groups, based upon patient records with heterogeneous data. See Ruiz et al. [98] for an example that diagnoses Alzheimer's disease through an analysis of MRI scans.

3. Online platforms use 'hypertargeting' of paid advertisements for political and commercial goals. Although an annoyance for some, and ethically dubious for others, such ads are nevertheless extraordinarily effective [9].

Clearly, sophisticated applications in pattern classification require distance measures that incorporate realistic representations of domain knowledge [56] [38]. Chapters 3, 4, and 5 will show that, for numerical patterns, that domain knowledge must include at least the second-order pattern statistics (i.e. the pattern covariance).

My overall goal, then, is to construct probabilistic models for observable events or sequences of events, and construct appropriate distance measures for these models. By "model", I mean a numerical structure and associated rule set that describes a set of patterns (see Section 2.1). My central question is: how can the principles of information theory and statistics guide us in formulating distance measures, specifically for probabilistic models obtained through maximum likelihood estimation? My thesis is that the principles of information theory and statistics say a lot about distance measures, but do not answer all possible questions. There is still room for human judgment.

Of course, we have to narrow down the topic. I will be considering probabilistic models derived through maximum likelihood estimation from large datasets, with no observational noise [46]. In other words, the datasets are only random in ways envisioned by the model. I will be looking at three specific types of models:

1. Histograms: this describes a dataset of random variates with categories, but no correlation.

2. Autoregressive moving average time series: this describes a dataset of random variates that have direct sequential correlation.

3. discrete hidden Markov model: this describes a dataset of discrete random variates, that have sequential correlation due to a hidden state, which is itself a discrete random variable.

We will also address the inevitable question of the "best" distance measure. I regard the question of distance measure comparison as a statistical one, namely: for a fixed type I error, which function leads to the smallest type II error when deciding whether a pattern belongs to a cluster? Chapter 2 will expand on this idea.

The remainder of this thesis is organized as follows: Chapter 2 will present background information about maximum likelihood estimation, distance measures in general, type I and type II errors, and the three types of model that I consider in this thesis (histograms, ARMA time series, and discrete hidden Markov models). In particular, for each model type, I describe the model in detail and describe distance measures that other researchers have used for that model type.

Chapter 3 will look at distance measures for histograms. In that chapter I derive the information matrix for histograms and demonstrate why we might prefer the L_2 distance measure for this case. I also look briefly at calculating the optimum number of bins for a histogram, and at distance measures based on sample cumulants and sample cumulative distribution functions.

Chapter 4 looks at a general formulation of an L_2 distance measure for maximum-likelihood models, and then applies it to ARMA(p, q) time series models. I show that the derived parameter covariance matches up with experiment, and give an example of the L_2 distance use with GARCH(1,1) models for financial time series. The chapter closes with observations concerning other candidate distance measures for ARMA(p, q) time series models.

Chapter 5 contains my derivation of an L_2 distance measure for stationary discrete hidden Markov models. In that chapter, I also derive the variance of empirical symbol block frequencies and verify the calculation through experiment. I look at the issues involved in doing parameter estimation by matching symbol block frequencies. The chapter concludes with an approximate distribution of the L_2 distance when the hidden Markov models are materially different.

Chapter 6 contains my conclusions, outline of contributions to knowledge, and suggestions for further work. The appendices contain all the mathematical derivations that the main chapters rely on.

Chapter 2

Background and previous work

2.1 Maximum likelihood estimation

Our starting point is maximum likelihood estimation. We suppose that there is a model, characterized by a set of parameters $\vec{\theta}_0$ which is not observed directly. Instead, the model gives rise to a possibly ordered set of observations or events, that have some element of randomness. In this case, probability comes into play. In particular, if the underlying model has parameters $\vec{\theta}_0$ then we may seek to estimate these parameters through the optimization problem:

$$\text{[MLE] Minimize } \mathcal{L} = -\log p(D, \vec{\theta}) \tag{2.1}$$

where the observed data D is fixed, the model parameters $\vec{\theta}$ are variable, and the likelihood $p(D, \vec{\theta})$ is the probability of observing the dataset D under the model with parameters $\vec{\theta}$ (see Figure 2.1). The likelihood function is the likelihood $p(D, \vec{\theta})$ with the dataset fixed, i.e. viewed as a function of $\vec{\theta}$ only. The total sample surprisal is $\mathcal{L} = -\log p(D, \vec{\theta})$. Since the logarithm function is monotonic increasing, minimizing the total sample surprisal amounts to maximizing the likelihood function. For the purposes of this thesis, I assume that the likelihood function is discoverable.

The MLE optimization problem of equation 2.1 may sound straightforward, but there are three issues which can severely affect the optimization result:

1. We chose the model incorrectly. For example, we may be estimating a Gaussian distribution, with no cumulants beyond the second, when the actual distribution is a beta or Pearson type 4 that definitely has third and fourth cumulants.

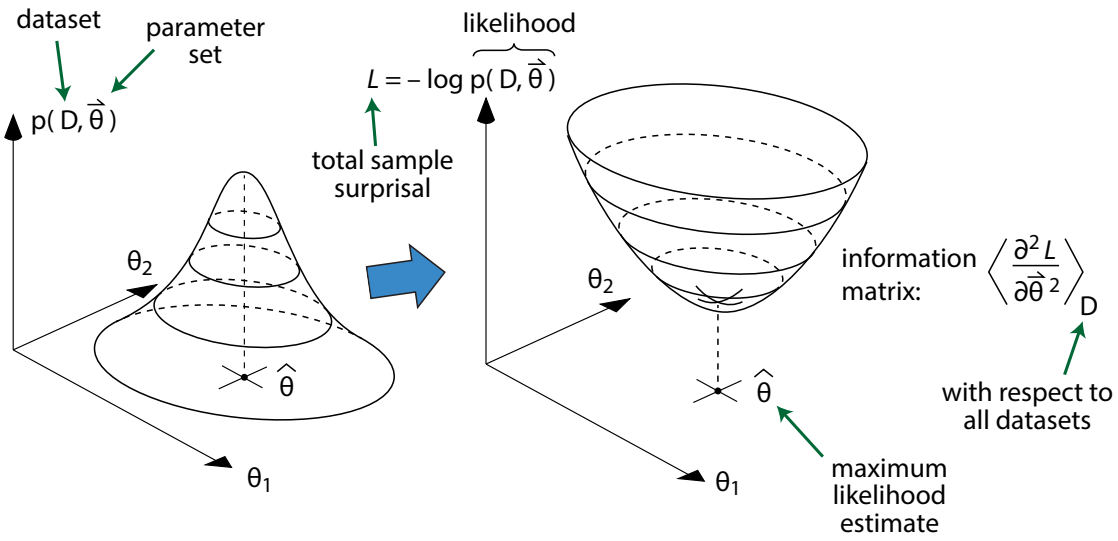


Figure 2.1: Maximum likelihood estimation - no constraints active

2. The underlying model is of the correct type, but it has too few or too many parameters.

3. The actual minimum of the total sample surprisal $\mathcal{L}$ occurs where a parameter constraint is active. In this case, we do not have the usual condition $\partial\mathcal{L}/\partial\vec{\theta} = 0$ at the minimum (see Figure 2.2).

I chose the examples in this thesis so that none of these basic issues occur. If the model is not of the correct type, maximum likelihood estimation is not a total loss: as the sample size becomes infinite, the resulting distribution will have the minimum Kullback-Leibler divergence with the true distribution [112]. I also assume that any prior distribution on the model parameters $\vec{\theta}$ is uninformative. In other words, given the model type and observed dataset D, I regard $\vec{\theta}_0$ as a parameter set to be discovered. For more information on the limitations of maximum likelihood estimation, see [20].

The goal of this thesis is to find out what information theory has to say about computing a numeric similarity between parameter set estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, not necessarily of the same size, derived from datasets D_1 and D_2. Such a similarity measure would ideally be

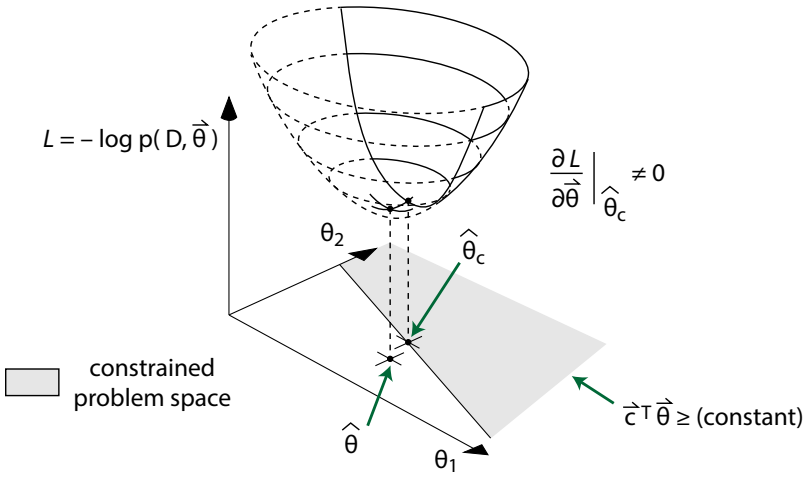


Figure 2.2: Maximum likelihood estimation with an active parameter constraint. The constrained parameter space is shown in grey. In this case, the gradient of the total sample surprisal may be non-zero at the minimum.

a function of $\hat{\theta}_1, \hat{\theta}_2, |D_1|$ and $|D_2|$ only, where $|D_k|$ refers to the size of dataset D_k. In this thesis I only consider the case of large datasets, i.e. $|D_1|, |D_2| \gg 1$.

We will review some basic results in maximum likelihood estimation; for more information, consult [29] and [105]. Suppose first that the data set D contains n observations, where $n \gg 1$ (typically $n > 1000$). As regularity conditions, we require that $\partial\mathcal{L}/\partial\vec{\theta}$, $\partial^2\mathcal{L}/\partial\vec{\theta}^2$, and $\partial^3\mathcal{L}/\partial\vec{\theta}^3$ all exist and are bounded in a neighborhood of the true model parameter set $\vec{\theta}_0$, and also that $\langle\partial\mathcal{L}/\partial\vec{\theta}\rangle_D = 0$ where the expectation is with respect to all possible data sets of size n. Let U be the single-observation score, i.e.

$$U = \left(\frac{1}{n}\right)\partial\mathcal{L}/\partial\vec{\theta} \tag{2.2}$$

and let the single-observation information matrix be

$$i(\vec{\theta}) = \langle\partial U/\partial\vec{\theta}\rangle_D = \frac{1}{n}\langle\partial^2\mathcal{L}/\partial\vec{\theta}^2\rangle_D \tag{2.3}$$

where the expectation is with respect to all possible data sets of size n. Then a minimum of the total sample surprisal $\mathcal{L}$ will occur at the parameter set estimate $\hat{\theta}$ where $\partial\mathcal{L}/\partial\vec{\theta} = 0$

and $i(\hat{\theta})$ is positive definite, and to lowest order

$$\langle \hat{\theta} - \vec{\theta}_0 \rangle_D = 0, \qquad \mathrm{var}(\hat{\theta}) = \frac{1}{n} i^{-1}(\vec{\theta}_0) \tag{2.4}$$

Thus, to lowest order, the maximum likelihood parameter estimates for data sets of size n form a Gaussian cluster around $\vec{\theta}_0$ with variance $i^{-1}(\vec{\theta}_0)/n$ (see Figure 2.1). As for the next-lowest order terms, they are

$$\langle \hat{\theta} - \vec{\theta}_0 \rangle_D = \frac{1}{n} b(\vec{\theta}_0), \qquad \mathrm{var}(\hat{\theta}) = \frac{1}{n} i^{-1}(\vec{\theta}_0) + \frac{1}{n^2} C(\vec{\theta}_0) \tag{2.5}$$

where $b(\vec{\theta}_0)$ and $C(\vec{\theta}_0)$ are functions of the third-order derivatives of U [30]. The functional forms of $b(\bullet)$ and $C(\bullet)$ are not essential here; the important feature of equation 2.5 is that its corrections to equation 2.4 are negligible when $n \gg 1$. Note that the total sample surprisal $\mathcal{L}$ may have several local minima. We are interested in the global minimum within the allowable parameter space. If it is not unique, then we need to apply some selection rule (see section 5.9 for an example involving hidden Markov models).

2.2 Distance measures

In the context of pattern matching and classification, distance means dissimilarity: the distance between two patterns is a real number that represents how dissimilar the two patterns are. A distance metric is a real-valued function $d(\bullet, \bullet)$ defined for all pairs of patterns from a set $W = \{\ldots x, y, z, \ldots\}$ with the following properties [109]

1. $d(x, y) \in [0, \infty)$ (non-negative real number)
2. $d(x, y) = 0 \Rightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

A distance function that has the first three properties but not the triangle inequality is a semimetric; I also refer to this kind of function as a distance measure [11]. We want to allow semimetrics, such as information divergence, since there are several classification algorithms in broad use that depend on them; see [32] for further information. Furthermore, we want to allow distance functions that mimic human judgments of dissimilarity, which are known

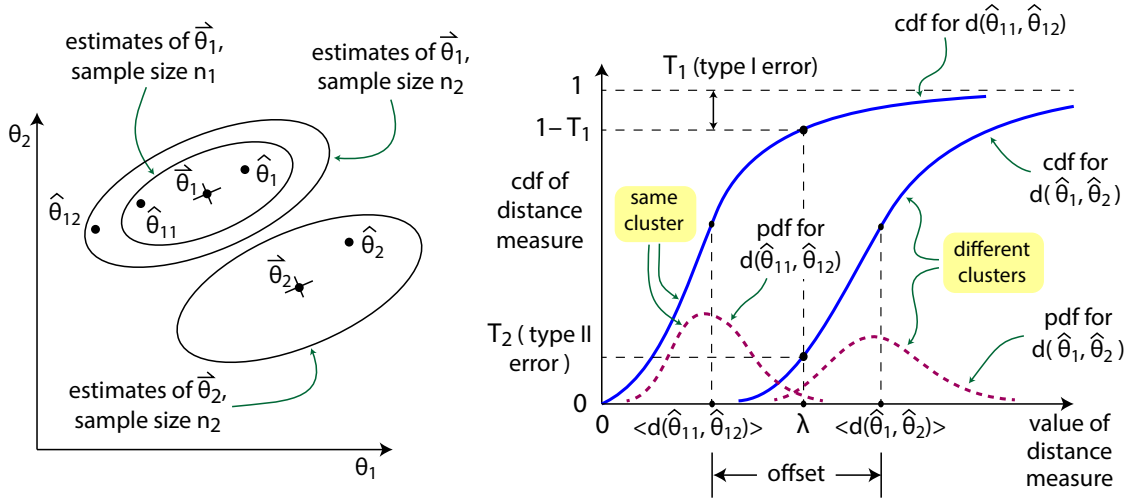


Figure 2.3: Maximum likelihood estimation: type I and type II errors. The null hypothesis is that parameter set estimates belong to the same cluster.

to be non-metric [110] [119] [50]. For the purposes of this thesis, distance measures have only one use, namely the classification of model parameter set estimates $\{\hat{\theta}_1, \hat{\theta}_2, \ldots\}$ into clusters. By cluster, I mean a collection of parameter set estimates derived from a single underlying model. I am interested specifically in the cluster separation issue: whether two parameter estimates belong to the same cluster or not. In particular, for maximum-likelihood parameter estimates $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ derived from datasets of fixed sizes n_1 and n_2 respectively, generated by a true model $\vec{\theta}_1$, we expect a spread for $d(\hat{\theta}_{11}, \hat{\theta}_{12})$ where $d(\bullet, \bullet)$ is a distance measure of interest (see Figure 2.3). In Figure 2.3, a cluster is represented by the underlying model and a curve of constant cdf (standardized at 0.68). Our null hypothesis is that $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ come from clusters based on $\vec{\theta}_1$.

In other words, if $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ are drawn randomly from the clusters around $\vec{\theta}_1$ corresponding to the dataset sizes, then $d(\hat{\theta}_{11}, \hat{\theta}_{12})$ will be a random quantity with its own probability distribution function (pdf) and cumulative distribution function (cdf). We have to decide the type I error allowable, that being the probability of deciding that $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ are based on different models, when in fact they are based on the same model $\vec{\theta}_1$. The corresponding threshold λ will be the value of $d(\hat{\theta}_{11}, \hat{\theta}_{12})$ at which the cdf reaches

the value $1 - T_1$, where T_1 is the permissible type I error. In this thesis, I use $T_1 = 0.05$ throughout.

Suppose now that $\vec{\theta}_1$ represents a base model, and that $\vec{\theta}_2$ represents a materially different model, that is one for which $p(D, \vec{\theta}_1) \neq p(D, \vec{\theta}_2)$ for some datasets D. If $\hat{\theta}_1$ is a maximum likelihood estimate of $\vec{\theta}_1$ derived from a dataset of size n_1, and $\hat{\theta}_2$ is a maximum likelihood estimate of $\vec{\theta}_2$ derived from a dataset of size n_2, then $d(\hat{\theta}_1, \hat{\theta}_2)$ will have a spread with its own pdf and cdf. The type II error T_2 is the probability of deciding that $\hat{\theta}_1$ and $\hat{\theta}_2$ are based on the same model $\vec{\theta}_1$ when in fact they are based on different models. It will be the value of the cdf of $d(\hat{\theta}_1, \hat{\theta}_2)$ at the type I threshold λ (see Figure 2.3). Note that the null hypothesis here is that $\hat{\theta}_1$ and $\hat{\theta}_2$ come from datasets based on $\vec{\theta}_1$. The type II error could possibly be different if $\vec{\theta}_2$ were the base model, i.e. if the null hypothesis were that $\hat{\theta}_1$ and $\hat{\theta}_2$ come from datasets based on $\vec{\theta}_2$.

In general, our goal in pattern classification is to minimize the total risk, that being $\alpha_1 T_1 + \alpha_2 T_2$ where α_1 is the risk of a type I error, and α_2 is the risk of a type II error [39]. The relative risk of type I and type II errors varies with the problem domain. In a health care setting, for example, the null hypothesis is that the patient is healthy. A type I error would be to classify the patient as ill, when in fact he is healthy. A type II error would be to classify the patient as healthy, when in fact he is sick. In this setting, a type II error is far more serious that a type I error. A criminal court of law is an example of the opposite case. There, the null hypothesis is that the accused is innocent. A type I error would be to find him guilty, when in fact he is innocent. A type II error would be to declare him innocent when in fact he is guilty, presumably due to a lack of convincing evidence. The British legal system has always held that a type I error is far more serious that a type II error; in fact the principle derives from Roman law and features prominently in the code of Justinian [107]. For the purposes of this thesis, I will assume that $\alpha_2 \gg \alpha_1$. In other words, we will fix T_1 and measure classification accuracy as $1 - T_2$. Our goal, then, in choosing $d(\bullet, \bullet)$ is to minimize T_2.

We can definitely say that if we keep $\langle d(\hat{\theta}_{11}, \hat{\theta}_{12}) \rangle$ and $\langle d(\hat{\theta}_1, \hat{\theta}_2) \rangle$ constant, but decrease the variance of $d(\hat{\theta}_1, \hat{\theta}_2)$, then the type II error will decrease (see Figure 2.4).

However, it may happen that an alternative distance measure $d_B(\bullet, \bullet)$ increases the variance for a given $\langle d_B(\hat{\theta}_1, \hat{\theta}_2) \rangle$ but also increases the offset $\langle d_B(\hat{\theta}_1, \hat{\theta}_2) \rangle - \langle d_B(\hat{\theta}_{11}, \hat{\theta}_{12}) \rangle$. In this case, we cannot say with certainty what happens to the type II error. It may still decrease (see Figure 2.5).

At times we will want to compare distance measures for classification accuracy, and we can boil that down to the question: which measure gives us the smallest type II error,

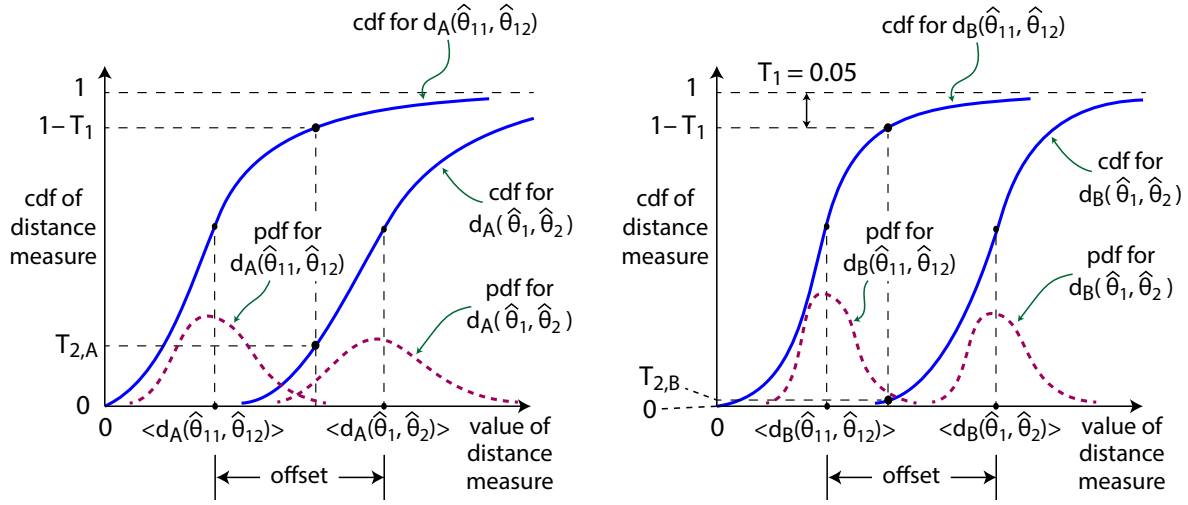


Figure 2.4: Illustration of how the type II error will decrease with constant offset and decreasing distance measure variance. $d_B(\bullet, \bullet)$ (on the right) has the lower distance measure variance and lower type II error.

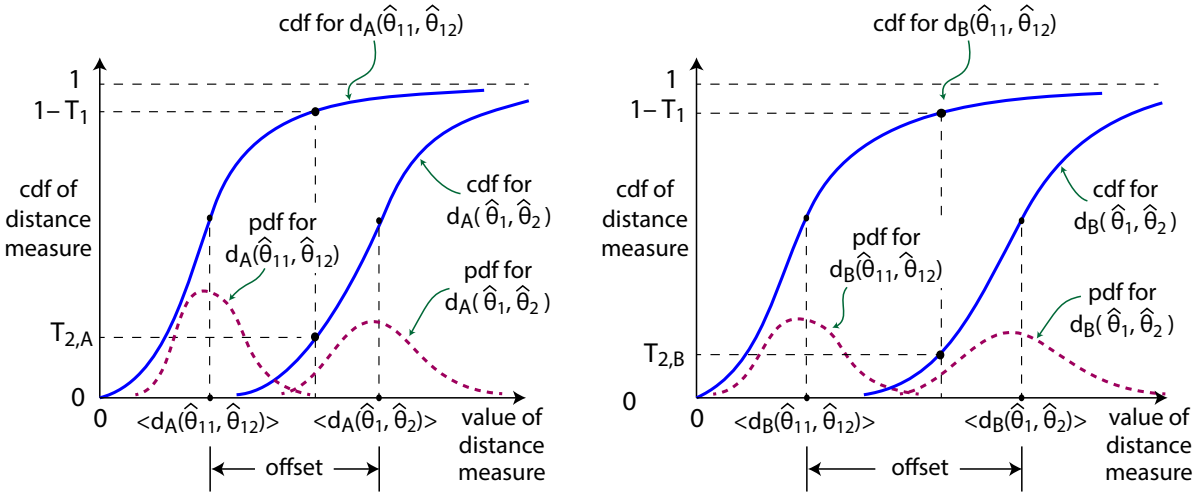


Figure 2.5: Illustration of how the type II error may decrease with increasing offset and increasing distance measure variance

holding the base models $(\vec{\theta}_1, \vec{\theta}_2)$ and dataset sizes (n_1, n_2) constant? Note that the only case of real interest is when the offset is comparable to the type I error threshold λ (see Figure 2.3). If the offset is much greater than the type I error threshold, then the type II error will be close to zero no matter what the variance of $d(\hat{\theta}_1, \hat{\theta}_2)$ is. Put simply: if the two clusters are far enough apart, any distance measure will separate them.

2.3 Histograms

A histogram is derived from a counting process. More specifically, we observe a set of n random events, each of which is unambiguously classifiable into one of c categories (or bins), labelled $\{1, 2, \ldots, c\}$. We can represent the histogram as a list of raw counts $[n_1, \ldots, n_c]$ where n_j is the number of events falling into category j, or as a list of proportions along with the sample size: $\{[p_1, \ldots, p_c], n\}$ where $p_j = n_j/n, 0 \leq p_j \leq 1$ and $\sum p_j = 1$. I will call a histogram normalized when its entries sum to one. Histograms are not vectors. There is no concept of rotation, for example; under any orthogonal transform other than a permutation, some histograms will become invalid (see Figure 2.6 for an example with

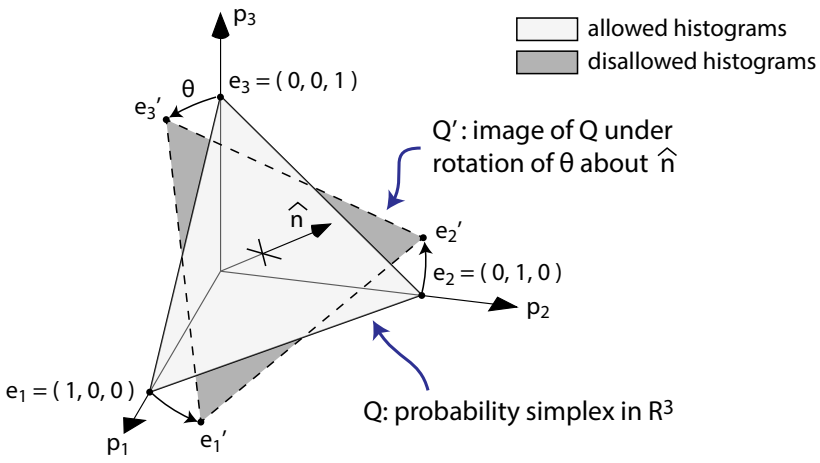


Figure 2.6: Histograms do not support arbitrary rotations

$c = 3$). Also, histograms may allow the operations of bin splitting and bin combination, which are not well defined for vectors. A histogram's bins need not be numeric, and they need not possess an order.

Comparing histograms is a non-parametric way of comparing pdfs, and there are many possible ways of performing the comparison; for a full list of these distance measures, see Cha's comprehensive survey [21]. There are three that I will mention throughout this thesis. With $\{p_1, \ldots, p_c\}$ and $\{q_1, \ldots, q_c\}$ being the normalized histograms in question, these are:

$$
\begin{aligned}
d_H^2 &= \sum_{j=1}^{c} (\sqrt{p_i} - \sqrt{q_i})^2 = 2(1 - \sum_{j=1}^{c} \sqrt{p_i q_i}) \quad [\text{ Hellinger, Cha \#35 }] \\
d_B^2 &= -\log \sum_{j=1}^{c} \sqrt{p_i q_i} \quad [\text{ Bhattacharyya, Cha \#33 }] \\
d_{KL}^2 &= \sum_{j=1}^{c} p_j \log (p_j / q_j) \quad [\text{ Kullback-Leibler, Cha \# 48 }]
\end{aligned}
\tag{2.6}
$$

Note that the Kullback-Leibler divergence d^2_{KL} is not a semimetric since it is not symmetric, however we will fix that up in Section 3.4. If a set of histograms is based on bins that are numeric and sortable (i.e. the bins can be ordered from least to greatest), then we can transform a normalized histogram $\{p_1, p_2, \ldots, p_c\}$ into a cumulative histogram $\{P_1, P_2, \ldots, P_c\}$ where $P_k = \sum_{j=1}^{k} p_j$. My thesis will mention three of the possible distance measures on cumulative histograms $\{P_1, \ldots, P_c\}$ and $\{Q_1, \ldots, Q_c\}$, namely:

$$\begin{aligned} d_{KS} &= \max_j |P_j - Q_j| \quad [\text{ Kolmogorov-Smirnov }] \\ d^2_{CM} &= \sum_{j=1}^{c-1} (P_j - Q_j)^2 \quad [\text{ Cramer - von Mises }] \\ d^2_{AD} &= 4 \sum_{j=1}^{c-1} \frac{(P_j - Q_j)^2}{(P_j + Q_j)[2 - (P_j + Q_j)]} \quad [\text{ Anderson-Darling }] \end{aligned} \tag{2.7}$$

Note that for the Anderson-Darling measure d^2_{AD}, we define as zero any term for which $P_j = Q_j = 0$, or $P_j = Q_j = 1$. For more information on distance measures for empirical cumulative functions and their distributions, see [42], [3], and [43].

The distance measures for histograms mentioned in equations 2.6 and 2.7 all require that the two histograms being compared have the same number of bins. If they do not, we have a choice: we can either perform bin splitting or bin combination to equalize the number of bins, or we can use an earth mover distance that does not require equal histogram sizes [95], [97]. The earth mover distance is the solution of a transportation problem that minimizes the "cost" of transforming one normalized histogram into another. In particular, if the source histogram is $\{x_1, \ldots, x_c\}$ and the target histogram is $\{y_1, \ldots, y_d\}$, and the ground distance (or "cost") between source bin i and target bin j is c_{ij}, then the earth mover distance is the target function optimum in

$$\begin{aligned} \text{[EMD] Minimize } T = \sum_{i=1}^{c} \sum_{j=1}^{d} c_{ij} f_{ij} \quad \text{subject to} \\ \sum_{j=1}^{d} f_{ij} = x_i, \sum_{i=1}^{c} f_{ij} = y_j, f_{ij} \geq 0 \end{aligned} \tag{2.8}$$

The ground distance c_{ij} is what determines the nature of the optimum flow $\{f_{ij}\}$ (see Figure 2.7). I treat the earth mover distance in more detail in section 3.5, and a brief survey of its use in pattern classification is in [34].

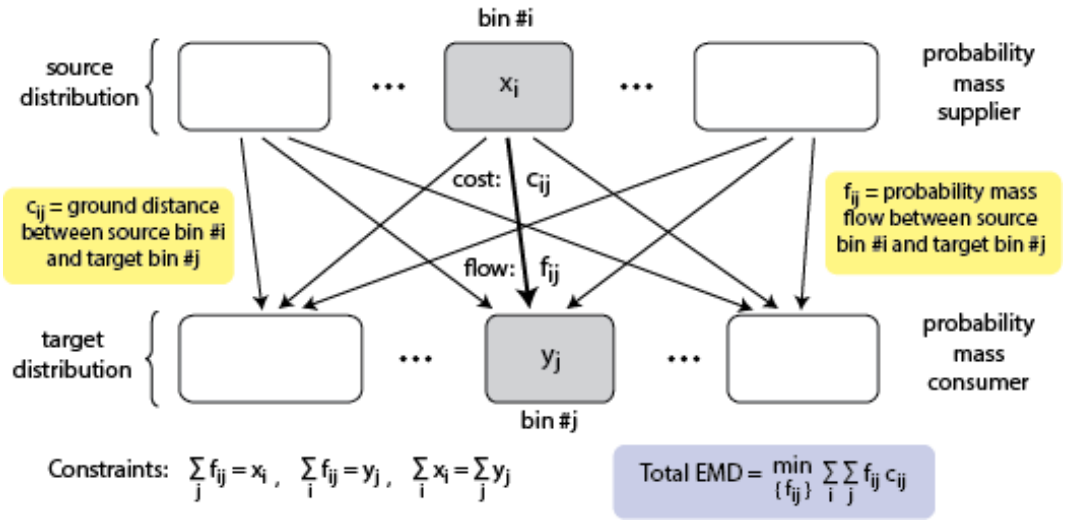


Figure 2.7: Earth mover distance between different histograms

2.4 ARMA time series

For our purposes, a time series is a sequence of random real variates that have a strict equidistant temporal ordering. Such series may be used to model records of natural phenomena such as rainfall and temperature, or man-made processes such as security prices and foot traffic [14]. I assume that such a series has been processed to remove non-stationary trends such as seasonality, exponential growth, or a deterministic polynomial trend. The series itself is observed noiselessly, but it depends on a parallel series of independent, identically distributed (IID) random innovations that is unobserved. We will be interested in stationary, invertible autoregressive moving average (ARMA) processes [15] [18]. More specifically, if μ is the long-term average of the series, then the defining equation is

$$u_t - a_1 u_{t-1} \ldots - a_p u_{t-p} = e_t - c_1 e_{t-1} \ldots - c_q e_{t-q} \tag{2.9}$$

where $u_t = y_t - \mu$, $\{y_t\}$ is the observed series, and $e_t \sim IID(0, v_e)$, meaning that the innovations $\{e_t\}$ are independent, are identically distributed, have zero mean, and have variance v_e. The integer pair (p, q) is the order of the ARMA model, and the innovations $\{e_t\}$ provide the randomness. A complete ARMA model is a 4-tuple $\{[a_1 \ldots a_p], [c_1 \ldots c_q], v_e, \mu\}$. The ARMA roots are defined by the factorizations

$$\begin{aligned}(1 - a_1 B - a_2 B^2 \ldots - a_p B^p) &= (1 - r_1 B)(1 - r_2 B) \ldots (1 - r_p B) \\ (1 - c_1 B - c_2 B^2 \ldots - c_q B^q) &= (1 - s_1 B)(1 - s_2 B) \ldots (1 - s_q B)\end{aligned} \tag{2.10}$$

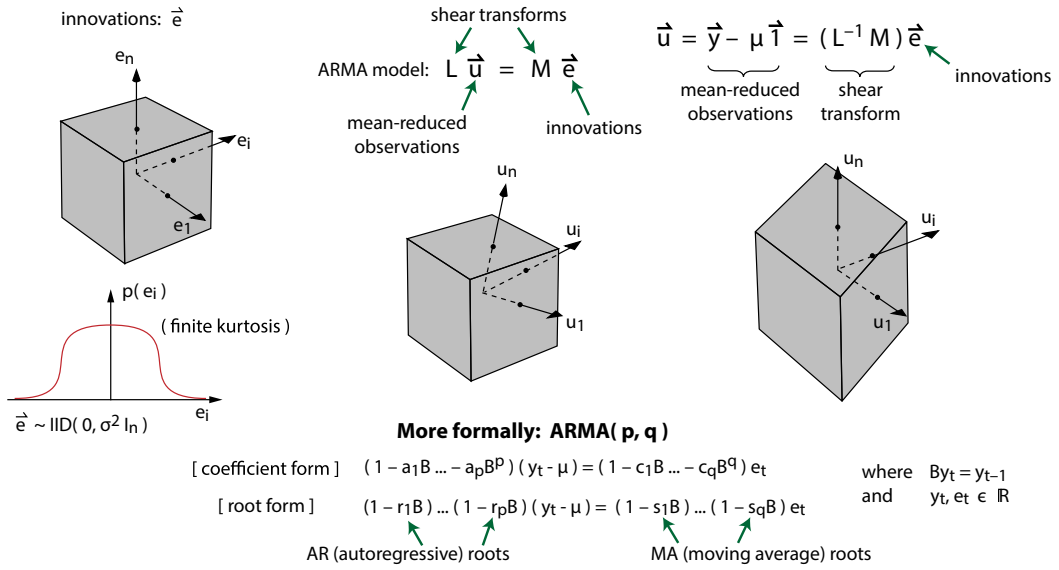


Figure 2.8: Stationary ARMA model for a time series, showing how the mean-reduced observations are a shear transform applied to the innovations.

In order for the process to be stationary and invertible, we require that the AR roots $\{r_1 \ldots r_p\}$ and MA roots $\{s_1 \ldots s_q\}$ have magnitude less than one. In order for ARMA models to be unique, we must also exclude any root that occurs as both an AR root and an MA root. The basic idea is that we apply a shear transform to the innovations $\vec{e} = [e_1, \ldots, e_{n-1}\, e_n]$ to get the mean-reduced observations $\vec{u} = [u_1, \ldots, u_{n-1}\, u_n]$ (see Figure 2.8).

In this thesis, I use the Whittle estimator to compute maximum likelihood estimates for ARMA time series models. This estimator is fully described in [113]. In particular, the maximum likelihood estimates of the ARMA parameters are those that minimize the

residual sample innovation variance:

$$
\begin{aligned}
[\text{ ARMA }]\ \text{Minimize } T &= \int_0^{2\pi} \frac{\hat{I}(\omega)}{PSD(\omega)} \frac{\mathrm{d}\omega}{2\pi} \quad \text{where} \\
\hat{I}(\omega) &= \sum_{-n+1}^{n-1} \hat{\gamma}(h) e^{-jh\omega} \quad (\text{ sample periodogram }) \\
\hat{\gamma}(h) &= \frac{1}{n} \sum_{j=1}^{n-h} u_j u_{j+h} \quad (\text{ sample autocovariance }) \\
PSD(\omega) &= \frac{g(s_1,\omega)\ldots g(s_q,\omega)}{g(r_1,\omega)\ldots g(r_p,\omega)} \quad (\text{ power spectral density }) \\
g(a,\omega) &= 1 + a^2 - 2a\cos\omega \quad [\omega = \text{ discrete frequency }]
\end{aligned}
\tag{2.11}
$$

Here, the mean-reduced sample is $\{u_1, \ldots, u_n\}$, and n is the sample size. The Whittle estimator assumes that the ARMA model size (p, q) is already known. In this thesis, I do not address the issue of how to choose the model size. That is generally done with the aid of the Akaike or Bayesian information criteria, and further information about this issue is in [60], [59], and [53].

The ARMA process that equation 2.9 describes has a constant variance, and there are many generalizations of that model that allow the series variance to fluctuate [63] [104]. In this thesis, I will make use of one such generalization, namely the generalized autoregressive conditionally heteroskedastic model of order (p, q), abbreviated as GARCH(p, q) [12] [99]. In this model, the innovation variance itself develops as an ARMA(p, q) process. The simplest such model would be:

$$
\begin{aligned}
x_t &= \sigma_t z_t \qquad z_t \sim IID(0,1) \\
\sigma_t^2 &= w + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2
\end{aligned}
\tag{2.12}
$$

where $\{x_t\}$ is an observed, zero-mean time series, $\{z_t\}$ are the innovations, and the model parameters are $\{[\alpha_1 \ldots \alpha_p], [\beta_1 \ldots \beta_q], w\}$. I describe the GARCH(1,1) process in more detail in section 4.5.

We are interested in distance measures between ARMA models; for distance measures between the datasets themselves, see [73] and [35]. As for the ARMA models, we have a choice: we can compare their parameter sets, or their power spectral densities (see equation

2.11). Piccolo [92] suggested a metric of the form

$$d^2_{PIC} = \sum_{j=1}^{\infty} (\pi_{1,j} - \pi_{2,j})^2 \tag{2.13}$$

between the AR(∞) representations of the two ARMA models [25]. The AR(∞) representation of an ARMA model $\{[a_1, \ldots, a_p], [c_1, \ldots, c_q], v_e\}$ is the solution of

$$1 - \pi_1 B - \pi_2 B^2 - \ldots = (1 - a_1 B \ldots - a_p B^p)(1 - c_1 B \ldots - c_q B^q)^{-1} \tag{2.14}$$

and will exist as long as the ARMA model is invertible. Otranto [88] extended that idea to GARCH(1,1) processes. Caiado et al. [19] proposed several distance measures based on the sample periodograms:

$$\begin{aligned} d^2_{NP} &= \sum_{j=1}^{[n/2]} \left[\frac{\hat{I}_1(\omega_j)}{\hat{\gamma}_{1,0}} - \frac{\hat{I}_2(\omega_j)}{\hat{\gamma}_{2,0}} \right]^2 \quad [\omega_j = 2\pi(j/n)] \\ d^2_{LNP} &= \sum_{j=1}^{[n/2]} \left[\log \frac{\hat{I}_1(\omega_j)/\hat{\gamma}_{1,0}}{\hat{I}_2(\omega_j)/\hat{\gamma}_{2,0}} \right]^2 \\ d^2_{KLFD} &= \sum_{j=1}^{[n/2]} \left[\frac{\hat{I}_1(\omega_j)/\hat{\gamma}_{1,0}}{\hat{I}_2(\omega_j)/\hat{\gamma}_{2,0}} - \log \frac{\hat{I}_1(\omega_j)/\hat{\gamma}_{1,0}}{\hat{I}_2(\omega_j)/\hat{\gamma}_{2,0}} - 1 \right] \end{aligned} \tag{2.15}$$

where $\hat{I}_k$ is the sample periodogram for dataset k, $\hat{\gamma}_{k,0}$ is the lag-zero sample autocovariance for that dataset, $\omega_j = 2\pi j/n$ is the discrete frequency, and the two datasets have the same size n. The sample periodogram is also calculable as the squared amplitude of the Fourier transform of the observed series [16]. Comparing two time series models by comparing their spectral densities makes a lot of sense: all time series models (ARMA or otherwise) have a corresponding spectral density.

Kalpakis et al. [66] proposed the Euclidean distance between time series model cepstra as an appropriate distance measure for time series models. The cepstrum of a time series is the inverse Fourier transform of the logarithm of the Fourier transform of the time series [86]. In particular, if $PSD_k(\omega)$ is the spectral density of model k, then its cepstral coefficients are defined by

$$\log PSD_k(\omega) = \sum_{n \in Z} c_{k,n} e^{jn\omega} \tag{2.16}$$

For further details on the cepstral coefficients, see [80]. We then have

$$\int_0^{\pi} |\log \frac{PSD_1(\omega)}{PSD_2(\omega)}|^2 \frac{\mathrm{d}\omega}{2\pi} = \sum_{n=1}^{\infty} |c_{1,n} - c_{2,n}|^2 \tag{2.17}$$

which is the metric that Kalpakis proposed; note that the absolute values in equation 2.17 refer to complex magnitude. Martin proposed a similar metric based on cepstral coefficients [80]:

$$d_M^2 = \int_0^\pi \left| \left(\frac{\mathrm{d}}{\mathrm{d}\omega} \right)^{1/2} \log \frac{PSD_1(\omega)}{PSD_2(\omega)} \right|^2 \frac{\mathrm{d}\omega}{2\pi} = \sum_{n=1}^{\infty} n |c_{1,n} - c_{2,n}|^2 \tag{2.18}$$

where $(\mathrm{d}/\mathrm{d}\omega)^{1/2}$ represents a fractional derivative.

2.5 Hidden Markov models

A time series is a sequence of random variates with serial correlation. The output of a hidden Markov model (HMM) is also a sequence of variates with serial correlation, but there's a wrinkle: the variates depend on a hidden state. Such models are commonly used in speech recognition, gesture analysis, bioinformatics, and control theory [94] [48] [69] [47]. I will be looking exclusively at discrete HMMs, in which the observable output is a sequence of symbols chosen from a finite set. Such a model has a set of states $\{\omega_1 \dots \omega_n\}$ which are not observable, and a set of symbols $\{v_1 \dots v_m\}$ which are unambiguously observable. While in state ω_i, the system will emit a symbol v_k with symbol emission probability $[B]_{ik} = b_{ik}$, then jump to state ω_j with state transition probability $[A]_{ij} = a_{ij}$. The state transition probability matrix A and emission probability matrix B are stochastic, i.e. $A\vec{1}_n^T = \vec{1}_n^T$ and $B\vec{1}_m^T = \vec{1}_m^T$ where $\vec{1}_n$ is the all-ones row vector of length n. The system's output is an infinite string $\{\dots v_{j_0} v_{j_1} v_{j_2} \dots v_{j_M} v_{j_{M+1}} \dots\}$ of which we observe a sample $\{v_{j_1} \dots v_{j_M}\}$ (see figure 5.1). For more information on the information theoretic properties of HMMs, consult [48] and [121].

We will be particularly interested in the Baum-Welch algorithm, a recursive procedure that takes the HMM model size (n, m) and output string $\{v_{j_1} \dots v_{j_M}\}$, and computes a maximum-likelihood estimate of the model parameters $\{\hat{A}, \hat{B}\}$ [7] [40]. Levinson et al. [72] showed that the algorithm does indeed yield maximum likelihood estimates. The algorithm works as follows: first, let $v(t)$ denote the symbol observed at time t. The observed sequence is then $\{v(1) \dots v(M)\}$. Using the current estimates $(\hat{A}_s, \hat{B}_s)$, calculate the following:

$\alpha(t-1, i)$ = probability of being in state ω_i at time $t-1$, having generated the symbols $\{v(1) \dots v(t-1)\}$ [forward probability]

$\beta(t, j)$ = probability of being in state ω_j at time t, and generating the symbols $\{v(t+1) \dots v(M)\}$ afterwards [backward probability]

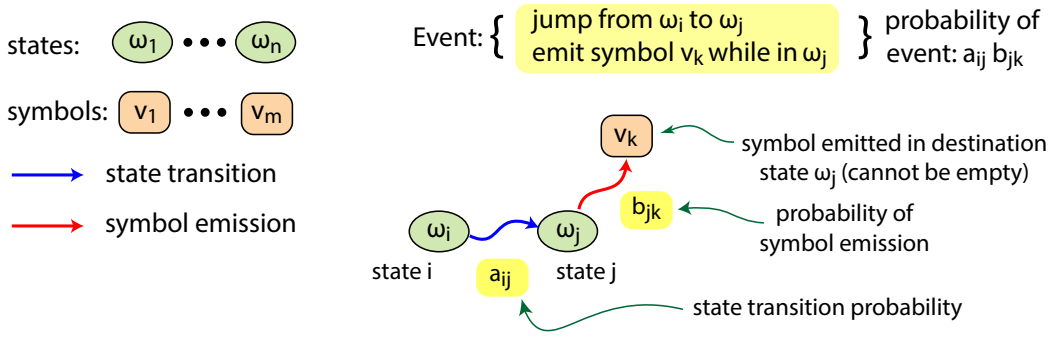


Figure 2.9: States, symbols, and transitions for a discrete hidden Markov model

$\gamma(i,j,t) = \alpha(t-1,i)[\hat{A}_s]_{ij}[\hat{B}_s]_{jk}\beta(t,j)$ = probability of being in state ω_i at time $t-1$, in state ω_j at time t, and generating the sequence $\{v(1)\dots v(M)\}$ where $v(t) = v_k$ [conditional transition probability]

In chapter 5, I show how to calculate the forward and backward probabilities. The next Baum-Welch iteration is

$$
\begin{aligned}
[\hat{A}_{s+1}]_{ij} &= \sum_{t=1}^{M} \gamma(i,j,t) / \sum_{t=1}^{M}\sum_{j=1}^{n} \gamma(i,j,t) \\
[\hat{B}_{s+1}]_{jk} &= \sum_{v(t)=v_k} \sum_{i=1}^{n} \gamma(i,j,t) / \sum_{t=1}^{M}\sum_{i=1}^{n} \gamma(i,j,t)
\end{aligned}
\tag{2.19}
$$

The algorithm is an example of an expectation maximization algorithm, and has linear convergence [40] [62]. See figure 2.10. The Baum-Welch algorithm is also known as the forward-backward algorithm. It has been extended to stochastic context-free grammars, where it is known as the inside-outside algorithm [5] [71] [57].

We are interested in distance measures between HMM models, where a model $\{A, B\}$ consists of the state transition and symbol emission probability matrices. For distance measures between the output strings themselves, see [70]. Levinson et al. [72] proposed

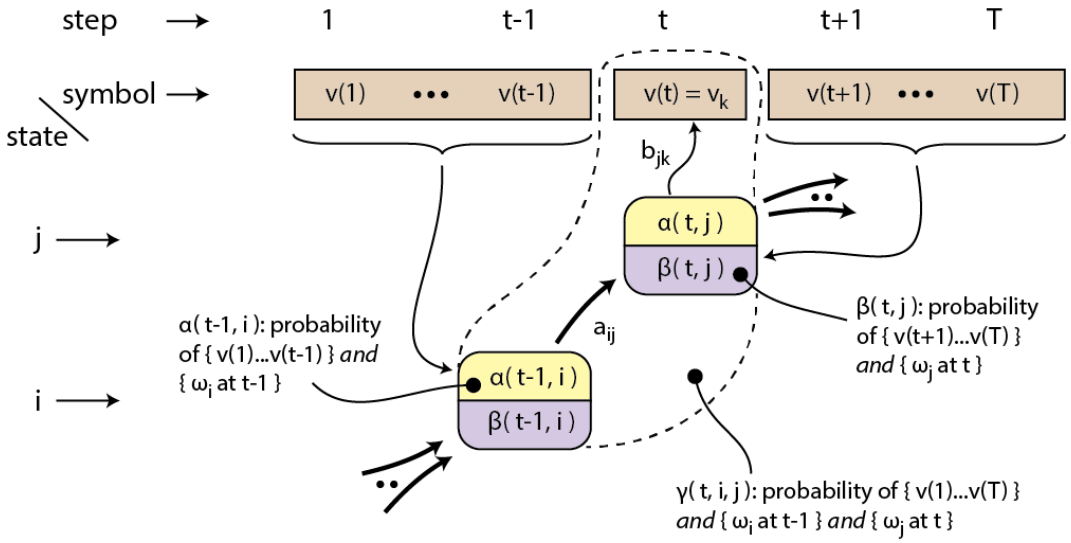


Figure 2.10: Forward, backward, and cell transition probabilities in the Baum-Welch algorithm

the first distance between HMM models, namely

$$d_B^2 = \frac{1}{nm} \sum_{j=1}^{n} \sum_{k=1}^{m} ([B_1]_{jk} - [B_2]_{\sigma(j)k})^2 \tag{2.20}$$

where the permutation on states σ is chosen so as to minimize d_B^2. In fairness, we should note that the purpose of this measure was to choose the best mapping of states between two HMM models, not to measure model similarity. Juang and Rabiner [64] proposed the following divergence measure for HMM models:

$$H(\theta_0, \theta) = \lim_{M \to \infty} \frac{1}{M} [\log p(j_1 \dots j_M | \theta_0) - \log p(j_1 \dots j_M | \theta)] \tag{2.21}$$

where $\{v_{j_1} v_{j_2} \dots\}$ is an infinitely long string generated by the true model $\theta_0 = \{A_0, B_0\}$, and $p(j_1 \dots j_M | \theta)$ is the probability of observing the finite sequence $\{v_{j_1} \dots v_{j_M}\}$ under model $\theta = \{A, B\}$. The limit $M \to \infty$ is important here; for finite M, the likelihood $\mathcal{L} = -\log p(j_1 \dots j_M | \theta)$ will achieve its minimum not at θ_0 but at some nearby model $\hat{\theta}_M$, in which case $H(\theta_0, \theta)$ may go negative (see figure 2.11). $H(\theta_0, \theta)$ is a distance function and should never be negative.

Equation 2.21 is clearly a divergence based on the marginal entropy of the true model, and for that reason I refer to the quantity as a marginal divergence. By marginal entropy,

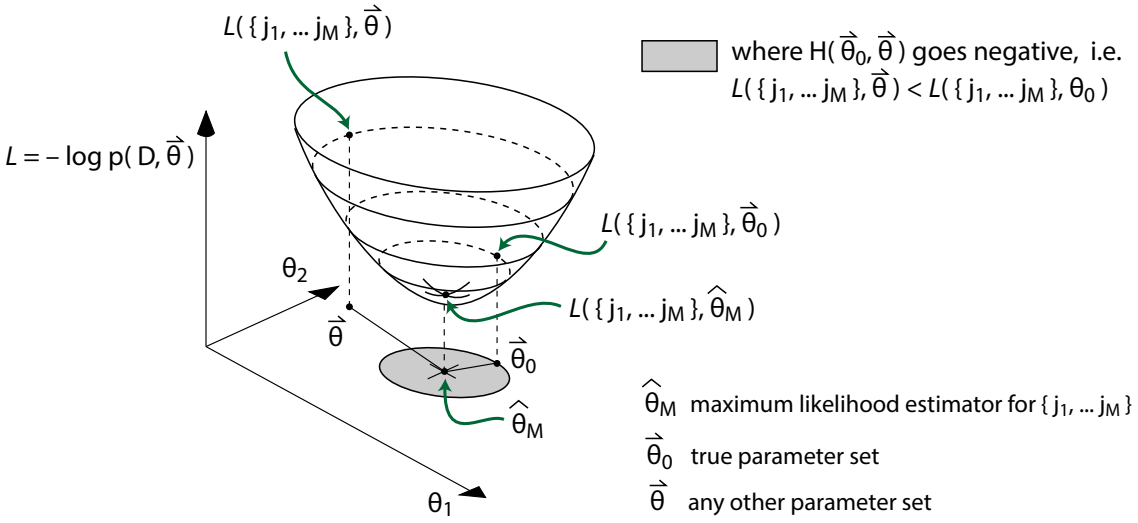


Figure 2.11: Divergence for finite output strings may go negative, in violation of the requirements for a distance function

I mean the average surprisal per symbol in the limit of an infinite output sequence. Hence, it makes sense to look for an approximation to an HMM's per-symbol entropy. Falkhausen et al. [51] proposed one such approximation, and used it to form the marginal divergence measure

$$d^2_{VIT} = \sum_{i=1}^{n} a_i \left[\sum_{j=1}^{n} a_{ij} \log \frac{a_{ij}}{\hat{a}_{ij}} + \sum_{k=1}^{m} b_{ik} \log \frac{b_{ik}}{\hat{b}_{ik}} \right] \tag{2.22}$$

where the base model is $\{A, B\}$ with $a_{ij} = [A]_{ij}$ and $b_{ik} = [B]_{ik}$, the target model is $\{\hat{A}, \hat{B}\}$, and $\{a_i\}$ are the components of the stationary distribution satisfying $\vec{a}\, A = \vec{a}$ (more about that in chapter 5). Equation 2.22 is actually an upper bound on the marginal divergence, as was shown by [36] and [118]. Silva and Narayanan [103] extended this marginal divergence idea to left-to-right HMMs, which are not ergodic but instead have definite initial and final states.

Discrete hidden Markov models have the same structure as probabilistic regular grammars with no final probabilities (i.e. no final states) [91] [41]. Thus, some distance measures for probabilistic regular grammars will be applicable to HMMs; we just have to take into account that discrete HMMs generate words of infinite length. A probabilistic regular

grammar with final probabilities will assign a probability to finite words (a word being any sequence of allowable symbols). For HMMs, we can use instead

$$\begin{aligned} p(u,\theta) &= p(j_1 \dots j_N, \theta) = p(\{v_{j_1} \dots v_{j_N}\}, \theta) \\ &= \text{probability of generating an infinite word with prefix} \\ &\quad \{v_{j_1} \dots v_{j_N}\} \text{ and model } \theta = \{A, B\} \end{aligned} \tag{2.23}$$

for finite words $u = \{v_{j_1} \dots v_{j_N}\}$ where $\sum_{|u|=N} p(u,\theta) = 1$ [41]. There are two obvious distance functions that we can define on words of finite size:

$$\begin{aligned} d_L^k &= \sum_{|u|=N} |p_1(u) - p_2(u)|^k \quad (L_k \text{ distance }) \\ d_{KL}^2 &= \sum_{|u|=N} p_1(u) \log \frac{p_1(u)}{p_2(u)} \quad (\text{ Kullback-Leibler }) \end{aligned} \tag{2.24}$$

where $p_1(u)$ and $p_2(u)$ are the probabilities that the two models assign to the prefix u, and N is the length of the prefix u. Note that the Kullback-Leibler divergence is not symmetric in p_1 and p_2; we will fix that up in Section 3.4. Chen and Kiefer [22] looked at the L_1 distance for probabilistic regular grammars, and Cortes et al. [26] [27] looked at both the L_k distances for even k, and the Kullback-Leibler distance for probabilistic regular grammars. In equation 2.24, note that the per-symbol limit of d_{KL}^2 as the prefix size increases indefinitely is just what Juang and Rabiner recommended as a general case (see equation 2.21). Assuming that large words are approximately independent, we have

$$\begin{aligned} H(\theta_0, \theta) &= \lim_{M\to\infty} \frac{1}{M} [\log p(j_1 \dots j_M|\theta_0) - \log p(j_1 \dots j_M|\theta)] \\ &\approx \lim_{N\to\infty} \frac{1}{N} \sum_{|u|=N} p(u,\theta_0)[\log p(u,\theta_0) - \log p(u,\theta)] \\ &\approx \lim_{N\to\infty} \frac{1}{N} \sum_{|u|=N} p(u,\theta_0) \log \frac{p(u,\theta_0)}{p(u,\theta)} \end{aligned} \tag{2.25}$$

Note also that defining a distance measure based on word probabilities has the advantage that we do not need to find a state mapping between the two HMMs, in fact they may have different numbers of states.

Chapter 3

Histograms

3.1 The statistics of histograms

The purpose of this chapter is to derive the information matrix for histograms prepared from large samples, and illustrate how L_2 distance functions based on that information matrix have the lowest type II error for cluster separation. In sections 3.6 and 3.7 I look at how to choose the number of bins, and in section 3.8 I look at alternate distance functions for datasets of numeric sortable variates, in particular those involving the empirical cumulative distribution function, and the sample cumulants.

The statistics of histograms are those of counting. Suppose that we have a total population of N events, each of which may be classified unambiguously into one of c classes labelled $\{1, 2, \ldots c\}$. This population has class sizes $\vec{N} = \{N_1, N_2, \ldots N_c\}$ where $\sum_{j=1}^{c} N_j = N$. Suppose further that we draw a sample of n events from this parent population, without replacement, and independently (i.e. without censoring). Then the probability of drawing a sample with composition $\vec{n} = \{n_1, n_2, \ldots n_c\}$ where n_j is the number of sample events in class j, and $\sum_{j=1}^{c} n_j = n$, is given by the discrete hypergeometric probability [115]

$$h(\vec{n}, \vec{N}) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_c}{n_c}}{\binom{N}{n}} \tag{3.1}$$

I assume here that draws are independent. If we assume further that $N_j \gg n_j$ for all classes, or equivalently that we draw events with replacement, then the discrete hyperge-

ometric probability becomes a discrete multinomial probability:

$$\begin{aligned} f(\vec{n},\vec{N}) &\approx \frac{n!}{N!}\prod_{j=1}^{c}\frac{N_j^{n_j}}{n_j!} \approx \frac{n!}{n_1!\dots n_c!}\left(\frac{N_1}{N}\right)^{n_1}\dots\left(\frac{N_c}{N}\right)^{n_c} \\ f(\vec{n},\vec{q}) &= \frac{n!}{n_1!\dots n_c!}q_1^{n_1}\dots q_c^{n_c} \end{aligned} \tag{3.2}$$

where $q_i = N_i/N$, $\sum_{i=1}^{c} q_i = 1$, and $\sum_{i=1}^{c} n_i = n$.

The characteristic function of that distribution is the expected value of $\exp(i\vec{t}\cdot\vec{n})$ over all possible values of $\vec{n}$:

$$\begin{aligned} \phi(\vec{t},\vec{n}) &= \langle\exp(i\vec{t}\cdot\vec{n})\rangle \text{ over } \vec{n} \text{ where } \sum_{i=1}^{c} n_i = n \\ &= \sum_{\vec{n}}\frac{n!}{n_1!\dots n_c!}\left(q_1e^{it_1}\right)^{n_1}\dots\left(q_ce^{it_c}\right)^{n_c} = \left(\sum_{j=1}^{c} q_je^{it_j}\right)^{n} \end{aligned} \tag{3.3}$$

and from that characteristic function, we may deduce the well-known mean and variance of the observed composition $\vec{n}$ [116]:

$$\begin{aligned} \langle n_i\rangle &= \left[-i\frac{\partial}{\partial t_i}\phi(\vec{t},\vec{n})\right]_{\vec{t}=0} = nq_i \\ \operatorname{var}(n_i,n_j) &= \left[-\frac{\partial^2}{\partial t_i\partial t_j}\phi(\vec{t},\vec{n})\right]_{\vec{t}=0} - \langle n_i\rangle\langle n_j\rangle = n(q_i\delta_{ij} - q_iq_j) \end{aligned} \tag{3.4}$$

Taking $\hat{q} = \vec{n}/n$ as an estimator for $\vec{q}$, we obtain:

$$\langle\hat{q}\rangle = \vec{q}, \quad \operatorname{var}(\hat{q}) = \frac{1}{n}(\operatorname{diag}(\vec{q}) - \vec{q}\,\vec{q}^{\,T}) \tag{3.5}$$

where the expectation is over all possible samples of size n.

That variance is singular since not all the $\{\hat{q}_i\}$ are independent. So let's define an estimator $\hat{p}$ as follows:

$$\begin{aligned} \vec{p} &= \{q_1,\dots,q_{c-1}\} \quad \text{with } p_c = q_c = 1 - \sum_{i=1}^{c-1} p_i \\ \hat{p} &= \{\hat{q}_1,\dots,\hat{q}_{c-1}\} = \{\frac{n_1}{n},\dots,\frac{n_{c-1}}{n}\} \quad \text{with } \hat{p}_c = 1 - \sum_{i=1}^{c-1}\hat{p}_i \end{aligned} \tag{3.6}$$

The estimator $\hat{p}$ is unconstrained except by virtue of the requirements that $\sum_{i=1}^{c} n_i = n$ and $n_i \geq 0$. Its variance and precision matrices are:

$$\begin{aligned} \text{var}(\hat{p}) &= \frac{1}{n}\left(\text{diag}(\vec{p}) - \vec{p}\vec{p}^{\,T}\right) = \frac{1}{n}R(\vec{p}) \\ \text{prec}(\hat{p}) &= \text{var}^{-1}(\hat{p}) = n\left[\text{diag}(\frac{1}{p_1}, \ldots, \frac{1}{p_{c-1}}) + \frac{1}{p_c}J_{c-1}\right] = nS(\vec{p}) \end{aligned} \tag{3.7}$$

where J_m is the all-ones matrix of size m by m.

Let's work out the Fisher information matrix for the discrete multinomial distribution. In section 3.2 we'll examine the Gaussian approximation to the multinomial distribution, and in section 3.3 we'll relate Fisher information to distance measures. The discrete multinomial distribution satisfies the Cramer-Rao regularity condition

$$\langle \frac{\partial}{\partial p_i} \log f(\vec{n}, \vec{p}) \rangle = \langle \frac{n_i}{p_i} - \frac{n_c}{p_c} \rangle = \frac{np_i}{p_i} - \frac{np_c}{p_c} = 0 \tag{3.8}$$

and hence the smallest possible covariance of an unbiased estimator $\hat{p}$ for $\vec{p}$ is $I^{-1}(\vec{p})$ where

$$[I(\vec{p})]_{ij} = -\langle \frac{\partial^2}{\partial p_i \partial p_j} \log f(\vec{n}, \vec{p}) \rangle = n\left(\frac{1}{p_i}\delta_{ij} + \frac{1}{p_c}\right) = [nS(\vec{p})]_{ij} \tag{3.9}$$

See [67] for more information about the Cramer-Rao bound. In particular, if $C_{\hat{p}}$ is the covariance matrix for an unbiased estimator $\hat{p}$, then

$$C_{\hat{p}} - I^{-1}(\vec{p}) \geq 0 \quad \text{i.e. is positive semidefinite} \tag{3.10}$$

$I(\vec{p})$ is the Fisher information matrix, i.e. the expected Hessian matrix of the total sample surprisal, where the second derivatives are with respect to the population model $\vec{p}$, and the expectation is with respect to all possible samples $\vec{n}$. I note here that the estimator $\hat{p}$ of equation 3.6 achieves the Cramer-Rao lower bound implied by equation 3.9.

3.2 Gaussian approximation to the multinomial distribution

Now we will consider large samples and show that the discrete multinomial distribution becomes approximately Gaussian in this case. Specifically, assume that $np_i \gg 1$ for $i = 1, \ldots c$ and also that

$$\frac{\triangle p_i}{p_i} = \frac{(\hat{p}_i - p_i)}{p_i} \ll 1 \text{ for } i = 1, \ldots, c \tag{3.11}$$

With these assumptions, and setting $\triangle\vec{p} = \{\hat{p}_1 - p_1, \ldots, \hat{p}_{c-1} - p_{c-1}\}$, we obtain:

$$f(\triangle\vec{p}, \vec{p}) \approx \frac{1}{(\sqrt{2\pi})^{c-1}} \frac{(\sqrt{n})^{c-1}}{\sqrt{p_1 p_2 \ldots p_c}} \exp\left[-\frac{1}{2}(\triangle\vec{p})^T n S(\vec{p}) \triangle\vec{p}\right] \frac{1}{n^{c-1}} \tag{3.12}$$

Since n^{c-1} is the density of possible estimates in $\triangle\vec{p}$ space and is large by virtue of our assumption that $n \gg 1$, we could consider $\triangle\vec{p}$ to be a continuous variable with density

$$g(\triangle\vec{p}, \vec{p})\mathrm{d}(\triangle\vec{p}) = \frac{1}{(2\pi)^{\frac{c-1}{2}}} \left[\frac{n^{c-1}}{p_1 p_2 \ldots p_c}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\triangle\vec{p})^T n S(\vec{p}) \triangle\vec{p}\right] \mathrm{d}(\triangle\vec{p}) \tag{3.13}$$

That is, the corresponding Gaussian density is

$$g(\triangle\vec{p}, \vec{p}) = \frac{1}{(2\pi)^{\frac{c-1}{2}}} \left[\det(nS(\vec{p}))\right]^{\frac{1}{2}} \exp\left[-\frac{1}{2}\triangle\vec{p}^T n S(\vec{p}) \triangle\vec{p}\right] \tag{3.14}$$

where

$$S(\vec{p}) = \mathrm{diag}(\frac{1}{p_1}, \ldots, \frac{1}{p_{c-1}}) + \frac{1}{p_c} J_{c-1} \tag{3.15}$$

The mean and variance are the same as for the discrete multinomial distribution, namely

$$\begin{aligned} \langle\triangle\vec{p}\rangle = 0 \Rightarrow \langle\hat{p}_i\rangle = p_i \quad (1 \le i \le c-1) \\ \mathrm{var}(\hat{p}_i, \hat{p}_j) = p_i\delta_{ij} - p_i p_j = R(\vec{p}) \quad (1 \le i \le c-1) \end{aligned} \tag{3.16}$$

but the Fisher information matrix is different. In particular, for the multivariate Gaussian density $g(\triangle\vec{p}, \vec{p})$,

$$\begin{aligned} [I(\vec{p})]_{kl} &= \langle \frac{\partial(\log g)}{\partial p_k} \frac{\partial(\log g)}{\partial p_l} \rangle \quad \text{(expectation over all } \triangle\vec{p}) \\ &= \frac{1}{2}\mathrm{tr}\left(R^{-1}\frac{\partial R}{\partial p_k} R^{-1} \frac{\partial R}{\partial p_l}\right) + nR_{kl}^{-1} \quad \text{where} R(\vec{p}) = \mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^T \end{aligned} \tag{3.17}$$

See [93] for a derivation of the Fisher information matrix for a multivariate Gaussian density. The Fisher information derived from the multivariate Gaussian approximation differs from the true value by a term of order $(1/n)$. For large sample size, we have

$$I(\vec{p}) \approx nR^{-1}(\vec{p}) = nS(\vec{p}) = n\left[\mathrm{diag}(\frac{1}{p_1}, \ldots, \frac{1}{p_{c-1}}) + \frac{1}{p_c} J_{c-1}\right] \tag{3.18}$$

which is the precision matrix for the estimator $\hat{p}$ of equation 3.6. That estimator is unbiased, and we know that the Fisher information matrix represents the largest possible

precision matrix for unbiased estimators of $\vec{p}$. So, for $n \gg 1$, the estimator $\vec{q} = \vec{n}/n$ has full first-order efficiency. Now, define a matrix L such that

$$L^T L = nS(\vec{p}) = n\left[\text{diag}(\frac{1}{p_1}, \ldots, \frac{1}{p_{c-1}}) + \frac{1}{p_c} J_{c-1}\right] \tag{3.19}$$

Then $L\triangle\vec{p}$ will be a $N(0, I_{c-1})$ random vector for $n \gg 1$, and hence $(\triangle\vec{p})^T(nS)(\triangle\vec{p})$ will be have an asymptotic chi-square$(c-1)$ distribution for $n \gg 1$. Note that

$$(\triangle\vec{p})^T(nS)(\triangle\vec{p}\,) = n\sum_{i=1}^{c}\frac{1}{p_i}(\triangle p_i)^2 = n\sum_{i=1}^{c}\frac{1}{p_i}(\hat{p}_i - p_i)^2 \tag{3.20}$$

which is already looking like a distance measure between $\hat{p}$ and $\vec{p}$. The next section will explore that idea further.

3.3 Relation between Fisher information and distance measures

Since the Fisher information matrix is the largest possible precision matrix for unbiased estimators, in the sense that $I(\vec{p}) - C^{-1}(\hat{p}) \geq 0$ for any unbiased estimator $\hat{p}$ with covariance $C(\hat{p})$, we might expect that for the domain of histograms, the quantity

$$d(\hat{p}, \vec{p}) = (\triangle\vec{p})^T(nS)(\triangle\vec{p}) = n\sum_{i=1}^{c}\frac{1}{p_i}(\triangle p_i)^2 \tag{3.21}$$

is in some sense "optimum" for measuring how $\hat{p}$ differs from $\vec{p}$. Let's show that! We have the estimator $\hat{q} = \vec{n}/n$ which computes $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_{c-1}\}$ for a sample of size n and composition $(n_1, \ldots, n_c)$. Suppose, first, that we know $\langle\hat{p}\rangle = \vec{p}$. Then, if we choose a matrix L according to equation 3.19 above, then $L(\hat{p} - \vec{p})$ has covariance I_{c-1}. The quantities $z_i = [L(\hat{p} - \vec{p})]_i$ are $N(0,1)$ variates, and there are $m = c - 1$ of them. I assume here that the purpose of a distance measure between $\hat{p}$ and $\vec{p}$ is for classification. Specifically, we want to optimize the classification of estimates $\hat{p}$ so as to minimize the type II error, i.e. probability of classifying an estimate $\hat{p}$ as being "close" to $\vec{p}$ when in fact it is not (see Figure 3.1, and also the discussion in Section 2.2). Suppose that our distance measure is $f(\vec{z}) = f(z_1, \ldots, z_m)$. We'll normalize it so that $\langle f(\vec{z})\rangle = 1$, where the expectation is with respect to $\vec{z}$.

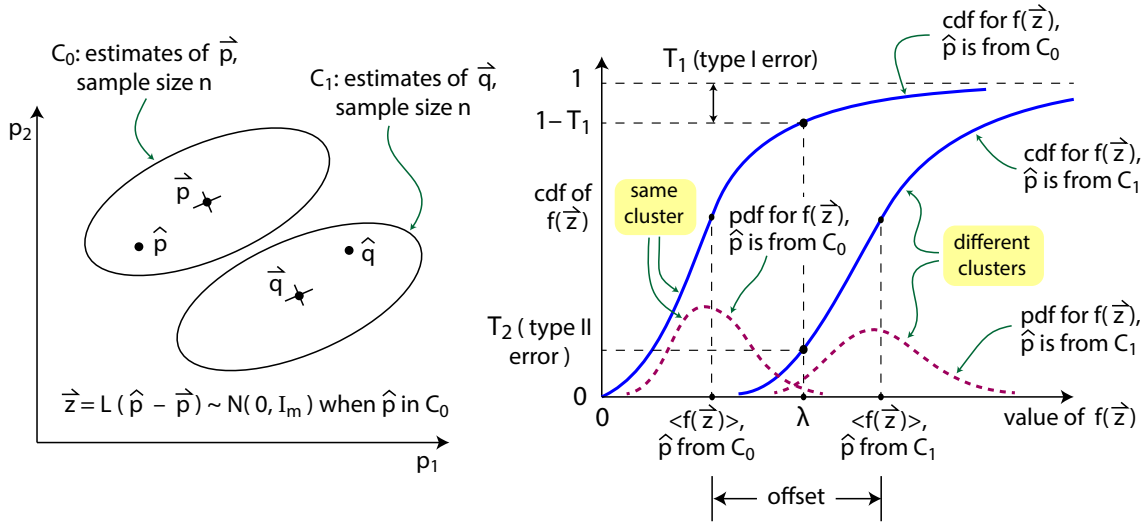


Figure 3.1: Illustration of type I threshold and type II error for an arbitrary distance function $f(\vec{z})$ on $\vec{z} \sim N(0, I_m)$. The null hypothesis is that histogram estimates belong to the same cluster.

We choose T_1, the type I error of saying that an estimate $\hat{p}$ is not derived from the model $\vec{p}$ when in fact it is, according to our classification needs. In this thesis, I use $T_1 = 0.05$ throughout. We want to minimize the type II error T_2 for a given T_1. In this system, the probability density is

$$w(\vec{z}) = \prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi}} \exp -(\frac{1}{2} z_i^2) \right] = \frac{1}{(\sqrt{2\pi})^m} \exp -(\frac{1}{2} |\vec{z}|^2) \tag{3.22}$$

So, our constraint that $\langle f(\vec{z}) \rangle = 1$ translates into

$$\langle f(\vec{z}) \rangle = \int f(\vec{z}) \frac{1}{(\sqrt{2\pi})^m} \exp -(\frac{1}{2} |\vec{z}|^2) \, \mathrm{d}^m \vec{z} = 1 \tag{3.23}$$

As a "first stab", let's look for a $f(\vec{z})$ that has minimum variance, given $\langle f(\vec{z}) \rangle = 1$. In other words:

$$\text{[MVD] Minimize } \langle f^2(\vec{z}) \rangle \text{ subject to } \langle f(\vec{z}) \rangle = 1 \text{ where } \vec{z} \sim N(0, I_m) \tag{3.24}$$

We will also require that $f(\vec{0}) = 0$ and that $f(\vec{z})$ be C^1 continuous. The weight function $w(\vec{z}) = \exp -\frac{1}{2}|\vec{z}|^2$ induces an orthonormal basis for $f(\vec{z})$, and since the weight function is pure radial, the basis functions will be products of radial functions and spherical harmonics for the sphere in m dimensions. Denote the resulting orthonormal basis functions as $h_{ij}(r)Y_{ij}(\vec{\theta})$, where $Y_{i0}(\vec{\theta}) = 1$ and $\vec{\theta}$ represents the $m-1$ angles of an m-dimensional spherical coordinate system [44]. Let $f_0(\vec{z})$ be the optimum we seek, with the expansion

$$f_0(\vec{z}) = \sum_{i \geq 0}^{\infty} \left[a_i h_{i0}(r) + \sum_{j \geq 1}^{\infty} b_{ij} h_{ij}(r) Y_{ij}(\vec{\theta}) \right] \quad (r = |\vec{z}|) \tag{3.25}$$

Since the $Y_{ij}(\vec{\theta})$ for $j \geq 1$ are orthogonal to any pure radial function, we have

$$\begin{aligned} \langle f_0(\vec{z}) \rangle &= \sum_{i \geq 0}^{\infty} a_i \langle h_{i0}(r) \rangle = 1 \\ \langle f_0^2(\vec{z}) \rangle &= \sum_{i \geq 0}^{\infty} \left[a_i^2 \langle h_{i0}^2(r) \rangle + \sum_{j \geq 1}^{\infty} b_{ij}^2 \langle h_{ij}^2(r) Y_{ij}(\vec{\theta}) \rangle \right] = \sum_{i \geq 0}^{\infty} \left[a_i^2 + \sum_{j \geq 1}^{\infty} b_{ij}^2 \right] \end{aligned} \tag{3.26}$$

Hence, we could achieve a smaller value of $\langle f^2(\vec{z}) \rangle$ by taking $f_0(\vec{z})$ as a purely radial function, i.e.

$$f_0(\vec{z}) = \sum_{i \geq 0}^{\infty} a_i h_{i0}(r) \tag{3.27}$$

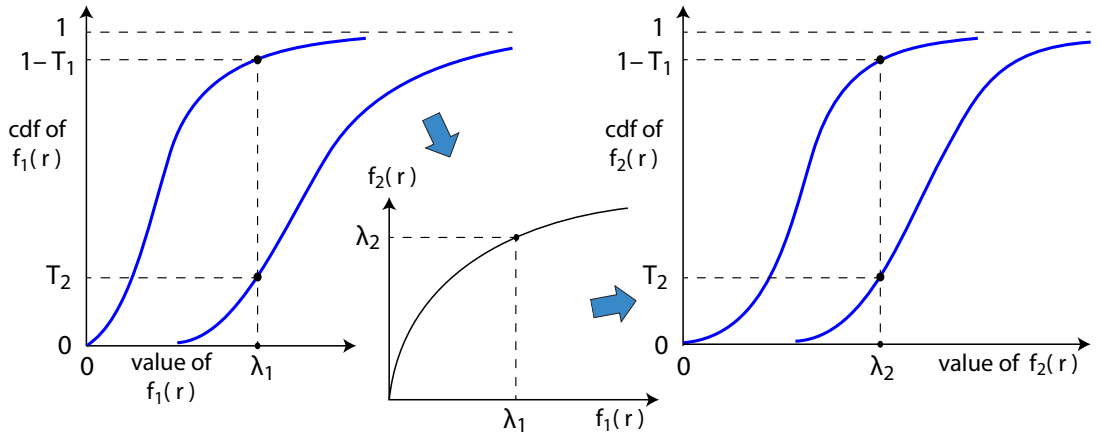


Figure 3.2: Correspondence between monotonic increasing functions of $|\vec{z}|$, showing that T_1 and T_2 do not change upon a monotonic transformation of a random variable

Which radial function to choose, though? Well, the absolute minimum of [MVD] is found at the indicator function $\{f_0(r) = 0 \text{ for } r = 0, f_0(r) = 1 \text{ for } r \neq 0\}$. But that is not practical for classification purposes. If we require that $f_0(\vec{z})$ be approximately scalable at $|\vec{z}| \approx 0$ (meaning that $f_0(a\vec{z}\,) = af_0(\vec{z}\,)$ for $a > 0$) then the only choice is the L_2 norm $f_0(r) = r$. By L_k norm, I mean

$$\| \vec{z} \|_k = \left[\sum_{i=1}^{m} |z_i|^k \right]^{1/k} \tag{3.28}$$

If we require only that $f_0(\vec{z})$ be a monotonically increasing function of $|\vec{z}|$ with $f_0(0) = 0$, and have two such functions $f_1(r)$ and $f_2(r)$, then there is a one-to-one correspondence between $\{r, f_1(r)\}$ and $\{r, f_2(r)\}$, which has the effect of warping the horizontal axis of Figure 3.1 while leaving the vertical axis intact (see Figure 3.2).

Hence, under this relaxed constraint, we can take $f_0(\vec{z})$ to be any convenient power of $|\vec{z}|$ other than a constant (since we need $f_0(0) = 0$). So let's take $f_0(\vec{z}) = \| \vec{z} \|_2 = |\vec{z}|$. As an illustration of the foregoing theory, which predicts that the L_2 distance has the lowest variance for a given mean, I made a numerical study in which I generate n samples of

Table 3.1: Variance and third central moment for $\| \vec{z} \|_k$ distributions with unit sample mean. Sample size is 8000, $m = 6$, and number of runs is 512. Both central moments are minimized for the L_2 distance.

k	m * variance		m^2 * third moment	
	mean	**std dev mean**	**mean**	**std dev mean**
0.50	0.694	0.001	0.745	0.005
1.00	0.571	0.001	0.430	0.003
1.50	0.528	0.001	0.317	0.003
2.00	0.519	0.001	0.292	0.003
2.50	0.524	0.001	0.308	0.003
3.00	0.535	0.001	0.343	0.003
3.50	0.547	0.001	0.384	0.003
4.00	0.558	0.001	0.422	0.003

$\| \vec{z} \|_k$ for $\vec{z} \sim N(0, I_m)$. For each such sample, I normalize to unit mean and compute the sample variance and third central moment. The results are in Table 3.1.

What Table 3.1 shows is that, as predicted, the $\| \vec{z} \|_k$ variance has a minimum at $k = 2$. The third central moment also appears to have a minimum at $k = 2$. Figure 3.3 shows, in chart format, the variance for L_k distributions of $\| \vec{z} \|_k$ for $\vec{z} \sim N(0, I_m)$ with unit mean. Note in particular that the variance reaches a minimum at $k = 2$ regardless of the number of dimensions m.

Now we need to show that for the L_k family of distance measures, there is a direct relationship between distance measure variance and type II error, as implied by Figures 2.4 and 3.1. I treat this subject in depth in Appendix G, and show that for a simple two-parameter approximation to the distribution function of an L_k distance measure, the directionally-averaged type II error is minimized at $k = 2$. I also verified this experimentally for representative values of m, k, and offset v^2. See Figure 3.4 for definitions of the offset vector $\vec{v}$ and type I threshold. Note that in Figure 3.4, a neighboring cluster is one where $(\triangle\vec{p})^T(nS)(\triangle\vec{p}) = O(1)$, where $\triangle\vec{p}$ is the difference in cluster centers, and S is defined by equation 3.18. It follows that $\triangle\vec{p}$ is of order $1/\sqrt{n}$. The cluster shape is a function of the cluster center, so the relative change in cluster shape is also of order $1/\sqrt{n}$. So, although in general the transform L defined in equation 3.19 will not diagonalize neighboring clusters exactly, it will nevertheless diagonalize them approximately to relative order $1/\sqrt{n}$, under the assumptions of Section 2.1.

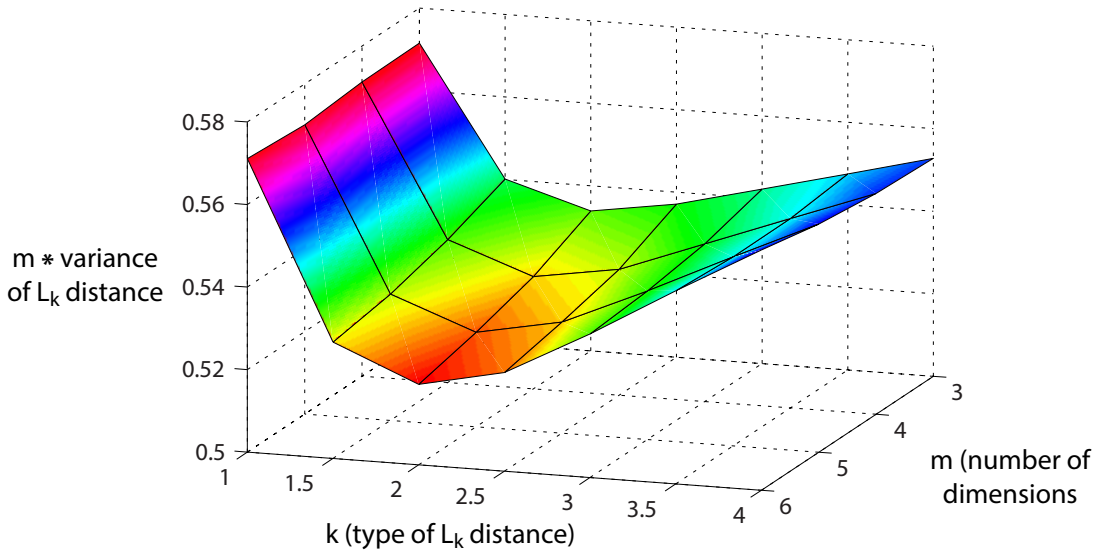


Figure 3.3: Variance of L_k distance measure when mean is held constant at one. The variance is minimized at $k = 2$.

Figure 3.5 shows the results of Appendix G (this is also Figure G.9 of Appendix G). For this figure, the experimental values of the type I threshold d_{95} are calculated by estimating the cdfs of 100 sets of 64000 randomly generated values of $d_0 = \sum_{i=1}^{m} |z_i|^k$ where all the $\{z_i\}$ are $N(0, 1)$. The T_2 values are calculated by estimating the cdfs of 400 sets of 64000 randomly generated values of $d_1 = \sum_{i=1}^{m} |z_i - v_i|^k$, where for each set the direction of $\vec{v}$ is chosen from a uniform distribution over all directions.

What Figure 3.5 shows is that the L_2 distance always has the lowest directionally-averaged type II error. Although not conclusive proof, that is strong evidence that the L_2 norm for $\vec{z} \sim N(0, I_m)$ minimizes the type II error for a specified type I error.

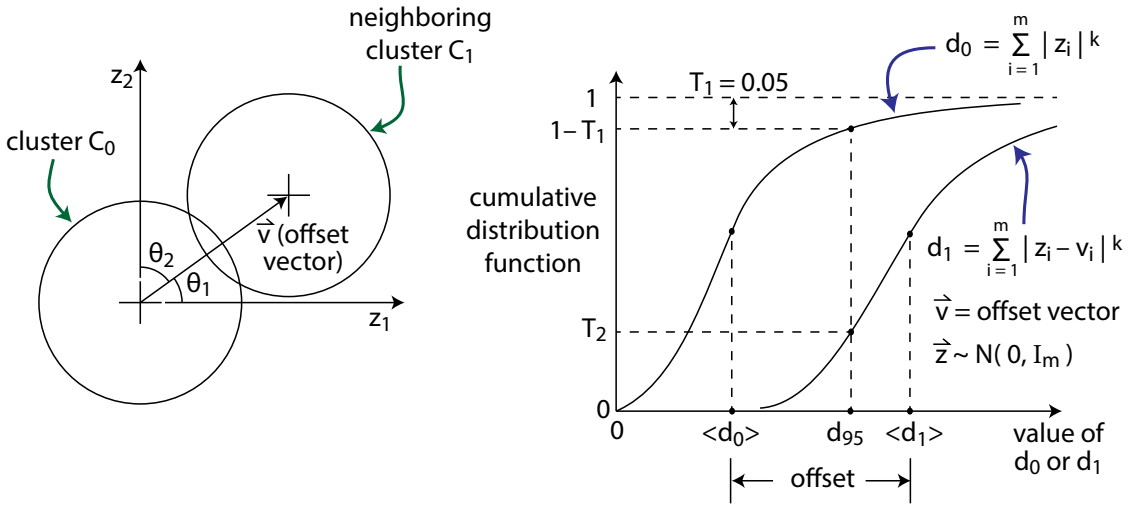


Figure 3.4: Definition of offset vector $\vec{v}$ and type I threshold d_{95} for L_k distance measures

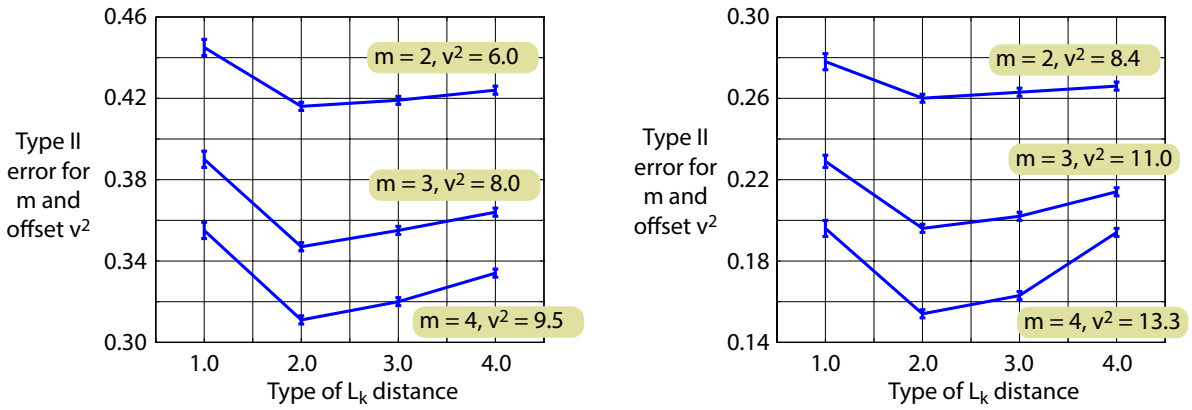


Figure 3.5: Empirical type II error rates for L_k distance measures, for representative values of m, k, and offset v^2. The error bars represent the 95% confidence limits.

3.4 Distance measures that approximate the Fisher information

What we have so far is that the quadratic form

$$d(\hat{p},\vec{p}) = (\triangle\vec{p})^T(nS)(\triangle\vec{p}) = n\sum_{i=1}^{c}\frac{1}{p_i}(\triangle p_i)^2 \tag{3.29}$$

where n is the sample size, and

$$S = \left[\text{diag}(\frac{1}{p_1},\ldots\frac{1}{p_{c-1}}) + \frac{1}{p_c}J_{c-1}\right] \tag{3.30}$$

appears to have the best classification accuracy, in the sense mentioned in the last section and illustrated in Figure 3.1. One problem, though, is that the measure becomes infinite when any of the p_i are zero. So, a reasonable question is: what functions $f(\vec{p},\hat{p},n)$ reduce to

$$f(\vec{p},\hat{p},n) \to n\sum_{i=1}^{c}\frac{1}{p_i}(\triangle p_i)^2 \text{ for } n\to\infty \text{ i.e. } \hat{p}\to\vec{p} \tag{3.31}$$

but which are more robust than equation 3.29 when any of the p_i are zero? Since we are interested in the behavior when $\hat{p}\approx\vec{p}$, we might start by looking at an f-divergence [6], say

$$\begin{gathered} d(\vec{p},\vec{q},n) = \sum_{i=1}^{c} f(\frac{p_i}{q_i})\,q_i \\ \text{where } f(1)=0,\ f''(1)>0, \quad \text{and } f(\frac{p_i}{q_i})\,q_i = 0 \text{ if } p_i = q_i = 0 \end{gathered} \tag{3.32}$$

I won't go into all the possible solutions of equation 3.32, since there are infinitely many. However, I will pursue the α-divergence family [23], which is an example of an f-divergence with

$$f(t) = \frac{4}{1-\alpha^2}\left[1 - t^{(1+\alpha)/2}\right] \tag{3.33}$$

which gives two well-known measures:

$$\begin{aligned} d_H^2(\vec{p},\vec{q},n) &= 8n\left[1-\sum_{i=1}^{c}\sqrt{p_i q_i}\right] \quad (\alpha=0) \\ d_{RE}^2(\vec{p},\vec{q},n) &= 2n\left[\sum_{i=1}^{c} p_i \log\frac{p_i}{q_i}\right] \quad (\alpha=+1) \\ d_{RE}^2(\vec{q},\vec{p},n) &= 2n\left[\sum_{i=1}^{c} q_i \log\frac{q_i}{p_i}\right] \quad (\alpha=-1) \end{aligned} \tag{3.34}$$

The first is the Hellinger metric, and the others are the two possible Kullbeck-Leibler relative entropies [21]. One way to symmetrize the Kullbeck-Leibler divergence is to take the arithmetic mean of its two versions, that is

$$d_{KL}^2 = n\left[\sum_{i=1}^{c} p_i \log\frac{p_i}{q_i} + \sum_{i=1}^{c} q_i \log\frac{q_i}{p_i}\right] \tag{3.35}$$

Yet another way to symmetrize the Kullbeck-Leibler divergence is to seek the histogram $\{R_1,\ldots R_c\}$ that solves

$$\begin{aligned} \text{[IKL] Minimize } T &= \sum_{i=1}^{c} R_i \log\frac{R_i}{p_i} + \sum_{i=1}^{c} R_i \log\frac{R_i}{q_i} \\ \text{subject to } &\sum_{i=1}^{c} p_i = \sum_{i=1}^{c} q_i = \sum_{i=1}^{c} R_i = 1 \end{aligned} \tag{3.36}$$

The solution to [IKL] is

$$R_i = \sqrt{p_i q_i}/\sum_{j=1}^{c}\sqrt{p_j q_j} \tag{3.37}$$

and the distance measure corresponding to the minimum objective value is

$$d_{DP}^2 = 8n\left[-\log\sum_{i=1}^{c}\sqrt{p_i q_i}\right] \tag{3.38}$$

which is also known as the Bhattacharyya distance [21]. The relation between the Hellinger and log dot product measures is

$$\frac{1}{8n}d_{DP}^2 = -\log\left[1-\frac{1}{8n}d_H^2\right] \tag{3.39}$$

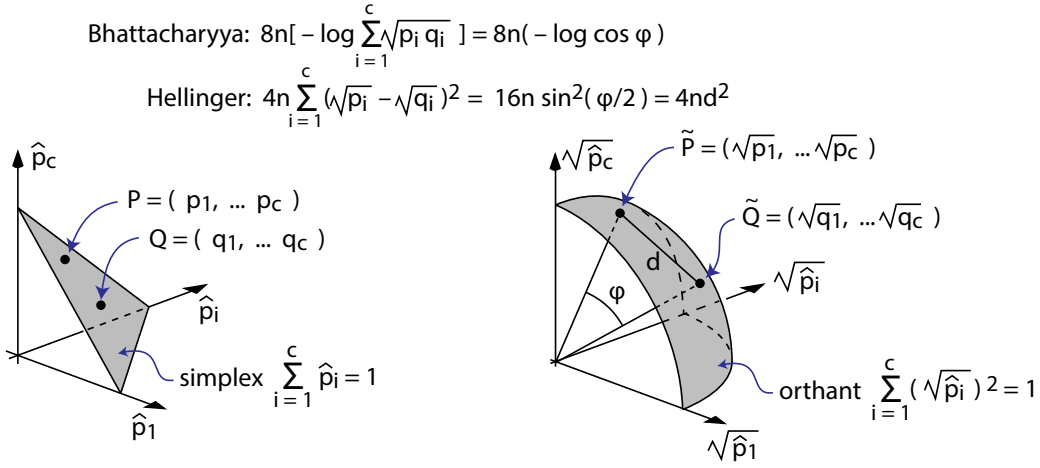


Figure 3.6: Geometric interpretation of Hellinger and Bhattacharyya distance measures for histograms

Both the Hellinger and log dot product measures have geometric interpretations (see Figure 3.6):

$$\begin{aligned} d_H^2 &= 4n\sum_{i=1}^{c}(\sqrt{p_i}-\sqrt{q_i})^2 = 4n(2\sin\frac{\phi}{2})^2 \approx 4n\phi^2(1-\frac{1}{6}\phi^2) \\ d_{DP}^2 &= 8n\left[-\log\sum_{i=1}^{c}\sqrt{p_i q_i}\right] = 8n(-\log\cos\phi) \approx 4n\phi^2(1+\frac{1}{6}\phi^2) \end{aligned} \tag{3.40}$$

where ϕ is the angle between the vectors $(\sqrt{p_1}\ldots\sqrt{p_c})$ and $(\sqrt{q_1}\ldots\sqrt{q_c})$.

The Hellinger metric is a Euclidean distance and hence a true metric, satisfying the triangle inequality. It has a value for any two histograms of equal size. The log dot product, however, does not satisfy the triangle inequality when the three points $(\sqrt{p_1}\ldots\sqrt{p_c})$, $(\sqrt{q_1}\ldots\sqrt{q_c})$, and $(\sqrt{r_1}\ldots\sqrt{r_c})$ are on, or close to being on, a great circle (see appendix A). It has a value for any two non-orthogonal histograms (orthogonal histograms being those that have $\sum_{i=1}^{c} p_i q_i = 0$). Both measures are robust, in the sense that they allow some $p_i = 0$, and the log dot product emphasizes the case when $\sum_{i=1}^{c} p_i q_i \approx 0$. A version of the relative entropy that does obey the triangle inequality is the Jensen-Shannon divergence with equal histogram weighting [74]:

$$d_{JS}^2 = n\left[\sum_{i=1}^{c} p_i \log\frac{2p_i}{p_1+q_i} + \sum_{i=1}^{c} q_i \log\frac{2q_i}{p_i+q_i}\right] \tag{3.41}$$

Finally, we may symmetrize the basic chi-square distance through

$$d^2_{CS} = 2n \sum_{i=1}^{c} \frac{1}{p_i + q_i}(p_i - q_i)^2 \tag{3.42}$$

where we must require that the summand be zero when $p_i = q_i = 0$. In order to compare these various measures, let's expand them all in powers of $\triangle p_i = q_i - p_i$:

$$\begin{aligned}
d^2_{CS} &= n \sum_{i=1}^{c} \left[\frac{(\triangle p_i)^2}{p_i} - \frac{1}{2}\frac{(\triangle p_i)^3}{p_i^2} + \frac{1}{4}\frac{(\triangle p_i)^4}{p_i^3} \ldots \right] \\
d^2_{H} &= n \sum_{i=1}^{c} \left[\frac{(\triangle p_i)^2}{p_i} - \frac{1}{2}\frac{(\triangle p_i)^3}{p_i^2} + \frac{5}{16}\frac{(\triangle p_i)^4}{p_i^3} \ldots \right] \\
d^2_{DP} &= n \sum_{i=1}^{c} \left[\frac{(\triangle p_i)^2}{p_i} - \frac{1}{2}\frac{(\triangle p_i)^3}{p_i^2} + \frac{5}{16}\frac{(\triangle p_i)^4}{p_i^3} \ldots \right] + \frac{1}{16}\left[\sum_{i=1}^{c} \frac{(\triangle p_i)^2}{p_i} \right]^2 \ldots \\
d^2_{KL} &= n \sum_{i=1}^{c} \left[\frac{(\triangle p_i)^2}{p_i} - \frac{1}{2}\frac{(\triangle p_i)^3}{p_i^2} + \frac{1}{3}\frac{(\triangle p_i)^4}{p_i^3} \ldots \right]
\end{aligned} \tag{3.43}$$

Bear in mind that for large sample size n, the quantities $\triangle p_i / p_i$ are of order $1/\sqrt{n}$, and so for each of the expressions above, the expected difference from the basic sum $n \sum_{i=1}^{c} (\triangle p_i)^2 / p_i$ is of order $1/n$. That expected difference will be the least for the symmetrized chi-square and Hellinger distances.

3.5 Other distance measures for histograms

The last few sections have demonstrated that in the domain of histograms, knowledge of the information matrix allows us to construct a distance measure that performs optimal classification, in the sense of having minimum type II error for a given type I error. In this section, we will show that weighted distances, measures based on the cumulative histogram, and earth-mover distances may all have sub-optimal classification performance, when averaged over direction.

Consider again functions of $\vec{z}$, where $\vec{z} = L\triangle\vec{p}$, $L^T L = nS(\vec{p})$, and $\triangle\vec{p} = \hat{p} - \vec{p}$. The L_2-squared function

$$u_2 = |\vec{z}\,|^2 = \sum_{i=1}^{m} z_i^2 \tag{3.44}$$

is a chi-square(m) variate to order $1/n$ in the sample size n. Assuming it is a true chi-square(m) variate, the corresponding density is

$$g(u_2) = \frac{1}{2^{m/2}\Gamma(m/2)} u_2^{m/2-1} \exp\left(-\frac{1}{2}u_2\right) \tag{3.45}$$

From this, we may find that the expected value of u_2^p is

$$\langle u_2^p \rangle = 2^p \Gamma(\frac{m}{2} + p)/\Gamma(\frac{m}{2}) \tag{3.46}$$

and using the properties of the gamma function, we may further deduce

$$\frac{\mathrm{var}(\sqrt{u_2})}{\langle\sqrt{u_2}\rangle^2} = \frac{\langle u_2\rangle}{\langle\sqrt{u_2}\rangle^2} - 1 = \frac{1}{m}\left(\frac{1}{2} + \frac{1}{8m} - \frac{1}{16m^2}\ldots\right) \tag{3.47}$$

which is verified by the $k = 2$ entry of Table 3.1 above. The L_1 version would be $u_1 = \sum_{i=1}^m |z_i|$, which gives

$$\frac{\mathrm{var}(u_1)}{\langle u_1\rangle^2} = \frac{1}{m}\left(\frac{\pi}{2} - 1\right) \tag{3.48}$$

which is in turn verified by the $k = 1$ entry of Table 3.1. Now suppose that v_1 and v_2 are convex combinations based on u_1 and u_2 respectively. Then

$$v_1 = \sum_{i=1}^m c_i|z_i| \text{ and } v_2 = \sum_{i=1}^m c_i|z_i|^2 \text{ where } c_i \geq 0 \text{ and } \sum_{i=1}^m c_i = 1 \tag{3.49}$$

In particular, set

$$c_i = \frac{1}{m} + \delta_i \quad \text{where } \sum_{i=1}^m \delta_i = 0 \tag{3.50}$$

Then, to first order in $\sum \delta_i^2$, we get

$$\frac{\mathrm{var}(v_1)}{\langle v_1\rangle^2} \approx \left(\frac{\pi}{2} - 1\right)\left(\frac{1}{m} + \sum_{i=1}^m \delta_i^2\right) \quad \text{and} \quad \frac{\mathrm{var}(\sqrt{v_2})}{\langle\sqrt{v_2}\rangle^2} \approx \frac{1}{2}\left(\frac{1}{m} + \sum_{i=1}^m \delta_i^2\right) \tag{3.51}$$

Thus, if $|\delta_i| \approx 1/m$, which is not a very large deviation from uniform weighting, then $\sum \delta_i^2 \approx 1/m$ and the variance of both L_1 and L_2 measures approximately doubles.

Now, let's make a quick examination of distance measures based on the cumulative histogram $\{P_1, \ldots, P_{c-1}\} = \{p_1, p_1 + p_2, \ldots, (p_1 + p_2 + \ldots + p_{c-1})\}$. Two points about the theory of cumulative histograms: (1) Not all histograms have a cumulative representation.

If the bins are categories, distinguishable but not sortable, then there's no cumulative histogram. (2) If the bins are sortable but circular, meaning that the ground distance between bins is periodic, then the cumulative histogram is not well defined. It could start at any bin. Examples of circular distributions are those based on angle or color.

Let's narrow our focus to the Cramer - von Mises distance [42], here defined as

$$d_{CM}^2 = n \sum_{i=1}^{c-1} (P_i - \hat{P}_i)^2 \tag{3.52}$$

where P is the true cumulative histogram $\{p_1, p_1 + p_2, \ldots (p_1 + p_2 + \ldots p_{c-1})\}$, $\hat{P}$ is the sample cumulative histogram $\{\hat{p}_1, \hat{p}_1 + \hat{p}_2, \ldots (\hat{p}_1 + \hat{p}_2 + \ldots \hat{p}_{c-1})\}$ and n is the sample size. Then

$$\begin{aligned} d_{CM}^2/n &= (p_1 - \hat{p}_1)^2 + [(p_1 - \hat{p}_1) + (p_2 - \hat{p}_2)]^2 + \ldots + [\sum_{i=1}^{c-1} (p_i - \hat{p}_i)]^2 \\ &= \sum_{i=1}^{c-1} \sum_{j=1}^{c-1} (p_i - \hat{p}_i)(p_j - \hat{p}_j)[c - \max(i,j)] \\ &= (\triangle \vec{p})^T M(c) (\triangle \vec{p}) \text{ where } [M(c)]_{ij} = c - \max(i,j) \end{aligned} \tag{3.53}$$

The matrix $M(c)$ is reducible by row operations to a lower-triangular matrix of all ones, so its determinant is one. Hence, as the number of bins gets larger, the condition number of $M(c)$ (the ratio of largest to smallest eigenvalue) also gets larger, as illustrated in Table 3.2. So, as the number of bins gets larger, the Cramer - von Mises distance becomes an ever more unevenly weighted L_2 distance based on $\triangle \vec{p}$. By our previous analysis, its classification accuracy gets correspondingly worse.

Next, we'll take a quick look at distance measures based on the earth mover distance for histograms. We need two concepts here: first, a "ground distance" between bins, i.e. a non-negative function $r(i,j)$ of bin labels i and j, representing a "cost" of shifting probability mass from one bin to the other. Here I assume that $r(i,j) = r(j,i)$ and that $r(i,i) = 0$. Second, we need the idea of a "coupling" between histograms $\{p_1, p_2, \ldots p_c\}$ and $\{\hat{p}_1, \hat{p}_2, \ldots \hat{p}_c\}$, being a non-negative function $\gamma(i,j)$ of bin labels such that

$$\sum_{j=1}^{c} \gamma(i,j) = p_i \text{ and } \sum_{i=1}^{c} \gamma(i,j) = \hat{p}_j \tag{3.54}$$

Then the earth mover distance is

$$d_{EM}(\vec{p}, \hat{p}, r) = \min_{\gamma} \sum_{i,j} \gamma(i,j) r(i,j) \tag{3.55}$$

Table 3.2: Growth of condition number of $M(c)$ with the number of bins c, for the Cramer von Mises distance

Number of bins	**Min eigenvalue**	**Max eigenvalue**	**Condition number**
3	0.382	2.618	6.85
4	0.308	5.049	16.4
5	0.283	8.291	29.3
6	0.272	12.34	45.5
7	0.265	17.21	64.9

where the minimum is over all possible couplings between $\vec{p}$ and $\hat{p}$. It clearly varies with the ground distance $r(i,j)$. One well-known result is that if the ground distance is linear, i.e. $r(i,j) = |i-j|$, then the earth mover distance is given by the Kantorovich-Monge metric [96]

$$d_{KM} = \sum_{i=1}^{c} \left| P_i - \hat{P}_i \right| \tag{3.56}$$

where P and $\hat{P}$ are the cumulative histograms of $\vec{p} = \{p_1, \ldots p_c\}$ and $\hat{p} = \{\hat{p}_1, \ldots \hat{p}_c\}$ respectively.

It is immediately clear that the earth mover distance is a weighted L_1 distance. We could have defined a coupling as a flow between the positive $(p_i - \hat{p}_i)$ elements and the negative ones, and would have obtained the same minimum. Such a flow would partition each $\triangle p_i = p_i - \hat{p}_i$ into pieces, and associate costs for each piece. But the overall effect is to assign a total cost to each $\triangle p_i$, so

$$d_{EM}(\vec{p}, \hat{p}, r) = \sum_{i=1}^{c} \lambda(\vec{p}, \hat{p}, r, i) |\triangle p_i| \tag{3.57}$$

We know by our previous analysis that such a measure has less classification accuracy than a uniformly weighted L_1 measure such as

$$d_{L1} = \sum_{i=1}^{c-1} |[L \triangle \vec{p}]_i| \quad \text{where } L^T L = nS(\vec{p}) \tag{3.58}$$

which in turn has less classification accuracy than a uniformly weighted L_2 measure:

$$d_{L2}^2 = (\triangle \vec{p})^T nS(\vec{p})(\triangle \vec{p}) = |\vec{z}|^2 \quad \text{where } \vec{z} = L \triangle \vec{p} \tag{3.59}$$

Table 3.3: Empirical variance and third central moment for distance measures with unit sample mean, for two histograms p_1 and p_2 (equation 3.61). Sample size is 8000, $m = 6$, and the number of runs is 512. The L_2 distance has the lowest values for both central moments.

example	**distance measure**	m * variance		m^2 * third moment	
		mean	**std dev**	**mean**	**std dev**
none	L_2 (from Table 3.1)	0.519	0.008	0.291	0.027
p_1	straight Euclidean	0.578	0.009	0.445	0.035
p_1	Kantorovich Monge	1.350	0.023	4.46	0.25
p_1	Cramer von Mises	1.148	0.020	3.05	0.17
p_2	straight Euclidean	0.706	0.011	0.935	0.062
p_2	Kantorovich Monge	1.495	0.025	5.11	0.27
p_2	Cramer von Mises	1.254	0.020	3.33	0.18

Table 3.3 illustrates this theory. The methodology of this diagnostic is as follows: sample a theoretical histogram $\vec{p}$ from its known distribution (equation 3.14 above), to get $\hat{p}$ for a specific sample size n. For each such pair $\{\hat{p}, n\}$, calculate the target distance measure to the model $\vec{p}$. For a set of n such distance values, normalize the mean to 1, and calculate variance and third central moment. Collect these (variance, 3rd central moment) pairs into a large dataset, and calculate both means and standard deviations for representative distance measures. In this table, the "straight Euclidean" distance is the weighted L_2 distance

$$d_E^2 = n \sum_{i=1}^{c} (p_i - \hat{p}_i)^2 = (\triangle \vec{p})^T n (I_{c-1} + J_{c-1})(\triangle \vec{p}) \tag{3.60}$$

and the two example histograms are:

$$\begin{aligned} p_1 &= \{0.17, 0.05, 0.09, 0.20, 0.17, 0.20\} \quad \text{with average } |\delta_i| = 0.05 \\ p_2 &= \{0.28, 0.09, 0.32, 0.04, 0.08, 0.04\} \quad \text{with average } |\delta_i| = 0.09 \end{aligned} \tag{3.61}$$

Table 3.3 illustrates that for these two histogram examples, the L_2 distance has the lowest variance when the sample mean is constrained to be one.

We can illustrate the superiority of the distance measures based on the information matrix in yet another way. This diagnostic test samples closely-spaced beta distributions into histograms with three bins, with boundaries [0.0, 0.4, 0.6, 1.0], and calculates the empirical type 2 classification error using four different distance measures. The beta probability

Table 3.4: Theoretical type II errors for histograms based on beta distributions. Sample size is $n_1 = n_2 = 8000$, $m = 2$, and cluster $C_0(a = 5.80, b = 4.20)$ is taken as primary.

Cluster	**Beta a**	**Beta b**	**Histogram model**	$(\triangle\vec{p})^T Q_{12}(\triangle\vec{p})$	**type II error**
C_0	5.80	4.20	[0.1246 0.4117 0.4638]	—	—
C_1	5.71	4.29	[0.1372 0.4229 0.4399]	11.16	0.142
C_2	5.12	3.68	[0.1371 0.3904 0.4725]	10.00	0.184

density [1] is given by:

$$p(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad a > 0, b > 0, x \in [0,1] \tag{3.62}$$

We can calculate the "ground truth" as follows: the 95% confidence level for a chi-square(2) variate is at 5.99, so the type II error for two 3-bin histograms $\vec{p}_1$ and $\vec{p}_2$ will be the non-central chi-square(2) CDF value at 5.99 with the non-centrality parameter $\lambda^2 = (\triangle\vec{p})^T Q_{12}(\triangle\vec{p})$, where $\triangle\vec{p} = \vec{p}_2 - \vec{p}_1$ and Q_{12} is given by

$$Q_{12}^{-1} = \frac{1}{n_1}[\mathrm{diag}(\vec{p}_1) - \vec{p}_1\,\vec{p}_1^T] + \frac{1}{n_2}[\mathrm{diag}(\vec{p}_2) - \vec{p}_2\,\vec{p}_2^T] \tag{3.63}$$

where n_1 and n_2 are the respective sample sizes. These theoretical values are in Table 3.4. The results of the diagnostic test itself are in table 3.5. The tolerance for the empirical type II error may be calculated from the DKW inequality [81], and is approximately ± 0.010 for this dataset.

What Table 3.5 shows is what we predicted earlier. The Hellinger and chi-square measures, which reduce to the optimal quadratic form for large sample size, give empirical T_2 values (i.e. type II error values) in accordance with the ground truth. The Euclidean measure has a demonstrably larger variance for the same mean, hence has positively biased T_2 values. The Cramer - von Mises measure has a larger variance, and distorts the offset $(\triangle\vec{p})^T Q_{12}(\triangle\vec{p})$ by virtue of equation 3.53, so its empirical T_2 values can be substantially far from the true ones.

3.6 Optimum number of bins for a histogram

The theory developed so far assumes that we know the number of bins in our histograms of interest. Is there a "best" number of bins Q for a given sample size N? In order to make this

Table 3.5: Empirical type II errors for histograms based on beta distributions. Sample size is $n_1 = n_2 = 8000$, $m = 2$, and number of runs is 16000.

Clusters	**Distance measure**	**mean**	**m * variance**	**arg** $(1 - T_1)$	T_2
C_0, C_0	Hellinger	1.00	0.536	1.95	—
C_0, C_0	Chi-square	1.00	0.542	1.96	—
C_0, C_0	Euclidean	1.00	0.612	2.03	—
C_0, C_0	Cramer Von Mises	1.00	0.633	2.06	—
C_0, C_1	Hellinger	2.79	1.22	—	0.145
C_0, C_1	Chi-square	2.80	1.22	—	0.144
C_0, C_1	Euclidean	2.82	1.45	—	0.183
C_0, C_1	Cramer Von Mises	3.30	1.88	—	0.103
C_0, C_2	Hellinger	2.67	1.19	—	0.183
C_0, C_2	Chi-square	2.65	1.21	—	0.192
C_0, C_2	Euclidean	2.57	1.31	—	0.261
C_0, C_2	Cramer Von Mises	2.07	0.74	—	0.501

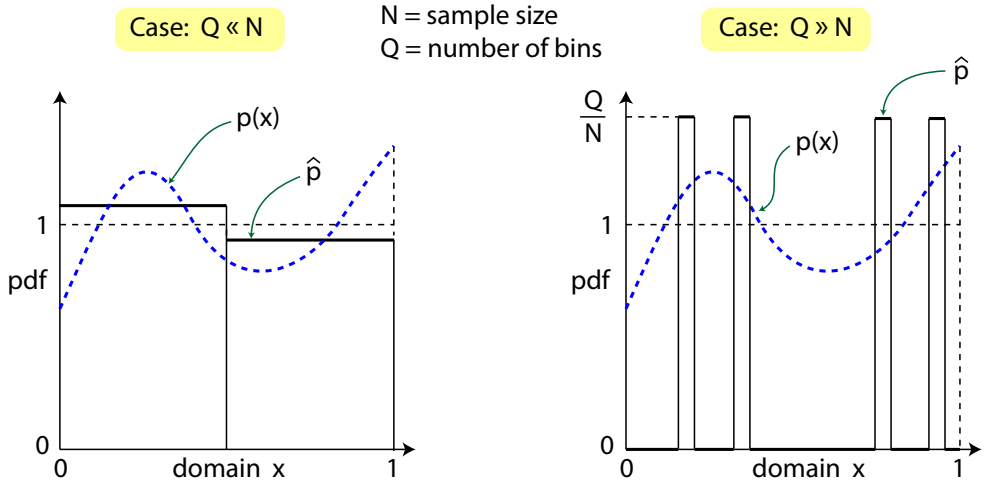


Figure 3.7: Approximating a pdf via histograms: the cases of too few bins, and too many bins

question precise, let's suppose that the purpose of the target histogram is to approximate an underlying continuous probability density distribution. Then the two extremes of "too few bins" and "too many bins" will be as illustrated in Figure 3.7. Suppose that our domain is of unit length, divided into Q bins of equal size, and that the underlying pdf is $p(x)$, smooth enough to allow a second-order Taylor expansion. When $Q \approx 1$, then the estimate $\hat{p}(x)$ derived from the histogram is too coarse. When $Q \gg N$, the bins end up having mostly just one count or none, so we get N "spikes" with pdf $(1/N)/(1/Q) = Q/N$, and zero elsewhere.

To capture this, we will use the continuous version of the Hellinger measure, namely

$$d_H^2 = 1 - \int_0^1 \sqrt{p(x)\hat{p}(x)}\,\mathrm{d}x \tag{3.64}$$

When $Q = 1$, then $\hat{p}(x) = 1$ and we get

$$d_H^2 = 1 - \int_0^1 \sqrt{p(x)}\,\mathrm{d}x \tag{3.65}$$

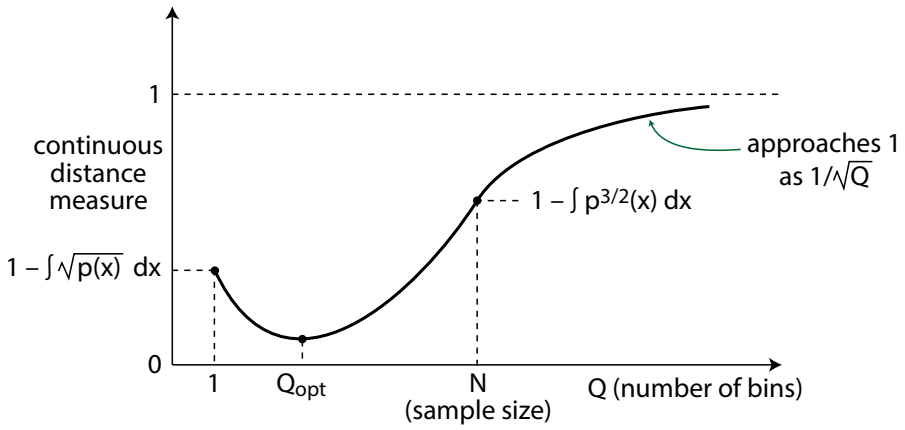


Figure 3.8: Conceptual variation of continuous Hellinger measure. N is the sample size and Q is the number of bins

which is positive as long as $p(x)$ is not constant. When $Q \gg N$, we get

$$d_H^2 \approx 1 - \sum_{\text{matches}} \sqrt{p(x_i)} \sqrt{\frac{Q}{N}} \left(\frac{1}{Q} \right) \tag{3.66}$$

where x_i is the center of bin i, and "matches" refers to those bins with a count. Assuming that $N \gg 1$,

$$\begin{aligned} \sum_{\text{matches}} \sqrt{p(x_i)} &\approx N \langle \sqrt{p(x)} \rangle \\ \Rightarrow d_H^2 &\approx 1 - N \langle \sqrt{p(x)} \rangle \sqrt{\frac{1}{QN}} \approx 1 - \sqrt{\frac{N}{Q}} \int_0^1 p^{3/2}(x)\, \mathrm{d}x \end{aligned} \tag{3.67}$$

So, we can expect the variation of d_H^2 with Q to be as illustrated in Figure 3.8.

For the intermediate case $1 \ll Q \ll N$, we will apply multinomial statistics, i.e. the theory of sampling from a very large parent population with several categories. Let i be

the bin number, so $1 \le i \le Q$, and x_i will be the bin center. The theoretical value of p_i, the total probability for the bin, is

$$p_i = \int_{x_i-1/2Q}^{x_i+1/2Q} p(x)\,\mathrm{d}x \approx p(x_i)\frac{1}{Q} + \frac{1}{24}p''(x_i)\frac{1}{Q^3} \text{ where } p''(x) = \frac{d^2}{dx^2}p(x) \tag{3.68}$$

We will also need the integral of $\sqrt{p(x)}$ over a bin, so we need the Taylor expansion of $\sqrt{p(x)}$, which I assume to be valid up to second order:

$$\sqrt{p(x)} \approx \sqrt{p(x_i)} + \frac{1}{2}\frac{p'(x_i)}{\sqrt{p(x_i)}}(x-x_i) + \left[\frac{1}{4}\frac{p''(x_i)}{\sqrt{p(x_i)}} - \frac{1}{8}\frac{p'^2(x_i)}{p^{3/2}(x_i)}\right](x-x_i)^2 \tag{3.69}$$

Integrating that over a bin yields

$$t_i = \int_{x_i-1/2Q}^{x_i+1/2Q} \sqrt{p(x)}\,\mathrm{d}x \approx \sqrt{p(x_i)}\frac{1}{Q} + \frac{1}{48}\frac{p''(x_i)}{\sqrt{p(x_i)}}\frac{1}{Q^3} - \frac{1}{96}\frac{p'^2(x_i)}{p^{3/2}(x_i)}\frac{1}{Q^3} \tag{3.70}$$

and assuming that $p(x)$ is smooth enough so that $p''(x)/p(x) \ll Q^2$ (recall that $Q \gg 1$), we get

$$t_i \approx \sqrt{\frac{p_i}{Q}}\left[1 - \frac{1}{96}\left(\frac{p'(x_i)}{p(x_i)}\frac{1}{Q}\right)^2\right] \tag{3.71}$$

Turning now to the estimates formed by calculating histograms from samples of size N, suppose that the observed counts over the Q bins for one such estimate are $\{f_1, \ldots f_Q\}$. Then

$$\begin{aligned} \mathrm{var}(f_i) &= Np_i(1-p_i) \quad \text{marginal distribution} \\ \langle f_i \rangle &= Np_i \quad \text{expectation over all } \{f_1, \ldots f_Q\} \end{aligned} \tag{3.72}$$

Since we are treating the intermediate case where $Q \ll N$, we can say that $\sqrt{\mathrm{var}(f_i)} \ll \langle f_i \rangle$, so

$$\sqrt{f_i} = [\langle f_i \rangle + f_i - \langle f_i \rangle]^{1/2} \approx \sqrt{\langle f_i \rangle}\left[1 + \frac{1}{2}\frac{f_i - \langle f_i \rangle}{\langle f_i \rangle} - \frac{1}{8}\frac{(f_i - \langle f_i \rangle)^2}{\langle f_i \rangle^2}\right] \tag{3.73}$$

Taking the expectation over all $\{f_1, \ldots f_Q\}$ yields

$$\langle \sqrt{f_i} \rangle \approx \sqrt{\langle f_i \rangle}\left[1 - \frac{1}{8}\frac{\mathrm{var}(f_i)}{\langle f_i \rangle^2}\right] \tag{3.74}$$

and putting that together with our formula for t_i gives

$$\begin{aligned}\langle d_H^2\rangle &= 1 - \sum_{\text{bins}} t_i \sqrt{\frac{Q}{N}} \langle\sqrt{f_i}\rangle = 1 - \sum_{\text{bins}} t_i \sqrt{Q p_i}\left[1 - \frac{1}{8}\frac{\text{var}(f_i)}{\langle f_i\rangle^2}\right] \\ &\approx \sum_{\text{bins}} p_i \left[\frac{1}{8}\frac{\text{var}(f_i)}{\langle f_i\rangle^2} + \frac{1}{96}\left(\frac{p'(x_i)}{p(x_i)}\frac{1}{Q}\right)^2\right]\end{aligned} \tag{3.75}$$

Using $\langle f_i\rangle = N p_i$ and $p_i \approx p(x_i)/Q$ (to first order), we then get

$$\begin{aligned}\langle d_H^2\rangle &\approx \frac{1}{8N}(Q-1) + \frac{1}{96}\sum_{\text{bins}} \frac{p(x_i)}{Q}\left(\frac{p'(x_i)}{p(x_i)}\frac{1}{Q}\right)^2 \\ &\approx \frac{1}{8N}(Q-1) + \frac{1}{96}\left[\int_0^1 \frac{p'^2(x)}{p(x)}\,\mathrm{d}x\right]\frac{1}{Q^2}\end{aligned} \tag{3.76}$$

and that gives us our final transformation of $\langle d_H^2\rangle$:

$$\begin{aligned}&\frac{d}{dx}\sqrt{p(x)} = \frac{1}{2}\frac{p'(x)}{\sqrt{p(x)}} \quad \Rightarrow \frac{p'^2(x)}{p(x)} = 4\frac{d}{dx}\sqrt{p(x)} \\ &\Rightarrow \langle d_H^2\rangle \approx \frac{1}{8N}(Q-1) + \frac{1}{24Q^2}\int_0^1 \left[\frac{d}{dx}\sqrt{p(x)}\right]^2 \mathrm{d}x\end{aligned} \tag{3.77}$$

Minimizing $\langle d_H^2\rangle$ with respect to Q gives us

$$Q_{\text{opt}}^3 = \frac{2}{3}N\int_0^1 \left[\frac{d}{dx}\sqrt{p(x)}\right]^2 \mathrm{d}x \tag{3.78}$$

I will illustrate the validity of that formula with two different distributions. For a beta distribution on $[0,1]$ with parameters (a,b) (equation 3.62), the membrane energy of $\sqrt{p(x)}$ is

$$\int_0^1 \left[\frac{d}{dx}\sqrt{p(x)}\right]^2 \mathrm{d}x = \frac{1}{4}(a+b-1)(a+b-2)\left(\frac{1}{a-2} + \frac{1}{b-2}\right) \tag{3.79}$$

so for the case $a = 4, b = 4, N = 2000$ we get $Q_{\text{opt}} \approx 24$. For a normal distribution over $[-0.5, +0.5]$ with zero mean and standard deviation $1/(2s)$, we get

$$\begin{aligned}&\int_0^1 \left[\frac{d}{dx}\sqrt{p(x)}\right]^2 \mathrm{d}x = s^2\left[1 - 2s\frac{g(s)}{\phi(s) - \phi(-s)}\right] \quad \text{where} \\ &g(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2) \text{ and } \phi(z) = \int_{-\infty}^{z} g(x)\,\mathrm{d}x\end{aligned} \tag{3.80}$$

For the case $s = 2, N = 2000$ that gives $Q_{\text{opt}} \approx 16$. Experimental results verifying these predictions are in Figures 3.9 and 3.10.

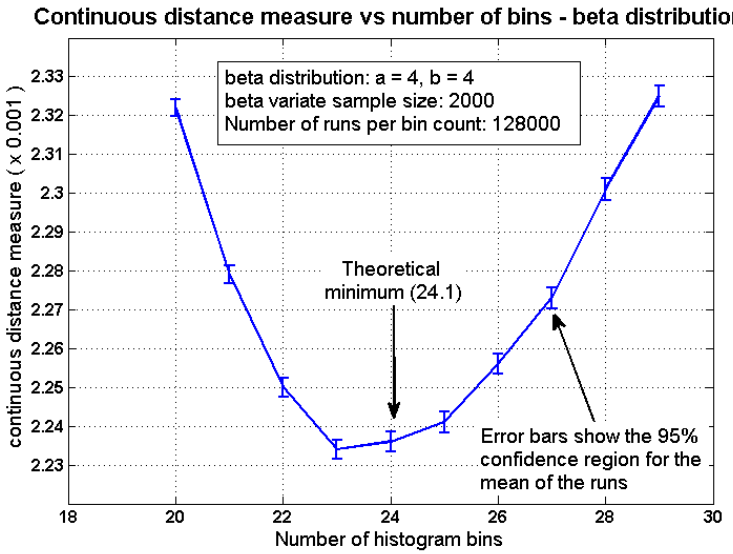


Figure 3.9: Optimal number of bins for samples from a beta distribution. The continuous distance measure is defined in equation 3.64.

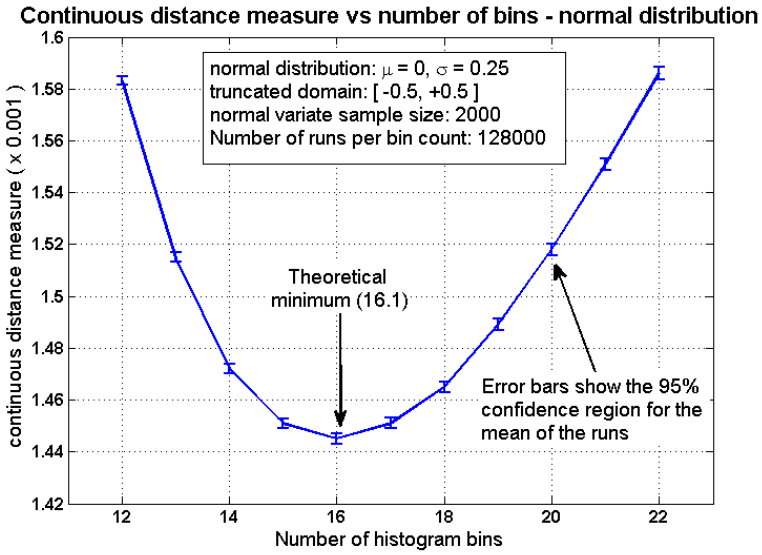


Figure 3.10: Optimal number of bins for samples from a truncated normal distribution. The continuous distance measure is defined in equation 3.64.

3.7 Histogram with a variable number of bins

In the last section, we condidered the "best" number of bins that could represent a distribution whose probability distribution function in known, or at least can be estimated, when the sample size is given. Now, I wish to examine the problem of using histograms to distinguish samples from two distributions, which are different but close to each other. Is there a "best" number of bins to use in this case?

I consider as solved the problem of deciding the bin boundaries, given the two probability distributions and the number of bins. Let $p_1(x)$ and $p_2(x)$ be the actual pdfs in question, let c be the number of bins, let n_1 and n_2 be the sample sizes, and let $\{b_1, \dots b_{c+1}\}$ be the bin boundaries (which may vary). Then, in the current context, that would amount to the following optimization problem:

$$\begin{aligned} &\text{[BB] Maximize } T = 1 - \sum_{j=1}^{c} \sqrt{p_{1,j} p_{2,j}} \quad \text{where } p_{w,j} = \int_{b_j}^{b_{j+1}} p_w(x)\,\mathrm{d}x \\ &\text{subject to } b_j < b_{j+1} \quad \text{and } n_w p_{w,j} > 5 \text{ for all } w = 1,2 \text{ and } j = 1, \dots, c \end{aligned} \tag{3.81}$$

where the last constraint on $p_{w,j}$ is a practical one, merely ensuring that the bin counts are not tiny [24].

Let the sampled histograms be $\{\hat{p}_{1,1}, \dots \hat{p}_{1,c}\}$ and $\{\hat{p}_{2,1}, \dots \hat{p}_{2,c}\}$, based on sample sizes of n_1 and n_2 respectively. Assuming that $p_1(x) \approx p_2(x)$, then

$$\begin{aligned} d_H^2 &= 8\frac{n_1 n_2}{n_1 + n_2}\left[1 - \sum_{j=1}^{c} \sqrt{\hat{p}_{1,j}\hat{p}_{2,j}}\right] \\ &\approx \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{c} \frac{(\hat{p}_{1,j} - \hat{p}_{2,j})^2}{\bar{p}_j} \text{ where } \bar{p}_j = \frac{n_1}{n_1 + n_2} p_{1,j} + \frac{n_2}{n_1 + n_2} p_{2,j} \end{aligned} \tag{3.82}$$

and $p_{w,j}$ represents the theoretical proportion of distribution p_w on bin j, for $w = 1, 2$. In this case, $\hat{p}_{1,j} - \hat{p}_{2,j}$ has a fixed part $p_{1,j} - p_{2,j}$ and a random part $(\hat{p}_{1,j} - p_{1,j}) - (\hat{p}_{2,j} - p_{2,j})$. We know from section 3.5 above that for $n_1, n_2 \gg 1$, the distribution of d_H^2 approaches a non-central chi-square with degrees of freedom $c - 1$ and offset given by

$$k = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{c} \frac{(p_{1,j} - p_{2,j})^2}{\bar{p}_j} \tag{3.83}$$

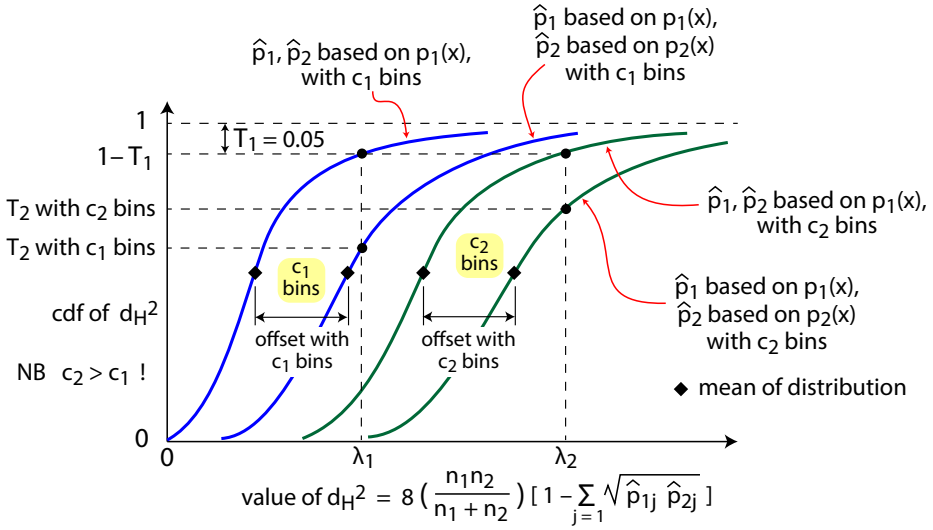


Figure 3.11: Variation of type II error with number of bins c. If the offset stays constant as c increases, then the type II error increases.

Assume for the moment that the domain of p_1 and p_2 is of size 1, with constant bin width $1/c$. Then

$$\begin{aligned} k &\approx \frac{n_1 n_2}{n_1+n_2} \sum_{j=1}^{c} \frac{[p_1(x_j) - p_2(x_j)]^2 c^{-2}}{\bar{p}(x_j) c^{-1}} = \frac{n_1 n_2}{n_1+n_2} \sum_{j=1}^{c} \frac{[p_1(x_j) - p_2(x_j)]^2}{\bar{p}(x_j)} \left(\frac{1}{c}\right) \\ &\approx n_1 n_2 \int_0^1 \frac{[p_1(x) - p_2(x)]^2}{n_1 p_1(x) + n_2 p_2(x)}\, dx \text{ where } \bar{p}(x_j) = \frac{n_1}{n_1+n_2} p_1(x_j) + \frac{n_2}{n_1+n_2} p_2(x_j) \end{aligned} \tag{3.84}$$

and the $\{x_j\}$ are the bin centers. Note that for $c \gg 1$, the offset k no longer depends on the number of bins. Thus, as the number of bins c increases, the "relative" distance between chi-square$(c-1)$ and non-central chi-square$(c-1, k)$ gets smaller and smaller. See Figure 3.11.

Suppose, for example, that $m = c - 1 \gg 1$. Then, both chi-square(m) and non-central chi-square(m, k) are well approximated by normal distributions. In this case, we have

$$\begin{aligned} \arg(1 - T_1) &\approx m + 1.65\sqrt{2m} \text{ where } T_1 = 0.05 \\ T_2 &\approx \phi\left(\frac{1.65\sqrt{2m} - k}{\sqrt{2m + 4k}}\right) \approx \phi\left(1.65 - \frac{k}{\sqrt{2m}}\right) \end{aligned} \tag{3.85}$$

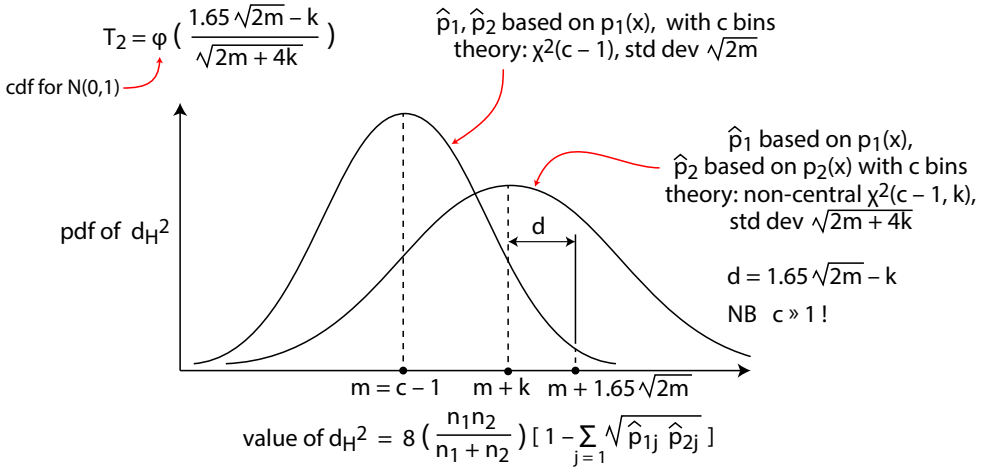


Figure 3.12: Normal approximation for type II error when number of histogram bins is large

where $\phi(x)$ is the cumulative distribution function for the standard normal distribution $N(0,1)$. See Figure 3.12 for an illustration of this case. Note that the type II error T_2 increases uniformly with m, and that we get the lowest T_2 when the number of bins c is as small as possible.

As an example, consider the two distributions in Figure 3.13. These are both derived from beta distributions (equation 3.62), but have been scaled and shifted so that they are zero-mean and have similar variances. The H_0 distribution is derived from beta(26, 34), and the H_1 distribution is derived from beta(16, 11). First, we'll calculate the expected values of type II error for histograms on 4, 6, and 8 bins, using the Hellinger distance measure, and using the pdf crossing points as our fixed bin boundaries. These are shown in Table 3.6. In this table, the last column is the approximation of equation 3.85 that is valid when $c \gg 1$.

Table 3.7 collects the experimental results for the same H_0, H_1 combinations, in which $n_1 = n_2 = 2048$, and the number of trials per combination is 16000. For this number of trials, the tolerance for the empirical type II error may be calculated from the DKW inequality and is approximately ± 0.010. The key observation from Table 3.7 is that, as predicted by Figure 3.11 and equation 3.85, increasing the number of bins in a histogram

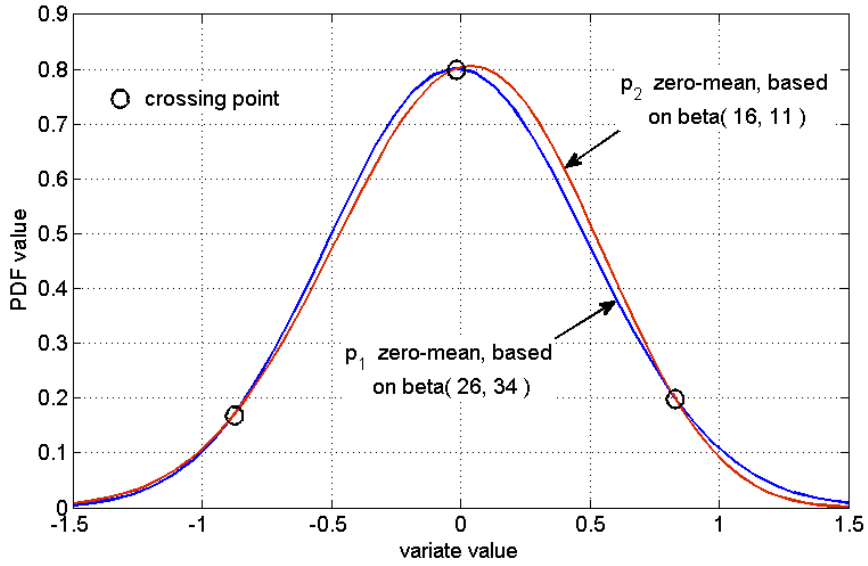


Figure 3.13: Example beta distributions with zero mean but different shapes

Table 3.6: Theoretical type II errors for histograms with different numbers of bins

scenario	**no. of bins**	**arg**$(1-T_1)$ **for** d_H^2	**offset k**	T_2	**large c**
H_0, H_1	4	7.815	4.537	0.597	0.42
H_0, H_1	6	11.07	4.887	0.645	0.54
H_0, H_1	8	14.07	4.927	0.691	0.63

Table 3.7: Empirical type II errors for histograms with different numbers of bins. Sample size is 2048 and number of runs is 16000. The type II error eventually increases with the number of bins.

scenario	**no. of bins**	**mean**	**variance**	**arg(1-T_1)**	$\hat{T}_2$
H_0, H_0	4	1.603	0.453	2.812	—
H_0, H_0	6	2.134	0.476	3.348	—
H_0, H_0	8	2.550	0.486	3.761	—
H_0, H_1	4	2.600	0.805	—	0.596
H_0, H_1	6	3.038	0.729	—	0.644
H_0, H_1	8	3.360	0.699	—	0.687

eventually results in a loss of classification accuracy.

What the foregoing analysis shows is that we should use the smallest number of bins that can separate the classes, i.e. the smallest value of c such that $n_w p_{w,j} > 5$ and $n_w(p_{1,j} - p_{2,j}) > 5$ for $w = 1, 2$ and $j = 1, \ldots, c$. Cochran [24] explains the significance of the minimum bin counts. After choosing the number of bins in the histogram, then the next step would be to optimize the bin boundaries using Eq. 3.81 (if required).

3.8 Other non-parametric and semi-parametric measures

Suppose now that we have samples from two very similar distributions, and that this time we have the samples themselves, not just histograms derived from them. I assume here that the variates in each sample are numeric, sortable, and IID (independent and identically distributed). Besides histograms, what else could we construct from those samples that may lead to a distance measure?

First up to bat are the non-parametric measures based on the empirical cumulative distribution function. If we have just one sample with n sortable IID variates $\{x_1, \ldots x_n\}$, then the empirical CDF is

$$F_n(x) = \{ \text{ proportion of sample with } x_j < x \} \tag{3.86}$$

If we have two such samples, with empirical CDFs $F_n(x)$ and $G_m(x)$, then the three most common distance measures involving those empirical CDFs are:

$$\begin{aligned} d_{KS} &= \sqrt{\frac{nm}{n+m}} \max_x |F_n(x) - G_m(x)| \quad \text{(Kolmogorov Smirnov)} \\ d^2_{CM} &= \frac{nm}{n+m} \int_0^1 [F_n(x) - G_m(x)]^2 \, \mathrm{d}H_{n+m}(x) \quad \text{(Cramer von Mises)} \\ d^2_{AD} &= \frac{nm}{n+m} \int_0^1 \frac{[F_n(x) - G_m(x)]^2}{H_{n+m}(x)[1 - H_{n+m}(x)]} \, \mathrm{d}H_{n+m}(x) \quad \text{(Anderson Darling)} \end{aligned} \tag{3.87}$$

where $H_{n+m}(x)$ is the empirical CDF of the joint sample. For more information on these non-parametric measures, see [42] and [31]. I also treat the Cramer von Mises and Anderson-Darling measures in more detail in Appendix C.

If we believe that the parent populations are centralized, and have finite cumulants up to say the 6th or 8th cumulant, then we may look at the sample cumulants and use their known covariance to build a distance measure from the corresponding information matrix. By "sample cumulants", I mean those derived from the sample central moments. For a sample $\{x_1, \ldots, x_n\}$, these are:

$$\begin{aligned} &\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{\mu}_r = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_1)^r \text{ for } r > 1, \text{ with} \\ &\hat{\kappa}_1 = \hat{\mu}_1, \quad \hat{\kappa}_2 = \hat{\mu}_2, \quad \hat{\kappa}_3 = \hat{\mu}_3, \quad \hat{\kappa}_4 = \hat{\mu}_4 - 3\hat{\mu}_2^2, \text{ etc} \end{aligned} \tag{3.88}$$

In other words, I am not worried about possible biases these quantities may have, when considered as estimators of the corresponding population cumulants. I regard them not as parameter estimates, but merely as statistics that characterize the sample.

Next, we'll need the covariance of sample cumulants, for samples of size n. In Appendix B, I describe how to calculate these and give some examples. The interested reader may also consult [52] and [108]. For the first four sample cumulants, the covariance matrix is:

$$\begin{aligned} n \, \mathrm{cov}(\hat{\kappa}_1, \hat{\kappa}_2, \hat{\kappa}_3, \hat{\kappa}_4) &= L + G, \quad L = \mathrm{diag}(\mu_2, 2\mu_2^2, 6\mu_2^3, 24\mu_2^4) \\ G_{12} &= \mu_3, \quad G_{13} = G_{22} = \mu_4 - 3\mu_2^2 \\ G_{23} &= \mu_5 - 4\mu_3\mu_2 \\ G_{33} &= \mu_6 - 6\mu_4\mu_2 - \mu_3^2 + 3\mu_2^3 \\ G_{14} &= \mu_5 - 10\mu_3\mu_2 \\ G_{24} &= \mu_6 - 7\mu_4\mu_2 - 4\mu_3^2 + 6\mu_2^3 \\ G_{34} &= \mu_7 - 9\mu_5\mu_2 - 5\mu_4\mu_3 + 36\mu_3\mu_2^2 \\ G_{44} &= \mu_8 - 12\mu_6\mu_2 - 8\mu_5\mu_3 - \mu_4^2 + 48\mu_4\mu_2^2 + 64\mu_3^2\mu_2 - 60\mu_2^4 \end{aligned} \tag{3.89}$$

Table 3.8: Parent distribution moments for H_0[from beta(26,34)]and H_1 [from beta(16,11)]. These two distributions are illustrated in Figure 3.13.

quantity	H_0	H_1
mean	0.0	0.0
variance	0.2415	0.2328
κ_3	0.00805	-0.01545
κ_4	-0.00516	-0.00935

where G is a symmetric matrix and μ_r refers to the r-th central moment of the parent population. G would be zero for a normally distributed parent population. Here I have neglected all the terms of order $1/n, 1/n^2$ etc, and assume that $\mu_1 = 0$ in the parent population. Using the same procedure as in section 3.3, and neglecting terms of order $1/n$, we get for a cumulant distance measure

$$\begin{aligned} d^2_{SC} &= \frac{n_1 n_2}{n_1 + n_2} (\triangle\kappa)^T (L + G)^{-1} (\triangle\kappa) \text{ where} \\ (\triangle\kappa)^T &= (\hat{\kappa}_{1,1} - \hat{\kappa}_{2,1}, \ldots \hat{\kappa}_{1,4} - \hat{\kappa}_{2,4}) \end{aligned} \tag{3.90}$$

and $\hat{\kappa}_{w,r}$ is the r-th sample cumulant for the sample with label w. When $n_1, n_2 \gg 1$ and the two samples are from the same parent distribution, then d^2_{SC} will follow a chi-square(4) distribution. If the samples are from different distributions, then d^2_{SC} will follow a non-central chi-square distribution (see section 3.5).

Now let's apply those ideas to the example of section 3.7, in which we looked at two parent distributions derived from beta distributions, but scaled and shifted so as to have zero means and similar variances. The H_0 distribution is derived from beta(26, 34), and the H_1 distribution is derived from beta(16, 11). Their theoretical cumulants are given in Table 3.8.

Note that the two distributions are close over their entire range (see Figure 3.13). The results for the cumulant measure are tabulated below, along with the corresponding results for a Hellinger measure based on histograms with 4 and 6 bins, and for the three non-parameteric measures based on the sample empirical CDFs. In each case, I have also included the type II error calculated from the theory developed so far. As for the type II error for the Cramer von Mises and Anderson Darling measures, I show how to calculate those in Appendix C. For this number of trials, the tolerance for the empirical type II error may be calculated from the DKW inequality and is approximately ± 0.010. The key

Table 3.9: Theoretical and experimental type II errors for measures based on cumulants, histograms, and empirical CDFs. H_0 and H_1 are illustrated in 3.13 and differ mainly in their skewness. Sample size is 2048 and number of trials is 16000.

scenario	**measure**	**mean**	**variance**	**arg(1-T_1)**	$\hat{T}_2$	**predicted T_2**
H_0, H_0	cumulants	1.879	0.451	3.070	—	—
H_0, H_0	histogram 4 bins	1.603	0.453	2.818	—	—
H_0, H_0	histogram 6 bins	2.134	0.476	3.348	—	—
H_0, H_0	Kolmogorov Smirnov	0.860	0.0676	1.358	—	—
H_0, H_0	Cramer von Mises	0.378	0.0233	0.678	—	—
H_0, H_0	Anderson Darling	0.943	0.106	1.573	—	—
H_0, H_1	cumulants	3.602	0.690	—	0.268	0.286
H_0, H_1	histogram 4 bins	2.600	0.805	—	0.596	0.597
H_0, H_1	histogram 6 bins	3.038	0.729	—	0.644	0.645
H_0, H_1	Kolmogorov Smirnov	0.999	0.0972	—	0.862	—
H_0, H_1	Cramer von Mises	0.464	0.0343	—	0.867	0.870
H_0, H_1	Anderson Darling	1.213	0.125	—	0.853	0.850

observation from Table 3.9 is that for this pair of distributions, differing mainly in their third central moment, the distance measure based on cumulants has a lower type II error than the distance measures based on histograms or sample cdfs. That makes sense since the cumulant-based distance is designed specifically to detect differences in skewness or kurtosis. That leads to the question: what about parent distributions that differ in other ways?

Table 3.10 attempts to answer that question, and is my final table for this section. It shows the theoretical type II errors for four distance measures (Cramer von Mises, Anderson Darling, histogram, and cumulants) over four types of parent distribution differences (mean shift, variance shift, skewness shift, and kurtosis shift). In this table, a "boxed

normal" distribution is

$$p(x,\mu,\sigma,L) = \frac{1}{\sigma[\phi(L)-\phi(-L)]} g\left(\frac{x-\mu}{\sigma}\right) \quad \text{where}$$
$$g(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) \quad \text{and } \phi(z) = \int_{-\infty}^{z} g(x)\ \mathrm{d}x, \quad x \in [\mu - L\sigma, \mu + L\sigma] \tag{3.91}$$

and a "scaled beta" distribution is

$$p(x,a,b) = \frac{1}{c}\beta(\frac{x}{c}+\mu,a,b) \quad \text{where } \mu = \frac{a}{a+b}, \quad c = \sqrt{a+b}, \quad \text{and}$$
$$\beta(z,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1-z)^{b-1}, \quad x \in [-c\mu, c(1-\mu)] \tag{3.92}$$

In Table 3.10, the "energy" quantity for the Cramer von Mises and Anderson-Darling measures is the value of the corresponding distance in equation 3.87, with the theoretical cumulative distributions in place of the sample cumulative distributions. The key observation from Table 3.10 is that the Cramer von Mises and Anderson-Darling measures are good at detecting a shift in parent distribution mean, while the cumulant based measure is great at detecting a shift in skewness or kurtosis.

The distance measure based on sample cumulants does have its limitations. As given above, it requires that the parent populations of the two samples have 8th central moments. In a pinch, we could drop the fourth cumulant and live with the first three, requiring only the 6th central moment in the parent populations. Also, as is well known, cumulants are susceptible to contamination from sample outliers [45]. But that weakness can be a strength: the example of Table 3.10 shows that a measure based on cumulants can distinguish distributions that differ only in their tails, and are difficult to distinguish in any other way. Moreover, such a measure is quick to calculate and does not require any decisions regarding number of bins or bin boundaries. My interpretation of the results so far is that when comparing samples from different probability distributions, the more information about the underlying PDF we can squeeze into the distance measure, the better we can separate neighboring clusters.

3.9 Summary

1. For histograms, maximum likelihood estimation gives an information matrix

$$L^T L = nS(\vec{p}) = n\left[\mathrm{diag}(\frac{1}{p_1},\ldots,\frac{1}{p_{c-1}}) + \frac{1}{p_c}J_{c-1}\right] \tag{3.93}$$

Table 3.10: Theoretical type II errors for measures based on cumulants, histograms, and empirical CDFs . Sample size is 2048.

category	**specific**	**mean shift**	**variance shift**	**skewness shift**	**kurtosis shift**
setup	distribution	boxed normal	boxed normal	scaled beta	scaled beta
	parameters	μ, σ, L	μ, σ, L	a, b	a, b
	H_0 distro	+0.04, 1, 2	0, 1.024, 2	12, 16	6, 6
	H_1 distro	-0.04, 1, 2	0, 0.976, 2	16, 12	28, 28
	n_1, n_2	2048	2048	2048	2048
CvM	energy	0.692	0.082	0.085	0.0068
	type 2 error	0.26	0.89	0.87	0.94
AD	energy	3.96	1.00	0.574	0.218
	type 2 error	0.20	0.74	0.86	0.93
histogram	number of bins	2	3	4	5
	PDF crossings	0.0	-1, 1	0, $\pm$0.86	$\pm$0.95, $\pm$0.25
	offset	4.58	2.98	4.66	3.69
	type 2 error	0.43	0.68	0.59	0.70
cumulants	offset	10.29	10.42	9.45	9.60
	type 2 error	0.27	0.26	0.31	0.30

where the cluster center is $\vec{p} = [p_1, \ldots, p_{c-1}]$, n is the sample size ($n \gg 1$), and J_m is the all-ones matrix of size m. If $\hat{p}$ is an estimate of $\vec{p}$ based on n sample variates, then the quantity $\vec{z} = L(\hat{p} - \vec{p})$ is approximately distributed as $N(0, I_{c-1})$. The function $f(\vec{z}\,)$ that minimizes var($f(\vec{z}\,)$) while holding $\langle f(\vec{z}\,)\rangle$ constant is a function of $|\vec{z}\,|$ only. Among the L_k family, that would be an L_2 distance. Any other L_k, or an unevenly weighted L_2, will have some directional dependence. If a proposed distance measure $g(\vec{z}\,)$ has directional dependence, then its type II error may be better for the boosted directions, but will be poorer when averaged over all directions, compared to an L_2 measure.

2. The information matrix of equation 3.93 is realized by the Hellinger and Bhattacharyya distances of 3.40, the symmetrized Kullbeck-Leibler distance of equation 3.35, and the chi-square distance of equation 3.42, in the sense that all these functions reduce to the quadratic form $(\hat{p} - \vec{p}\,)^T n S(\vec{p}\,)(\hat{p} - \vec{p}\,)$ for $\hat{p} \approx \vec{p}$.

3. As we add bins to a histogram, eventually the type II error will increase (section 3.7). This happens when a new bin's contribution to the offset of equation 3.83 is not enough to counteract the increased degrees of freedom (see Figure 3.11).

4. Sample distance measures other than those involving histograms may have better classification accuracy for specific types of parent distribution difference. Table 3.10 illustrates how the Anderson-Darling measure easily detects a shift in mean, while a cumulant-based measure is good at detecting differences in skewness and kurtosis. That, in turn, illustrates how there is no "one size fits all" for distance measures: each one needs to incorporate knowledge about the problem domain [57].

Chapter 4

Stationary ARMA time series

The purpose of this chapter is to develop a theory of distance measures for stationary, invertible ARMA time series models that are derived from observed samples. We'll start with an information-theoretic treatment, derive appropriate distance measures, and then illustrate their use with synthetic and real-life time series.

4.1 Generalized distance measure

In section 3.2 of chapter 3 on histograms, I mentioned that for histograms that are almost equal, the information-theoretic ideal quadratic form for histogram discrimination is based on the information matrix

$$I(\vec{p}) = n\left[\text{diag}(\frac{1}{p_1}, \ldots \frac{1}{p_{c-1}}) + \frac{1}{p_c} J_{c-1}\right] \tag{4.1}$$

where $\vec{p}$ is the normalized histogram, c is the number of bins, J_m is the all-ones matrix of size m, and n is the sample size (see 3.18). That ideal quadratic form is realized by a family of measures based on the F-divergence [6]

$$T = \sum_{\text{events v}} f\left(\frac{p(v)}{q(v)}\right) \qquad \text{with } f(1) = 0, \frac{d^2 f}{dx^2}(1) > 0 \tag{4.2}$$

where $\{v\}$ represents the decomposition of the sample into independent events. Suppose now that we have a sample of n observed events $\{v_1, \ldots v_n\}$, and that these events are

independent to order $1/n$, meaning that their correlation is at most of that order. Suppose also that that we have two possible models, M_1 and M_2, that can generate those events. Then a distance measure between M_1 and M_2 that has the form of a F-divergence is

$$T = \sum_{\text{events s}} f\left(\frac{p(v_s, M_1)}{p(v_s, M_2)}\right) \tag{4.3}$$

where $f(1) = 0, d^2f/dx^2(1) > 0$, and $p(v_s, M_j)$ is the probability of event v_s under model M_j. The summation is over all events in the sample. We would like the target T to treat the probability ratio $p(v_s, M_1)/p(v_s, M_2)$ symmetrically, in other words

$$f\left(\frac{p(v_s, M_1)}{p(v_s, M_2)}\right) = f\left(\frac{p(v_s, M_2)}{p(v_s, M_1)}\right) \tag{4.4}$$

One way is to take the function f as an L_k distance on $\log p(v_s, M_j)$:

$$D^k = \sum_s \left|\log \frac{p(v_s, M_1)}{p(v_s, M_2)}\right|^k \tag{4.5}$$

The simplest distance that leads to a differentiable optimization target is

$$D^2 = \sum_s \left[\log \frac{p(v_s, M_1)}{p(v_s, M_2)}\right]^2 \tag{4.6}$$

In words: for each independent (or nearly independent) observation v_s, compute the relative surprisal $\log\left(p(v_s, M_1)/p(v_s, M_2)\right)$, square it, and sum them all up.

Let's show that this gives meaningful results for histograms. Suppose that we have two sample histograms, $\{\hat{p}_1, \ldots \hat{p}_c\}$ and $\{\hat{q}_1, \ldots \hat{q}_c\}$ on the same set of bins, with sample sizes $n_1 = n_2 = n$. Then

$$T = \sum_s \log^2 \frac{p(v_s, M_1)}{p(v_s, M_2)} \approx \sum_{j=1}^{c} (\log^2 \frac{\hat{p}_j}{\hat{q}_j})(n\bar{p}_j) \tag{4.7}$$

Under the null hypothesis that $\hat{p}_j$ and $\hat{q}_j$ are samples from the same distribution, we may take

$$\bar{p}_j = \frac{n_1}{n_1 + n_2}\hat{p}_j + \frac{n_2}{n_1 + n_2}\hat{q}_j \tag{4.8}$$

If $n \gg 1$, then under the null hypothesis $\hat{p}_j \approx \hat{q}_j$, and we get

$$T \approx \sum_{j=1}^{c} \left(\frac{\hat{p}_j - \hat{q}_j}{\bar{p}_j}\right)^2 (\bar{p}_j n) = n \sum_{j=1}^{c} \frac{(\hat{p}_j - \hat{q}_j)^2}{\bar{p}_j} \tag{4.9}$$

which is just the two-sample version of equation 3.20.

Suppose we take the approximation $\hat{p}_j \approx \hat{q}_j$ only "halfway", that is,

$$T \approx \sum_{j=1}^{c} [\frac{\hat{p}_j - \hat{q}_j}{\bar{p}_j}](\log \frac{\hat{p}_j}{\hat{q}_j})(n\bar{p}_j) = n \left[\sum_{j=1}^{c} \hat{p}_j \log \frac{\hat{p}_j}{\hat{q}_j} + \sum_{j=1}^{c} \hat{q}_j \log \frac{\hat{q}_j}{\hat{p}_j} \right] \tag{4.10}$$

In this case, we recover the symmetrized Kullbeck-Leibler distance of equation 3.35. That makes the expression of equation 4.6 plausible, but we still need to show its underlying value (i.e. some connection to reality). Suppose that $p(v_s, M_1)$ and $p(v_s, M_2)$ are both functions of a parameter set θ, and that the differences between M_1 and M_2 are due to small changes in θ, that is

$$\frac{|\theta_1 - \theta_2|}{|\theta_1 + \theta_2|} \ll 1 \tag{4.11}$$

Then, setting

$$l_k = -\log p(v_k, \theta) = \text{ surprisal of event } k \text{ under model } \theta \tag{4.12}$$

we get

$$\begin{aligned} \log \frac{p(v_k, \theta_1)}{p(v_k, \theta_2)} &\approx (\frac{\partial l_k}{\partial \theta})^T (\Delta\theta) + \frac{1}{2}(\Delta\theta)^T \frac{\partial^2 l_k}{\partial \theta^2}(\Delta\theta) \\ \Rightarrow \ \log^2 \frac{p(v_k, \theta_1)}{p(v_k, \theta_2)} &\approx (\Delta\theta)^T (\frac{\partial l_k}{\partial \theta})(\frac{\partial l_k}{\partial \theta})^T (\Delta\theta) \end{aligned} \tag{4.13}$$

to second order in $\Delta\theta$. Substituting in equation 4.7, we get

$$\begin{aligned} \sum_s \log^2 \frac{p(v_s, \theta_1)}{p(v_s, \theta_2)} &\approx \sum_{i,j} \Delta\theta_i \sum_s \frac{\partial p_s}{\partial \theta_i} \frac{\partial p_s}{\partial \theta_j} \frac{1}{p_s^2} \Delta\theta_j \quad \text{where } p_s = p(v_s, \theta) \\ &\approx n \sum_{i,j} \Delta\theta_i \left[\sum_x \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \frac{1}{p_x} \right] \Delta\theta_j \quad \text{where } p_x = p(v_x, \theta) \end{aligned} \tag{4.14}$$

where x now represents a distinct, independent (or nearly independent) event that is possibly observable. Thus, for large samples, the curvature of

$$T = \sum_s \log^2 \frac{p(v_s, \theta_1)}{p(v_s, \theta_2)} \tag{4.15}$$

is proportional to

$$[M]_{ij} = \sum_x \frac{1}{p(x,\theta)} \frac{\partial p(x,\theta)}{\partial \theta_i} \frac{\partial p(x,\theta)}{\partial \theta_j} \tag{4.16}$$

where the sum is over all observable events x. Let's see why that is significant. We started off with a sample $\{v_1, \ldots v_n\}$ of quasi-independent events (i.e. events whose correlation is at most of order $1/n$). The total sample surprisal is

$$\mathcal{L} = \sum_s -\log p(v_s, \theta) \tag{4.17}$$

where the parameter set θ describes the true underlying model. The expectation of the Hessian (i.e. matrix of second derivatives) of the total sample surprisal is

$$[Q]_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \langle \sum_s -\log p(v_s, \theta) \rangle = -n \langle \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(v_x, \theta) \rangle \tag{4.18}$$

where the expectation is over all possible nearly independent events. But, we have that

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_x = \frac{1}{p_x} \frac{\partial^2 p_x}{\partial \theta_i \partial \theta_j} - \frac{1}{p_x^2} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \quad \text{where } p_x = p_x(\theta) = p(v_x, \theta) \tag{4.19}$$

and consequently

$$\begin{aligned} -\langle \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_x \rangle &= \langle \frac{1}{p_x^2} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \rangle - \langle \frac{1}{p_x} \frac{\partial^2 p_x}{\partial \theta_i \partial \theta_j} \rangle \\ &= \langle \frac{1}{p_x^2} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \rangle - \frac{\partial^2}{\partial \theta_i \partial \theta_j} \sum_x p_x(\theta) \\ &= \sum_x \frac{1}{p_x} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \quad \text{Note: } \sum_x p_x(\theta) = 1 \end{aligned} \tag{4.20}$$

where the sum is over all quasi-independent events. So, our main result becomes

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} = n \sum_x \frac{1}{p_x(\theta)} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \quad \text{where } p_x = p(v_x, \theta) \tag{4.21}$$

That result is true in general, and more specifically where $\partial \mathcal{L}/\partial \theta = 0$. We took an expectation over all samples, so the result depends on the parameter set θ. If the total sample surprisal $\mathcal{L}$ is minimized at a particular parameter set θ_0, then

$$\left. \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right|_{\theta_0} = n \sum_x \left. \frac{1}{p_x(\theta)} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \right|_{\theta_0} \quad \text{where } \left. \frac{\partial \mathcal{L}}{\partial \theta} \right|_{\theta_0} = 0. \tag{4.22}$$

However, the proposed measure

$$T = \sum_s \log^2 \frac{p(v_s, \theta_1)}{p(v_s, \theta_2)} \tag{4.23}$$

has the same curvature as the total surprisal $\mathcal{L}$, where $\partial\mathcal{L}/\partial\theta = 0$. So, if we can calculate a closed form or approximation for

$$Q_{ij} = \sum_x \frac{1}{p_x(\theta)} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \quad \text{where } p_x = p(v_x, \theta) \tag{4.24}$$

for a suitable set of nearly independent events $\{v_x\}$, then we will have a distance measure for θ that automatically achieves the desired quadratic form for neighboring parameter sets θ_1 and θ_2. That quadratic form would be $(\theta_1 - \theta_2)^T Q(\theta_1 - \theta_2)$.

4.2 ARMA process description

An autoregressive-moving average time series is a series of sequential numerical observations, made without random noise, but depending on a parallel series of identically distributed random innovations [15] [18]. Let μ represent the long-term average of the series. Then the defining equation is

$$x_t - a_1 x_{t-1} - \ldots - a_p x_{t-p} = e_t - c_1 e_{t-1} - \ldots - c_q e_{t-q} \tag{4.25}$$

where the integer pair (p, q) is the order of the ARMA model, $x_t = y_t - \mu$, $\{y_t\}$ is the observed series, $e_t \sim IID(0, v_e)$, and all quantities in the defining equation are real. The innovations $\{e_t\}$ are the only source of randomness [15]. They are independent, are identically distributed, have zero mean, and have variance v_e. An ARMA model has four parts: the autoregressive coefficients $\{a_1 \ldots a_p\}$, the moving average coefficients $\{c_1 \ldots c_q\}$, the innovation variance v_e, and finally the long-term average μ. In this chapter, I will assume that the ARMA models under scrutiny are stationary and invertible. In practice, that means that when we compute the factorizations

$$\begin{aligned} (1 - a_1 B - a_2 B^2 \cdots - a_p B^p) &= (1 - r_1 B) \ldots (1 - r_p B) \\ (1 - c_1 B - c_2 B^2 \cdots - c_q B^q) &= (1 - s_1 B) \ldots (1 - s_q B) \end{aligned} \tag{4.26}$$

then the AR roots $\{r_i\}$ and MA roots $\{s_j\}$ have magnitude less than 1. Also, we must exclude from the model any root that occurs as both an AR root and an MA root. The

typical goals of ARMA model estimation are to forecast future values of the series with minimal variance, and to design a feedback control scheme to minimize the variance of the $\{y_t\}$ [14].

The AR(2) models (meaning ARMA(2,0) models) are simple enough that we may express their parameter covariance in closed form, so let's look at those briefly. For an AR(p) series, let a bundle of p successive mean-reduced observations be denoted by $\vec{x}_t = [x_t \dots x_{t-p+1}]^T$, and let $E[\bullet]$ denote the expectation over an ensemble of time series for the specific AR(p) model. Furthermore, define

$$\begin{aligned} \Gamma(h) = \Gamma(-h)^T = E[\vec{x}_t \vec{x}_{t-h}^T] \quad & \text{NB depends only on } \{a_1 \dots a_p\} \text{ and } v_e \\ A = \begin{bmatrix} \vec{a} \\ I_{p-1} \quad \vec{0} \end{bmatrix} \quad & \text{NB } \vec{a} \text{ is a row vector, } \vec{0} \text{ is a column vector} \end{aligned} \tag{4.27}$$

Then, from the defining equation of the model (Eqn 4.25 with $q = 0$) we get

$$\Gamma(h) = A\Gamma(h-1) + v_e J_{11}(p)\delta_{h0} \tag{4.28}$$

where $J_{11}(p)$ represents a p by p matrix with a one in the upper left corner, and zeros elsewhere. Thus we have

$$\begin{aligned} \Gamma(0) &= A\Gamma(0)A^T + v_e J_{11}(p) \\ \Rightarrow \text{vec } \Gamma(0) &= (I_{p^2} - A \bigotimes A)^{-1}[v_e \, 0 \dots 0]^T \end{aligned} \tag{4.29}$$

where the "vec" operation, when applied to a matrix, means to assemble its columns (in order) into one long column vector. $\Gamma(0)$ is the asymptotic covariance of a bundle of p successive mean-reduced observations (asymptotic as series size $n \to \infty$), and is important because the covariance of the AR parameter estimates depends on its inverse. In particular,

$$\Gamma(0) = \begin{bmatrix} \gamma(0) & \dots & \gamma(p-1) \\ \dots & & \dots \\ \gamma(p-1) & \dots & \gamma(0) \end{bmatrix} \tag{4.30}$$

where

$$\gamma(h) = E[x_t x_{t-h}] = E[(y_t - \mu)(y_{t-h} - \mu)] \tag{4.31}$$

The well-known body of AR(p) estimation theory [89] [77] gives

$$\text{cov}(\hat{a}^T - \vec{a}^T) = \frac{v_e}{n}\Gamma^{-1}(0), \quad \hat{a} = [\hat{a}_1 \dots \hat{a}_p], n \to \infty \tag{4.32}$$

where the sample size n approaches infinity and $\hat{a}$ is the set of parameters that minimizes the residual sample variance

$$T = \frac{1}{n}(\vec{x} - \hat{a}Z)(\vec{x}^T - Z^T\hat{a}^T) \quad \text{where } \vec{x} = [x_1 \dots x_n] \quad \text{and } Z = [\vec{x}_0 \dots \vec{x}_{n-1}] \tag{4.33}$$

For a fuller development of AR(p) estimation, see [77]. The AR(2) system is small enough that equation 4.29 may be solved directly, which yields

$$\Gamma(0) = \frac{1}{\Delta}\begin{bmatrix} 1-a_2 & a_1 \\ a_1 & 1-a_2 \end{bmatrix} v_e, \quad \Delta = (1+a_2)(1-a_1-a_2)(1+a_1-a_2) \tag{4.34}$$

Now consider $\vec{a} = [a_1\ a_2]$ as a function of $\vec{r} = [r_1\ r_2]$, where $a_1 = r_1 + r_2$ and $a_2 = -r_1 r_2$. Also, let $\hat{r} = [\hat{r}_1\ \hat{r}_2]$ be the AR root estimates corresponding to the parameter estimates $\hat{a} = [\hat{a}_1\ \hat{a}_2]$. Then

$$\begin{aligned}
\text{cov}(\hat{a}^T - \vec{a}^T) &= \left(\frac{\partial \vec{a}}{\partial \vec{r}}\right) \text{cov}(\hat{r}^T - \vec{r}^T) \left(\frac{\partial \vec{a}}{\partial \vec{r}}\right)^T \\
&= \begin{bmatrix} 1 & 1 \\ -r_1 & -r_2 \end{bmatrix} \text{cov}(\hat{r}^T - \vec{r}^T) \begin{bmatrix} 1 & -r_2 \\ 1 & -r_1 \end{bmatrix} \\
\Rightarrow \text{cov}^{-1}(\hat{r}^T - \vec{r}^T) &= \begin{bmatrix} 1 & -r_2 \\ 1 & -r_1 \end{bmatrix} \text{cov}^{-1}(\hat{a}^T - \vec{a}^T) \begin{bmatrix} 1 & 1 \\ -r_1 & -r_2 \end{bmatrix} \\
&= \frac{n}{v_e}\begin{bmatrix} 1 & -r_2 \\ 1 & -r_1 \end{bmatrix} \frac{1}{\Delta} \begin{bmatrix} 1+r_1 r_2 & r_1 + r_2 \\ r_1 + r_2 & 1 + r_1 r_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -r_1 & -r_2 \end{bmatrix} v_e \\
&= n \begin{bmatrix} 1/(1-r_1^2) & 1/(1-r_1 r_2) \\ 1/(1-r_1 r_2) & 1/(1-r_2^2) \end{bmatrix}
\end{aligned} \tag{4.35}$$

since $\Delta = (1-r_1^2)(1-r_2^2)(1-r_1 r_2)$. Thus, for an AR(2) system, the covariance of the AR root estimates $[\hat{r}_1\ \hat{r}_2]$ takes a particularly simple form. The general case for an ARMA(p, q) model, with $\theta = [r_1 \dots r_p\ s_1 \dots s_q]$ being the true ARMA roots and $\hat{\theta} = [\hat{r}_1 \dots \hat{r}_p\ \hat{s}_1 \dots \hat{s}_q]$ being the corresponding sample estimates, is

$$\frac{1}{n}[\text{cov}^{-1}(\hat{\theta}^T - \vec{\theta}^T)]_{ij} = \begin{cases} 1/(1-\theta_i\theta_j) \text{ if both are AR or MA} \\ -1/(1-\theta_i\theta_j) \text{ if one is AR, one is MA} \end{cases} \tag{4.36}$$

In Appendix D, I prove that formula using the series innovations, and in Appendix E, I obtain the same result using equation 4.24 and taking the sample periodogram variates as the nearly independent quantities. Equation 4.36 does not depend on the distribution of the series innovations, but does require that their fourth moment exist. The interested reader will find more about this result in Box & Jenkins [17].

4.3 Prediction of ARMA(p, q) parameter covariance

We may use equation 4.36 "in reverse", that is, from the easily computable covariance of the ARMA roots, we may calculate the covariance of the ARMA parameters themselves. For example, let's represent an ARMA(p, q) model as

$$\begin{aligned}(1 - a_1B - a_2B^2 \ldots - a_pB^p)x_t &= (1 - c_1B - c_2B^2 \ldots - c_qB^q)e_t \\ \Rightarrow (1 - r_1B)\ldots(1 - r_pB)x_t &= (1 - s_1B)\ldots(1 - s_qB)e_t\end{aligned} \tag{4.37}$$

where $\{r_1 \ldots r_p \; s_1 \ldots s_q\}$ are all less than 1 in magnitude, $e_t \sim IID(0, v_e)$, and B is the backshift operator defined as $Bx_t = x_{t-1}$. We'll collect the model ARMA parameters into a vector $\vec{g} = [\vec{a} \; \vec{c}]^T = [a_1 \ldots a_p \; c_1 \ldots c_q]^T$, and put the associated ARMA roots into another vector $\vec{\theta} = [\vec{r} \; \vec{s}]^T = [r_1 \ldots r_p \; s_1 \ldots s_q]^T$. Then

$$\frac{\partial \vec{g}}{\partial \vec{\theta}} = \begin{bmatrix} \partial \vec{a}/\partial \vec{r} & 0 \\ 0 & \partial \vec{c}/\partial \vec{s} \end{bmatrix}, \quad \text{cov}(\hat{g} - \vec{g}) = \left(\frac{\partial \vec{g}}{\partial \vec{\theta}}\right)^T \text{cov}(\hat{\theta} - \vec{\theta}) \left(\frac{\partial \vec{g}}{\partial \vec{\theta}}\right) \tag{4.38}$$

where $\hat{g} = [\hat{a}_1 \ldots \hat{a}_p \; \hat{c}_1 \ldots \hat{c}_q]^T$ is the set of ARMA parameters that minimizes the residual sample innovation variance, acccording to the Whittle estimator (see section 2.4). That residual variance is given by

$$\begin{aligned} T &= \int_0^{2\pi} \frac{\hat{I}(\omega)}{\text{PSD}(\omega)} \frac{\text{d}\omega}{2\pi} \quad \text{where } \omega \text{ is the discrete frequency,} \\ \hat{I}(\omega) &= \sum_{-n+1}^{n-1} \hat{\gamma}(h) e^{-jh\omega} \quad \text{(sample periodogram)} \\ \hat{\gamma}(h) &= \frac{1}{n} \sum_{j=1}^{n-h} x_j x_{j+h} \quad \text{(sample covariance)} \\ \text{PSD}(\omega) &= \frac{g(s_q, \omega)\ldots g(s_1, \omega)}{g(r_p, \omega)\ldots g(r_1, \omega)} \quad \text{(normalized power spectral density)} \\ g(a, \omega) &= (1 - ae^{j\omega})(1 - ae^{-j\omega}) = 1 + a^2 - 2a\cos(\omega) \end{aligned} \tag{4.39}$$

I tested the validity of equation 4.36 with the following diagnostic: choose a sample size n and number of trials n_T. For each trial, synthesize an ARMA(p, q) series of length n, based on a known model $\{\vec{a}, \; \vec{c}, \; v_e\}$ and discarding enough initial values to ensure stationarity. Find the model parameters that minimize the residual variance of equation 4.39. After the trials are done, compute the empirical variance of those parameter estimates, with

Table 4.1: Near agreement between experimental and predicted ARMA parameter covariances for an ARMA(2,1) model $\{\vec{a},\ \vec{c},\ v_e\} = \{[0.6\ -0.08],\ [-0.3],\ 1\}$ with $n = 4096$ and $n_T = 2048$. Batch size is 277.

n * **covariance**	**theoretical**	**mean**	**95% confidence interval**
a_1, a_1	10.98	11.09	[11.04, 11.14]
a_1, a_2	-7.87	-7.96	[-7.99, -7.93]
a_1, c_1	10.22	10.36	[10.32, 10.40]
a_2, a_2	6.36	6.42	[6.39, 6.45]
a_2, c_1	-7.50	-7.60	[-7.63,-7.57]
c_1, c_1	10.47	10.63	[10.59, 10.67]

respect to the true parameters. Run several hundred batches, then compare the average and spread to the predicted covariance of equation 4.35. Table 4.1 shows the results I got with $n = 4096,\ n_T = 2048,\ \{\vec{a},\ \vec{c},\ v_e\} = \{[0.6\ -0.08],\ [-0.3],\ 1\}$ and a batch size of 277.

Table 4.1 shows that all the empirical parameter covariances are biased away from zero. That is not entirely unexpected, and is due to the nature of the Whittle estimator that I am using (equation 4.39, see also [113]). Suppose, for example, that we are estimating an AR(p) autoregressive process. If n is the sample size and $\{x_1, \ldots x_n\}$ are the mean-reduced observations, then the Whittle estimator has an embedded observation covariance estimate

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h} x_j x_{j+h} \qquad \text{(sample covariance)} \tag{4.40}$$

which is biased toward zero by a factor $(1 - h/n)$. Thus, the observation covariance matrix

$$\hat{\Gamma}(0) = \begin{bmatrix} \hat{\gamma}(0) & \ldots & \hat{\gamma}(p-1) \\ \ldots & & \ldots \\ \hat{\gamma}(p-1) & \ldots & \hat{\gamma}(0) \end{bmatrix} \tag{4.41}$$

is on the whole biased toward zero by some factor $(1 - u/n)$ where u is a multiple of the autoregressive order p. Note that the observation covariance matrix $\hat{\Gamma}(0)$ is specific to a sample, and depends on the sample size n. In any case, the AR parameter covariance

$$\text{cov}(\hat{a}^T - \vec{a}^T) = \frac{v_e}{n}\hat{\Gamma}^{-1}(0), \quad \hat{a} = [\hat{a}_1 \ldots \hat{a}_p] \tag{4.42}$$

where v_e is the innovation variance, is biased away from zero by the factor $(1 + u/n)$. As a diagnostic, I performed a parameter covariance study on an ARMA(1,1) model, using

the same Whittle estimator, but varying the sample size from $n = 480$ to $n = 2400$. The model was ARMA(1,1)$[a = 0.4, c = -0.3]$, number of trials is 2048, and number of runs per sample size is 400. Figure 4.1 shows the results along with the theoretical parameter covariances predicted by equation 4.38. The figure shows that there is indeed a bias away from zero that is approximately linear in $1/n$ for large sample size n. Thus, equation 4.38 gives us the parameter covariance for ARMA models when $n \to \infty$.

4.4 Distance measure from precision matrix

Equation 4.36 gives us a precision matrix for the roots of a stationary, invertible ARMA time series model, and from that we can formulate a corresponding quadratic form and integrate it to get our first candidate distance measure for ARMA(p, q) models. The quadratic form corresponding to equation 4.36 is

$$D^2 = \sum_{i,j}^{p} \frac{\Delta r_i \Delta r_j}{1 - r_i r_j} + \sum_{i,j}^{q} \frac{\Delta s_i \Delta s_j}{1 - s_i s_j} - \sum_{i}^{p} \sum_{j}^{q} \frac{\Delta r_i \Delta s_j}{1 - r_i s_j} \tag{4.43}$$

For the moment, we'll concentrate on the first term. It is

$$\begin{aligned} \sum_{i,j}^{p} \frac{\Delta r_i \Delta r_j}{1 - r_i r_j} &= \sum_{i,j}^{p} \Delta r_i \Delta r_j + \sum_{i,j}^{p} r_i r_j \Delta r_i \Delta r_j + \sum_{i,j}^{p} r_i^2 r_j^2 \Delta r_i \Delta r_j \ \ldots \\ &= (\Delta \sum_{i}^{p} r_i)^2 + (\frac{1}{2} \Delta \sum_{i}^{p} r_i^2)^2 + (\frac{1}{3} \Delta \sum_{i}^{p} r_i^3)^2 + \ \ldots \end{aligned} \tag{4.44}$$

which suggests that we construct a vector

$$v(\vec{r}) = [\sum_{i}^{p} r_i \quad \frac{1}{2} \sum_{i}^{p} r_i^2 \quad \frac{1}{3} \sum_{i}^{p} r_i^3 \ \ldots] \tag{4.45}$$

in which case we would have for that first term

$$[\Delta v(\vec{r})]^2 = \sum_{i,j}^{p} \frac{\Delta r_i \Delta r_j}{1 - r_i r_j} \tag{4.46}$$

In order to match equation 4.36 exactly, we would need

$$v(\vec{r}, \vec{s}) = [(\sum_{i}^{p} r_i - \sum_{j}^{q} s_j) \quad \frac{1}{2}(\sum_{i}^{p} r_i^2 - \sum_{j}^{q} s_j^2) \quad \frac{1}{3}(\sum_{i}^{p} r_i^3 - \sum_{j}^{q} s_j^3) \ \ldots] \tag{4.47}$$

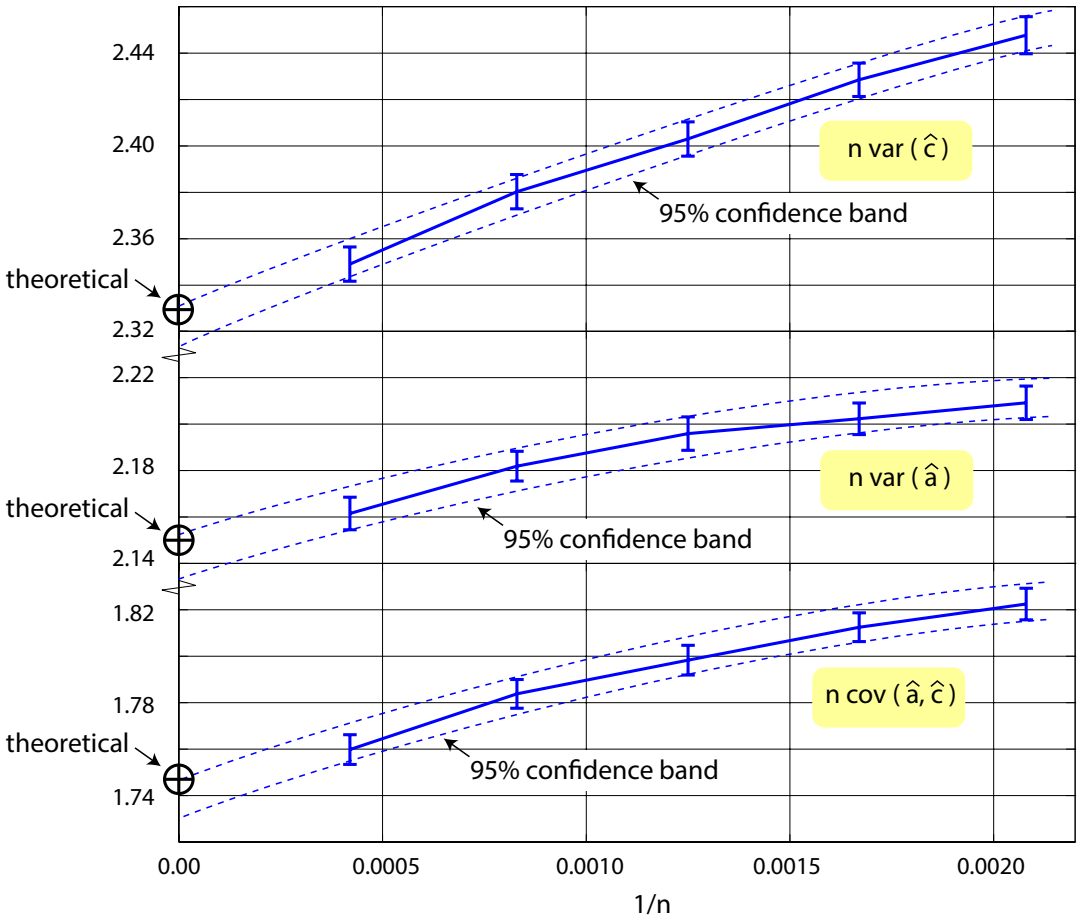


Figure 4.1: Experimental ARMA parameter covariance for an ARMA(1,1) model $\{a = 0.4, c = -0.3\}$ with varying n, $n_T = 2048$, and batch size 400.

Hence, we can define a distance measure between a stationary ARMA(p_1, q_1) model $\vec{\theta}_1 = [\vec{r}_1\ \vec{s}_1]$ and another ARMA(p_2, q_2) model $\vec{\theta}_2 = [\vec{r}_2\ \vec{s}_2]$ as

$$\begin{aligned}\frac{1}{n}D^2(\theta_1, \theta_2) &= [v(\vec{r}_1,\ \vec{s}_1) - v(\vec{r}_2,\ \vec{s}_2)]^2 \\ &= \sum_{k=1}^{\infty} \frac{1}{k^2}\left[\sum_{i=1}^{p_1} r_{1i}^k - \sum_{j=1}^{q_1} s_{1j}^k - \sum_{i=1}^{p_2} r_{2i}^k + \sum_{j=1}^{q_2} s_{2j}^k\right]^2\end{aligned} \tag{4.48}$$

Note that the ARMA models do not have to be of the same order. Our final step is to allow different sample sizes. For two models close together, the precision matrix for ARMA roots is nQ where Q_{ij} is $1/(1-\theta_i\theta_j)$ for roots of the same type, and $-1/(1-\theta_i\theta_j)$ for roots of differing type (i.e. one AR, one MA), and n is the sample size (see eqn 4.36). So, the deviations $\Delta\theta = \hat{\theta} - \vec{\theta}$ will have covariance Q^{-1}/n_1 for one sample, and Q^{-1}/n_2 for the other. The deviation difference $\Delta\theta_1 - \Delta\theta_2$ will have covariance $Q^{-1}(1/n_1 + 1/n_2)$, under the assumption that the two samples are independent. So our two-sample distance measure should be

$$D_I^2(\theta_1, \theta_2) = \frac{n_1 n_2}{n_1 + n_2}\sum_{k=1}^{\infty} \frac{1}{k^2}\left[\sum_{i=1}^{p_1} r_{1i}^k - \sum_{j=1}^{q_1} s_{1j}^k - \sum_{i=1}^{p_2} r_{2i}^k + \sum_{j=1}^{q_2} s_{2j}^k\right]^2 \tag{4.49}$$

A case which shall occupy us several times is that of comparing an AR(1) model and an MA(1) model. Figure 4.2 shows a plot of the integrated ARMA distance measure between an AR(1) model with AR parameter r, and an MA(1) model with parameter s. Here, I have left out the factor $n_1 n_2/(n_1 + n_2)$ which refers to the sample sizes. Figure 4.2 shows the expected behavior, namely that the models are very similar when $r + s = 0$. I treat this case in more detail in section 4.8.

4.5 Application of ARMA distance measure to financial time series

In the world of econometrics, the important time series for freely-traded securities are the logarithmic returns, namely

$$r_t = r(t) = \log\left[\frac{\text{security price at period } t}{\text{security price at period } t-1}\right] \tag{4.50}$$

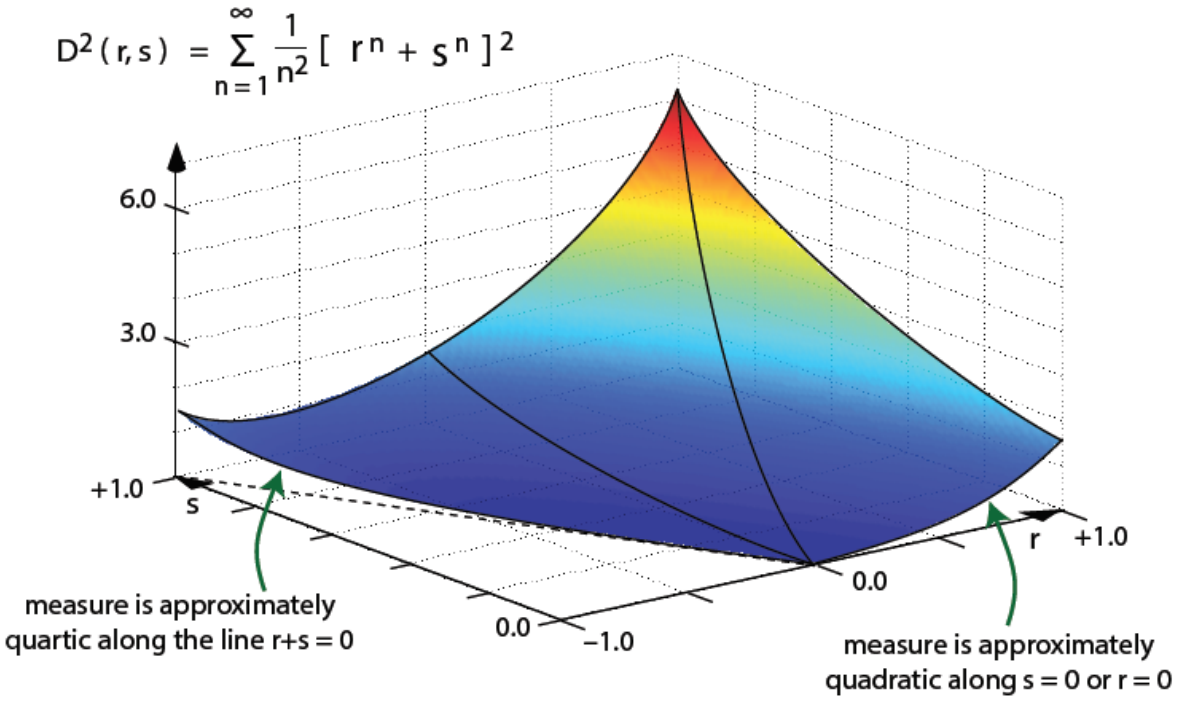


Figure 4.2: Integrated ARMA distance measure for AR(1)[r] vs MA(1)[s]

where the time period is constant, and is the one suitable for the problem under investigation. Day traders, for example, would be interested in hourly returns, whereas fund managers would be interested in daily, weekly, or even monthly returns. ARMA models are not suitable for modeling financial returns directly, since they assume a constant variance for the innovations. Actual series of financial returns show that the series variance is itself correlated, in the sense that periods of high volatility tend to persist.

Being able to predict security price movements is clearly a valuable skill, and as a result there are very sophisticated models of financial returns [63] [90]. For my purposes, I will restrict myself to GARCH(p, q) models, where GARCH stands for "generalized autoregressive conditionally heteroskedastic". A GARCH(p, q) model for financial returns would be as follows [99]:

$$\begin{aligned} r_t &= \log\left[P(t)/P(t-1)\right] \quad P(t) = \text{price at period } t \\ (1-\phi B)(r_t - \mu_r) &= x_t \quad \mu_r = \text{ mean return, } \phi \text{ is typically near zero} \\ x_t &= \sigma_t z_t \quad z_t \sim IID(0,1), \quad \sigma_t^2 = \text{ volatility} \\ \sigma_t^2 &= w + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{4.51}$$

which says that the series variance σ_t^2 itself undergoes an ARMA(p, q) process. We'll concentrate on the simplest GARCH model, namely the GARCH(1,1) model

$$\begin{aligned} x_t &= \sigma_t z_t \quad z_t \sim IID(0,1) \\ \sigma_t^2 &= w + \alpha_1 x_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned} \tag{4.52}$$

Now set $v_t = x_t^2 - \sigma_t^2 = \sigma_t^2(z_t^2 - 1)$. By equation 4.52, σ_t^2 depends on $\{z_{t-1}^2, z_{t-2}^2, \ldots\}$ but not z_t^2. Thus, σ_t^2 and z_t^2 are independent:

$$\langle v_t \rangle = \langle \sigma_t^2 \rangle \langle z_t^2 - 1 \rangle = 0 \quad \text{since } \langle z_t^2 \rangle = 1 \tag{4.53}$$

Also from equation 4.52, we get

$$\langle \sigma_t^2 \rangle = \frac{w}{1 - (\alpha_1 + \beta_1)} = \langle x_t^2 \rangle = \mu_x \quad \text{NB we need } \alpha_1 + \beta_1 \leq 1! \tag{4.54}$$

Furthermore, we also have

$$\begin{aligned} \sigma_t^2 &= w + \alpha_1 x_{t-1}^2 + \beta_1 (x_{t-1}^2 - v_{t-1}) \quad \text{NB } v_t = x_t^2 - \sigma_t^2 \\ \Rightarrow x_t^2 &= \sigma_t^2 + v_t = w + (\alpha_1 + \beta_1) x_{t-1}^2 + v_t - \beta_1 v_{t-1} \\ &\Rightarrow [1 - (\alpha_1 + \beta_1)B](x_t^2 - \mu_x) = (1 - \beta_1 B) v_t \end{aligned} \tag{4.55}$$

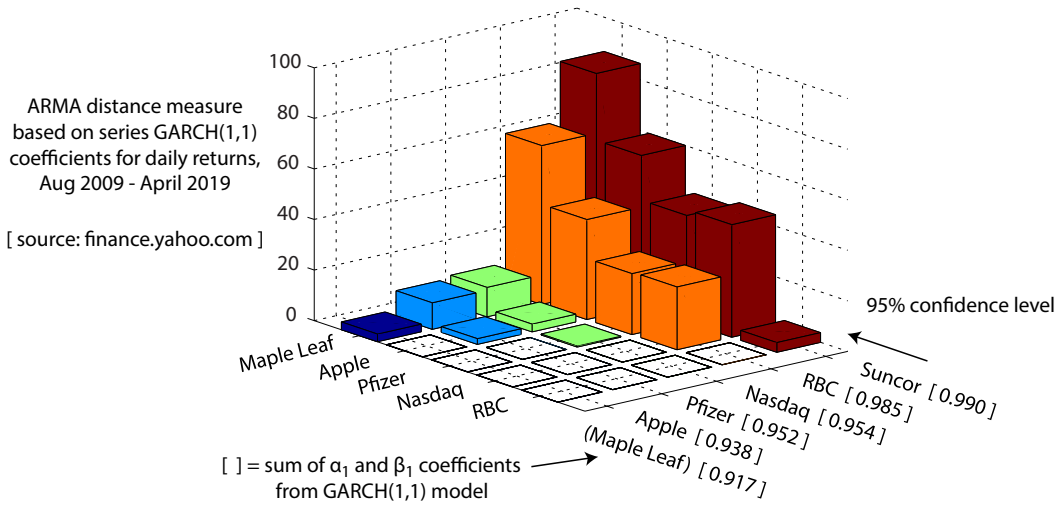


Figure 4.3: Integrated ARMA distance measure for six equity returns

which shows that $(x_t^2 - \mu_x)$ follows an ARMA(1,1) process with residuals $v_t = x_t^2 - \sigma_t^2$. The residuals are uncorrelated but not independent. They are uncorrelated since $(x_t^2 - 1)$ and $(x_{t-1}^2 - 1)$ are uncorrelated. They are not independent, though, because σ_t^2 and σ_{t-1}^2 are correlated by equation 4.52. The equivalent ARMA(1,1) process (equation 4.55) has an AR parameter $\alpha_1 + \beta_1$ and an MA parameter β_1. Using equation 4.49 as a distance measure, we can then compare two different sets of financial returns by looking at their GARCH models. For Figure 4.3, I analyzed the daily returns of six securities in different financial sectors, using the time period August 2009 - April 2019 inclusive (the source is finance.yahoo.com), and using MATLAB's econometrics toolbox to estimate the GARCH(1,1) parameters. The figure shows that with a 95% confidence level, we can reliably distinguish ten pairs of financial series over that nine-year interval, using equation 4.49 as a distance measure.

When looking at the distance measure results of Figure 4.3, a natural question is: could we have done better with another distance measure? I treat this question in depth in Appendix G, where I show that a quadratic form other than one based on the information matrix (in this case, equation 4.36) may have better classification accuracy in a specific direction, but will have poorer performance when averaged over all directions; see in particular Figures G.6 and G.7. Let's illustrate that for our GARCH(1,1) models. Martin [80]

proposed a distance measure for ARMA(p, q) models based on the cepstral coefficients, in particular

$$d_M^2 = (\text{constant}) \sum_{n=1}^{\infty} n|c_{1,n} - c_{2,n}|^2 \tag{4.56}$$

where the cepstral coefficients for ARMA model k are defined by

$$\log \text{PSD}_k(\omega) = \sum_{n \in Z} c_{k,n} e^{jn\omega} \tag{4.57}$$

and $\text{PSD}_k(\omega)$ is the power spectral density for ARMA model k. In terms of the autoregressive roots $\{r_1 \dots r_p\}$ and moving average roots $\{s_1 \dots s_q\}$,

$$\begin{aligned} \text{PSD}(\omega) &= \frac{g(s_1, \omega) \dots g(s_q, \omega)}{g(r_1, \omega) \dots g(r_p, \omega)} \quad \text{where } \omega \text{ is the discrete frequency and} \\ g(a, \omega) &= (1 - ae^{j\omega})(1 - ae^{-j\omega}) = 1 + a^2 - 2a\cos(\omega) \end{aligned} \tag{4.58}$$

Using equations 4.56, 4.57, and 4.58, and inserting a factor $n_1 n_2/(n_1 + n_2)$ to account for sample sizes, we get

$$\begin{aligned} d_M^2 &= \frac{n_1 n_2}{n_1 + n_2} (\Delta\theta)^T Q_M (\Delta\theta) \quad \text{where } \theta = [r_1 \dots r_p \, s_1 \dots s_q]^T \text{ and} \\ [Q_M]_{ij} &= \begin{cases} +1/(1 - \theta_i\theta_j)^2 \text{ if both are AR or MA} \\ -1/(1 - \theta_i\theta_j)^2 \text{ if one is AR, one is MA} \end{cases} \end{aligned} \tag{4.59}$$

Note how this differs from equation 4.36 for the Fisher information matrix for ARMA(p, q) models. For the ARMA(1,1) case, with AR parameter r and MA parameter s, the Martin distance measure becomes

$$d_M^2 = \frac{n_1 n_2}{n_1 + n_2} [\Delta r \; \Delta s] \begin{bmatrix} 1/(1 - r^2)^2 & -1/(1 - rs)^2 \\ -1/(1 - rs)^2 & 1/(1 - s^2)^2 \end{bmatrix} \begin{bmatrix} \Delta r \\ \Delta s \end{bmatrix} \tag{4.60}$$

The following algorithm translates a pair of GARCH(1,1) models into the terminology and variables of Appendix G.

1. Given: (α_1, β_1, n_1) and (α_2, β_2, n_2), where (α_k, β_k) are the GARCH(1,1) parameters for model k, and n_k is the sample size for model k.

2. Set $r_1 = \alpha_1 + \beta_1, s_1 = \beta_1, r_2 = \alpha_2 + \beta_2$, and $s_2 = \beta_2$.

3. Calculate the sample-weighted averages $r = (n_1 r_1 + n_2 r_2)/(n_1 + n_2)$ and $s = (n_1 s_1 + n_2 s_2)/(n_1 + n_2)$.

Table 4.2: Comparison of type II error rates between integrated distance measure d_I^2 and Martin distance measure d_M^2 for GARCH(1,1) models

Series	$T_{2,1}$ for d_I^2	δ	v^2	θ (degrees)	$T_{2,2}$ for d_M^2	$\langle T_{2,2}\rangle_d$
RBC/Suncor	0.599	0.953	3.843	59.7	0.828	0.719
Apple/Nasdaq	0.669	0.827	3.097	17.5	0.608	0.755

4. Find the Cholesky decomposition of the information matrix of equation 4.36, i.e. solve

$$L^T L = \begin{bmatrix} 1/(1-r^2) & -1/(1-rs) \\ -1/(1-rs) & 1/(1-s^2) \end{bmatrix} \tag{4.61}$$

5. Transform Q_M into the coordinate system in which the information matrix of 4.36 is a multiple of I_2:

$$\bar{Q} = L^{-T} \begin{bmatrix} 1/(1-r^2)^2 & -1/(1-rs)^2 \\ -1/(1-rs)^2 & 1/(1-s^2)^2 \end{bmatrix} L^{-1} \tag{4.62}$$

6. Find the eigenvectors and eigenvalues of $\bar{Q}$, i.e. solve $\bar{Q}V = VD$ where V is orthogonal and D is diagonal. The columns of V are the eigenvectors, and the entries of D are the eigenvalues λ_1 and λ_2.

7. Assuming that $\lambda_1 \geq \lambda_2$, calculate the boost as $\delta = (\lambda_1 - \lambda_2)/(\lambda_1 + \lambda_2)$.

8. The offset is

$$\vec{v} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} V^T L [\delta r \, \delta s]^T \tag{4.63}$$

where $\delta r = r_1 - r_2 = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)$, and $\delta s = s_1 - s_2 = \beta_1 - \beta_2$. From the offset $\vec{v}$, we can deduce the square magnitude $v^2 = |\vec{v}|^2$ and angle θ between $\vec{v}$ and the boosted axis.

Table 4.2 shows the results of applying that algorithm to the (RBC, Suncor) and (Apple, Nasdaq) pairs of series from Figure 4.3. For these series, the sample sizes are $n_1 = n_2 = 2453$. In Table 4.2, $T_{2,1}$ represents the type II error that we get using a quadratic form based on the information matrix of equation 4.36. $T_{2,2}$ represents the type II error that we get using the quadratic form of equation 4.60. The last column, $\langle T_{2,2}\rangle_d$, represents the type II error that we get by using equation 4.60, but averaged over all directions. I calculated

the type II error values using the Imhof procedure [61], as implemented in the R package CompQuadForm [33]. Table 4.2 reinforces the conclusions of Appendix G, namely that using a quadratic form not based on the information matrix may yield a lower type II error in specific directions, but always gives a larger type II error when averaged over all directions of the offset vector $\vec{v}$.

4.6 Issues with the ARMA precision matrix

Two questions spring immediately to mind when looking at the precision matrix for ARMA(p, q) models (equation 4.36), namely: (1) is the integrated distance measure still real when some of the AR or MA roots are complex? (2) the precision matrix becomes singular if any two AR roots or any two MA roots are the same, is this a problem?

As for the first question, the short answer is this: the AR and MA roots are the roots of polynomial equations with real coefficients (see equation 4.26), so they are either real or occur in conjugate pairs. In expressions such as $\sum r_i^k$ and $\sum s_j^k$, the imaginary parts of the conjugate pairs cancel, so the results are always real. We can do better, though. Let's express $\sum r_i^k$ explicitly in terms of the AR parameters $\{a_1 \dots a_p\}$, and $\sum s_j^k$ in terms of the MA parameters $\{c_1 \dots c_q\}$. For this, we need the Newton identities that relate power sums to symmetric polynomials [65]. We'll treat the case of the autoregressive roots $\{r_1 \dots r_p\}$, and use the notation

$$u_k = \sum_{i=1}^{p} r_i^k, \quad t_k = \sum_{1 \le j_1 < j_2 \dots < j_k \le p} r_{j_1} \dots r_{j_k} \tag{4.64}$$

For $p = 3$, for example, we would have

$$\begin{aligned} u_1 &= t_1 = r_1 + r_2 + r_3 \\ u_2 &= r_1^2 + r_2^2 + r_3^2, \quad t_2 = r_1 r_2 + r_1 r_3 + r_2 r_3 \\ u_3 &= r_1^3 + r_2^3 + r_3^3, \quad t_3 = r_1 r_2 r_3 \end{aligned} \tag{4.65}$$

Note that in the notation of Appendix B,

$$t_k = \frac{1}{k!} \sum_{1 \le j_1 \ne j_2 \dots \ne j_k \le p} r_{j_1} \dots r_{j_k} = \frac{1}{k!} m_{11\dots1} \quad (k \text{ ones }) \tag{4.66}$$

Our goal is to express the $\{u_k\}$ in terms of the $\{t_k\}$. The Newton identities state that

$$k t_k = \sum_{i=1}^{k} (-1)^{i-1} t_{k-i} u_i \quad \text{with } t_0 = 1 \tag{4.67}$$

From these, we get

$$\begin{aligned} u_1 &= t_1, \quad u_2 = t_1 u_1 - 2t_2 = t_1^2 - 2t_2 \\ u_3 &= t_1 u_2 - t_2 u_1 + 3t_3 = t_1^3 - 3t_2 t_1 + 3t_3 \end{aligned} \tag{4.68}$$

and so forth. From the defining equation 4.26, we also have $a_j = (-1)^{j-1} t_j$. Therefore, we get

$$\begin{aligned} \sum_{i=1}^{p} r_i = t_1 = a_1, \quad \sum_{i=1}^{p} r_i^2 = t_1^2 - 2t_2 = a_1^2 + 2a_2 \\ \sum_{i=1}^{p} r_i^3 = t_1^3 - 3t_2 t_1 + 3t_3 = a_1^3 + 3a_2 a_1 + 3a_3 \end{aligned} \tag{4.69}$$

and so forth. Clearly, all the power sums $\sum r_i^k$ are expressible as polynomials in $\{a_1 \ldots a_p\}$ and are real. The same reasoning gives us $\sum s_j^k$ in terms of the MA parameters $\{c_1 \ldots c_q\}$, with $\sum s_j = c_1, \sum s_j^2 = c_1^2 + 2c_2, \sum s_j^3 = c_1^3 + 3c_2 c_1 + 3c_3$, and so on. Thus, the integrated ARMA distance measure of equation 4.49 is expressible in terms of polynomials involving the ARMA parameters of both models.

As for the second question, involving repeated AR or MA roots: it is true that the precision matrix of equation 4.36 becomes singular when $r_i = r_j$ or $s_i = s_j$ for $i \neq j$. We excluded the case $r_i = s_j$ when we defined ARMA models, so that these models would be unique. However, the precision matrix for the ARMA coefficients themselves does not become singular. I will illustrate this for the AR(p) case. Here, the covariance matrix for the parameter estimates is [77]

$$\mathrm{cov}(\hat{a}^T - \vec{a}^T) = \frac{1}{n} \hat{\Gamma}^{-1}(0) v_e \tag{4.70}$$

where n is the finite sample size, v_e is the innovation variance, and

$$\hat{\Gamma}(0) = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T, \quad \vec{x}_i = [x_i \quad x_{i-1} \ldots x_{i-p+1}]^T \tag{4.71}$$

Note that equation 4.32 involves the theoretical value of $\hat{\Gamma}(0)$, which we obtain in the limit as $n \to \infty$, whereas in equation 4.71, n is large but finite. If $\hat{\Gamma}(0)$ had determinant zero, then there would exist a non-zero vector $\vec{u}$ such that

$$\begin{aligned} \hat{\Gamma}(0)\vec{u} = 0, \quad \text{hence } \vec{u}^T \hat{\Gamma}(0) \vec{u} = 0 \\ \Rightarrow \vec{u}^T \hat{\Gamma}(0) \vec{u} = \frac{1}{n} \sum_{i=1}^{n} (\vec{u}^T \vec{x}_i)(\vec{x}_i^T \vec{u}) = \frac{1}{n} \sum_{i=1}^{n} (\vec{u} \bullet \vec{x}_i)^2 = 0 \end{aligned} \tag{4.72}$$

But that requires that $\vec{u}$ be orthogonal to all the $\{\vec{x}_j\}$, hence the rank of $\{\vec{x}_1 \dots \vec{x}_n\}$ must be less than p. But the $\{\vec{x}_j\}$ are random vectors. The chance that n random vectors of size p do not span $\mathbb{R}^p$ for $n \gg p$ is vanishingly small; see, for example, Bourgain et al. [13]. So, for sure $\hat{\Gamma}(0)$ is non-singular, hence $\text{cov}(\hat{a}^T - \vec{a}^T)$ exists.

4.7 Approximate distribution of the integrated measure for different ARMA models

In section 4.4, I derived a distance measure for ARMA time series based upon the information matrix of equation 4.36. That integrated distance measure for ARMA samples was (see equation 4.49):

$$\begin{aligned}
\frac{n_1+n_2}{n_1 n_2} D_I^2(\theta_1,\theta_2) &= \sum_{k=1}^{\infty} \frac{1}{k^2} \left[\sum_{i=1}^{p_1} r_{1i}^k - \sum_{j=1}^{q_1} s_{1j}^k - \sum_{i=1}^{p_2} r_{2i}^k + \sum_{j=1}^{q_2} s_{2j}^k \right]^2 \\
&= \sum_{i=1}^{p_1} h(r_{1i}^2) + \sum_{j=1}^{q_1} h(s_{1j}^2) + \sum_{i=1}^{p_2} h(r_{2i}^2) + \sum_{j=1}^{q_2} h(s_{2j}^2) + \\
&\quad 2\Big[- \sum_{i=1}^{p_1}\sum_{j=1}^{q_1} h(r_{1i}s_{1j}) - \sum_{i=1}^{p_1}\sum_{m=1}^{p_2} h(r_{1i}r_{2m}) + \sum_{i=1}^{p_1}\sum_{j=1}^{q_2} h(r_{1i}s_{2j}) \\
&\quad + \sum_{j=1}^{q_1}\sum_{m=1}^{p_2} h(s_{1j}r_{2m}) - \sum_{j=1}^{q_1}\sum_{m=1}^{q_2} h(s_{1j}s_{2m}) - \sum_{i=1}^{p_2}\sum_{j=1}^{q_2} h(r_{2i}s_{2j}) \Big]
\end{aligned} \tag{4.73}$$

where the first sample has size n_1 and estimated ARMA roots $\theta_1 = \{r_{11}, \dots r_{1,p_1}, s_{11}, \dots s_{1,q_1}\}$, the second sample has size n_2 and estimated ARMA roots $\theta_2 = \{r_{21}, \dots r_{2,p_2}, s_{21}, \dots s_{2,q_2}\}$, and finally

$$h(x) = \sum_{k=1}^{\infty} \frac{1}{k^2} x^k = x + \frac{1}{4}x^2 + \frac{1}{9}x^3 + \dots \tag{4.74}$$

For reference, the important properties of $h(x)$ are:

$$\begin{aligned} h'(x) &= -\frac{1}{x}\log(1-x) \quad \text{with } h'(0) = 1 \\ h''(x) &= \frac{1}{x(1-x)} + \frac{1}{x^2}\log(1-x) \quad \text{with } h''(0) = 1/2 \\ h(-x) &= -h(x) + \frac{1}{2}h(x^2) \\ h(1) &= \zeta(2) = \frac{\pi^2}{6} \approx 1.64 \quad \zeta(s) \text{ is the Riemann zeta function} \end{aligned} \tag{4.75}$$

Thus, for example, if we have an AR(1) sample of size n and estimated AR parameter $\hat{a}$, and another AR(1) sample of size m and estimated AR(1) parameter $\hat{b}$, then

$$\left(\frac{n+m}{nm}\right) D_I^2 = h(\hat{a}^2) + h(\hat{b}^2) - 2h(\hat{a}\hat{b}) \tag{4.76}$$

which has the expected result that $D_I^2 = 0$ when $\hat{a} = \hat{b}$. However, if we have an AR(1) sample of size n and estimated AR parameter $\hat{a}$, and an MA(1) sample of size m and estimated MA(1) parameter $\hat{b}$, then

$$\left(\frac{n+m}{nm}\right) D_I^2 = h(\hat{a}^2) + h(\hat{b}^2) + 2h(\hat{a}\hat{b}) \tag{4.77}$$

We will use that last case as a test case for estimating the spread of D_I^2. Let the first model be AR(1) with parameter a, and the second model be MA(1) with parameter b. Then, by equation 4.36, we would expect

$$\text{var}(\hat{a}) = \frac{1}{n}(1-a^2), \qquad \text{var}(\hat{b}) = \frac{1}{m}(1-b^2) \tag{4.78}$$

and we would also expect the second-order ensemble mean value of D_I^2 to be

$$\begin{aligned} \left(\frac{n+m}{nm}\right) \langle D_I^2 \rangle = h(a^2) + h'(a^2)\ \text{var}(\hat{a}) + \frac{1}{2}h''(a^2)[4a^2\ \text{var}(\hat{a})] + \\ 2h(ab) + h''(ab)[b^2\ \text{var}(\hat{a}) + a^2\ \text{var}(\hat{b})] + \\ h(b^2) + h'(b^2)\ \text{var}(\hat{b}) + \frac{1}{2}h''(b^2)[4b^2\ \text{var}(\hat{b})] \end{aligned} \tag{4.79}$$

As for the expected variance of D_I^2, to second order that is

$$\text{var}(D_I^2) = \left[\frac{\partial(D_I^2)}{\partial\hat{a}}\right]_a^2 \text{var}(\hat{a}) + \left[\frac{\partial(D_I^2)}{\partial\hat{b}}\right]_b^2 \text{var}(\hat{b}) \tag{4.80}$$

where

$$\begin{aligned}\left(\frac{n+m}{nm}\right)\left[\frac{\partial(D_I^2)}{\partial\hat{a}}\right]_a &= 2ah'(a^2)+2bh'(ab)\\ \left(\frac{n+m}{nm}\right)\left[\frac{\partial(D_I^2)}{\partial\hat{b}}\right]_b &= 2bh'(b^2)+2ah'(ab)\end{aligned} \tag{4.81}$$

Figure 4.4 shows experimental results for the case just mentioned, namely AR(1)[0.5] against MA(1)[-0.5] with $n = m = 1024$. The methodology of this diagnostic is as follows: for each trial, generate two samples of the known distributions, find the estimated ARMA roots via the Whittle estimator [113], and then compute the integrated ARMA distance measure of equation 4.49. Figure 4.4 also shows a plot of the normal CDF having the mean and variance given by equations 4.79 and 4.80. The fit is good but not exact, since the distribution of D_I^2 for unequal models is similar to a non-central chi-square, not a normal distribution. By contrast, Figure 4.4 shows experimental results for the case where both samples are based on the same ARMA(2,1) model with $n = m = 2048$. In this case, we expect a chi-square(3) distribution since there are $p + q = 3$ ARMA parameters in the model. The figure shows that the Kolmogorov-Smirnov statistic is well within the 95% confidence interval.

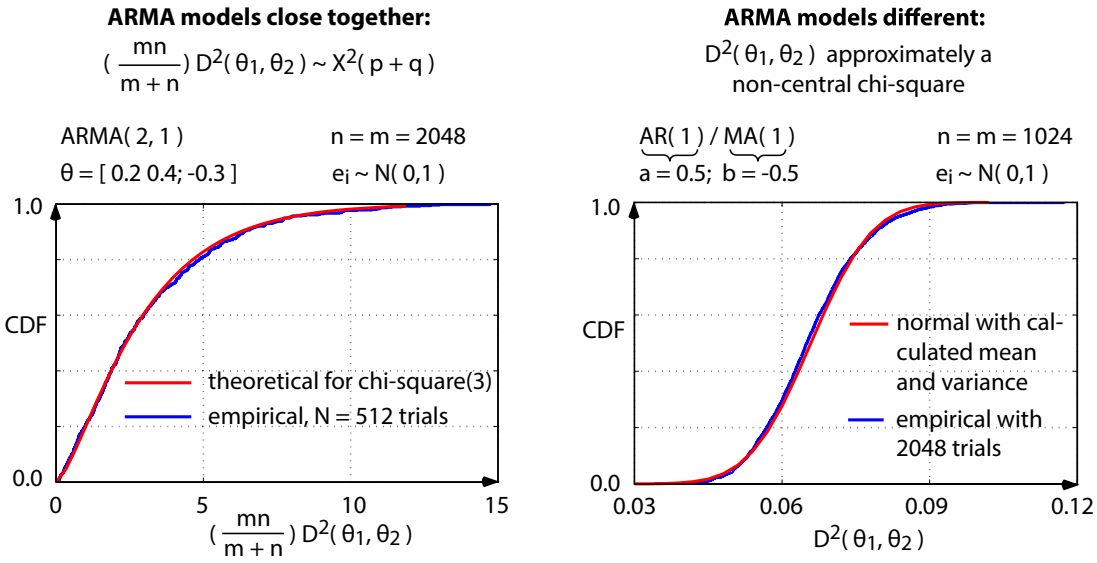
4.8 Other candidate distance measures for ARMA time series models

In section 4.2, we derived the precision matrix for ARMA(p, q) parameters estimated from a large but finite sample, and in section 4.4 we integrated the corresponding quadratic form to get a candidate distance measure for ARMA(p, q) models. However, any distance measure that takes the form

$$D^2 = \frac{n_1 n_2}{n_1+n_2}\left[\sum_{i,j=1}^{p}\frac{\Delta r_i \Delta r_j}{1-r_i r_j}+\sum_{i,j=1}^{q}\frac{\Delta s_i \Delta s_j}{1-s_i s_j}-\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\Delta r_i \Delta s_j}{1-r_i s_j}\right] \tag{4.82}$$

for models $M_1 = [\vec{r}_1, \vec{s}_1]$ and $M_2 = [\vec{r}_2, \vec{s}_2]$ that are close together, with

$$\begin{aligned}&\Delta\vec{r} = \vec{r}_1 - \vec{r}_2, \quad \Delta\vec{s} = \vec{s}_1 - \vec{s}_2, \quad p = \max(\text{size}(\vec{r}_1), \text{size}(\vec{r}_2)),\\ &q = \max(\text{size}(\vec{s}_1), \text{size}(\vec{s}_2)), \quad \vec{r} = \frac{n_1\vec{r}_1 + n_2\vec{r}_2}{n_1+n_2}, \quad \vec{s} = \frac{n_1\vec{s}_1 + n_2\vec{s}_2}{n_1+n_2}\end{aligned} \tag{4.83}$$



N	$\sqrt{N}$ [Kolmogorov Smirnov]	$\sqrt{N}$ [critical value]
512	1.09	1.35
1024	1.02	1.35
2048	0.59	1.35

N	$\sqrt{N}$ [Kolmogorov Smirnov]	$\sqrt{N}$ [critical value]
512	1.98	1.35
1024	1.20	1.35
2048	1.61	1.35
2048	1.96	1.35

Figure 4.4: Comparison of expected and actual distributions for integrated ARMA distance measure, along with the corresponding Kolmogorov-Smirnov statistics.

is a candidate distance measure. By "close together", I mean that $|\Delta\vec{r}\,|/|\vec{r}\,| \ll 1$, and $|\Delta\vec{s}\,|/|\vec{s}\,| \ll 1$. Clearly, there are an infinite set of such distance measures. When the subject ARMA models are close together, the candidate distance measures will have similar classification performance, since they reduce to the same quadratic form in that case. When the subject ARMA models are far apart, the candidate distance measures will also have similar classification performance, since any reasonable distance measure will be successful in separating models that are far apart. Consequently, if we need to choose a "best" candidate measure among those that satisfy equation 4.82, it will have to be on the basis of theoretical properties.

What might those alternative candidate measures look like? Well, first let's describe our models using their ARMA roots, i.e. $M_1 \to \theta_1 = [\vec{r}_1, \vec{s}_1]$ and $M_2 \to \theta_2 = [\vec{r}_2, \vec{s}_2]$. For a general model $\theta = [r_1 \dots r_p, s_1 \dots s_q]$, the normalized spectral density is

$$\mathrm{PSD}(\omega, \theta) = \frac{g(s_q, \omega) \dots g(s_1, \omega)}{g(r_p, \omega) \dots g(r_1, \omega)} \qquad \text{where } g(a, \omega) = 1 + a^2 - 2a\cos\omega \tag{4.84}$$

and in Appendix E, I demonstrate that

$$\frac{1}{2}\int_0^{2\pi} \log^2 \frac{\mathrm{PSD}(\omega, \theta_1)}{\mathrm{PSD}(\omega, \theta_2)} \frac{\mathrm{d}\omega}{2\pi} \approx \int_0^{2\pi} \frac{\mathrm{PSD}(\omega, \theta_1)}{\mathrm{PSD}(\omega, \theta_2)} \frac{\mathrm{d}\omega}{2\pi} - 1 \tag{4.85}$$

Since the integrated ARMA distance measure of equation 4.49 comes from the left-hand side (see Appendix E), we get right away three possible distance measures having the same curvature at $\theta_1 \approx \theta_2$:

$$\begin{aligned}
d^2_{YW} &= \left(\frac{n_1 n_2}{n_1 + n_2}\right) \left[\int_0^{2\pi} \frac{\mathrm{PSD}(\omega, \theta_1)}{\mathrm{PSD}(\omega, \theta_2)} \frac{\mathrm{d}\omega}{2\pi} - 1\right] \\
d^2_{RYW} &= \left(\frac{n_1 n_2}{n_1 + n_2}\right) \left[\int_0^{2\pi} \frac{\mathrm{PSD}(\omega, \theta_2)}{\mathrm{PSD}(\omega, \theta_1)} \frac{\mathrm{d}\omega}{2\pi} - 1\right] \\
d^2_{SYW} &= \frac{1}{2}\left(\frac{n_1 n_2}{n_1 + n_2}\right) \left[\int_0^{2\pi} \left[\frac{\mathrm{PSD}(\omega, \theta_1)}{\mathrm{PSD}(\omega, \theta_2)} + \frac{\mathrm{PSD}(\omega, \theta_2)}{\mathrm{PSD}(\omega, \theta_1)}\right] \frac{\mathrm{d}\omega}{2\pi} - 2\right]
\end{aligned} \tag{4.86}$$

where n_1 and n_2 are the sample sizes corresponding to models θ_1 and θ_2 respectively. Another idea is that of an intermediate model, in the spirit of equation 3.36: we choose an intermediate ARMA model $\bar{\theta}$ so as to minimize

$$d^2_{INT} = 2\left(\frac{n_1 n_2}{n_1 + n_2}\right) \left[\int_0^{2\pi} \left[\frac{\mathrm{PSD}(\omega, \theta_1)}{\mathrm{PSD}(\omega, \bar{\theta})} + \frac{\mathrm{PSD}(\omega, \theta_2)}{\mathrm{PSD}(\omega, \bar{\theta})}\right] \frac{\mathrm{d}\omega}{2\pi} - 2\right] \tag{4.87}$$

Here, the order of the intermediate ARMA model $\bar{\theta}$ must be sufficient to cover a linear combination of PSD(ω, θ_1) and PSD(ω, θ_2). In particular, if M_1 has order (p_1, q_1) and M_2 has order (p_2, q_2), then the intermediate model may need the ARMA order $(p_1 + p_2,\ \max(p_1 + q_2, p_2 + q_1))$ [54]. For example, if we are comparing two theoretical models AR(1)$[a]$ and AR(1)$[b]$, then we would expect the intermediate model to be an ARMA(2,1) with $\bar{M} \to \bar{\theta} = \{[a, b], c\}$. The intermediate measure of equation 4.87 is then

$$\begin{aligned} d^2_{INT} &= 2\left[\int_0^{2\pi} \frac{g(a,\omega)g(b,\omega)}{g(c,\omega)} \bullet \left[\frac{1}{g(a,\omega)} + \frac{1}{g(b,\omega)}\right]\frac{\mathrm{d}\omega}{2\pi} - 2\right] \\ &= 2\left[\int_0^{2\pi} \frac{2 + a^2 + b^2 - 2(a+b)\cos\omega}{1 + c^2 - 2c\cos\omega}\frac{\mathrm{d}\omega}{2\pi} - 2\right] \end{aligned} \tag{4.88}$$

which will be minimized when the integrand is a constant, i.e. where

$$\frac{2 + a^2 + b^2}{1 + c^2} = \frac{a + b}{c} \tag{4.89}$$

which in turn leads to

$$c = \frac{(2 + a^2 + b^2) - \sqrt{(2 + a^2 + b^2)^2 - 4(a+b)^2}}{2(a+b)} \qquad \text{assuming } a + b \neq 0 \tag{4.90}$$

with corresponding measure

$$d^2_{INT} = \frac{2}{1 - c^2}[(c - a)^2 + (c - b)^2] \tag{4.91}$$

A natural question is, would these alternative candidate measures have the same classification performance? As a diagnostic check, I took the AR(1)$[a]$ / AR(1)$[b]$ system as an example. For this system, the three symmetric distances mentioned so far are:

$$\begin{aligned} d^2_{SYW} &= \frac{n_1 n_2}{2(n_1 + n_2)}(a - b)^2\left(\frac{1}{1 - a^2} + \frac{1}{1 - b^2}\right) \\ d^2_{I} &= \frac{n_1 n_2}{n_1 + n_2}\left[h(a^2) + h(b^2) - 2h(ab)\right] \qquad (\text{ see equation } 4.74\ ) \\ d^2_{INT} &= \frac{n_1 n_2}{n_1 + n_2}\frac{2}{1 - c^2}\left[(c - a)^2 + (c - b)^2\right] \qquad \text{where} \\ c &= \frac{(2 + a^2 + b^2) - \sqrt{(2 + a^2 + b^2)^2 - 4(a+b)^2}}{2(a+b)} \end{aligned} \tag{4.92}$$

In the diagnostic test, I find the empirical T_2 value (i.e. type II error) for those three symmetric distance measures over 16000 pairs of samples, where the first sample is drawn

Table 4.3: Empirical T_2 values for the AR(1)[a] / AR(1)[b] system with $n_1 = n_2 = 2048$ over 16000 runs

Scenario	**Symmetric Yule-Walker**	**Integrated Euclidean**	**Intermediate**
0.59 : 0.61	0.8732	0.8738	0.8740
0.58 : 0.62	0.6334	0.6344	0.6347
0.57 : 0.63	0.3242	0.3251	0.3255
0.56 : 0.64	0.1121	0.1126	0.1127
0.55 : 0.65	0.0260	0.0260	0.0261

from AR(1)[a] with size n_1, and the second sample is drawn from AR(1)[b] with size n_2. I vary the $a : b$ split from 0.59:0.61 (hard to resolve at $n_1 = n_2 = 2048$) to 0.55:0.65 (easy to resolve at $n_1 = n_2 = 2048$). The results are given in Table 4.3. The tolerance for the empirical Type 2 errors may be calculated from the DKW inequality [81], and is approximately ± 0.010 for this diagnostic at the 95% confidence level.

Table 4.3 clearly shows that within the experimental tolerance, the three symmetric distance measures for the AR(1)[a] / AR(1)[b] system have identical classification performance over the chosen spread of ARMA models. Figure 4.5 indicates why this is so. For any specific distance measure, the type II error value comes from the part of the cumulative distribution function arising from models that are very similar and model distance is $(\hat{\theta}_1 - \hat{\theta}_2)^T I(\bar{\theta}\ )(\hat{\theta}_1 - \hat{\theta}_2)$, where $\bar{\theta}$ is an "average" model between $\hat{\theta}_1$ and $\hat{\theta}_2$, and $I(\bar{\theta}\ )$ is the information matrix at $\bar{\theta}$. The candidate distance measures were chosen to have this information matrix in common, hence the type II error does not depend on which specific model we are considering.

I will further illustrate the concept of alternate candidate distance measures with a particular example. We'll look at candidate distance measures between a theoretical AR(1) model with autoregressive root a, and a theoretical MA(1) model with moving average root b, and ask the question: for a given value of a, what value of b minimizes the AR(1)[a] / MA(1)[b] distance, for each of the three symmetric measures mentioned above (see equation 4.93 for example). In the case of the symmetric Yule-Walker distance,

$$\begin{aligned} T_{SYW} = 2d_{SYW}^2 &= \int_0^{2\pi} \left[\frac{1}{g(a,\omega)g(b,\omega)} + g(a,\omega)g(b,\omega) \right] \frac{\mathrm{d}\omega}{2\pi} - 2 \\ &= \frac{1}{(1-a^2)(1-b^2)} \frac{1+ab}{1-ab} + (1+a^2)(1+b^2) + 2ab - 2 \end{aligned} \tag{4.93}$$

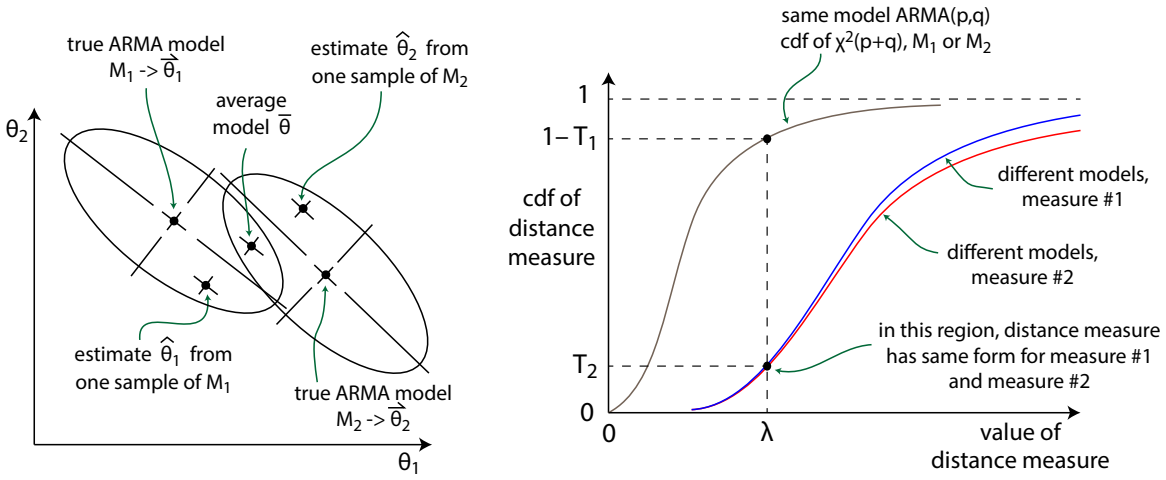


Figure 4.5: Classification performance of candidate ARMA distance measures

Minimizing T_{SYW} with respect to b gives

$$b_{SYW} = -a + a^3 + a^5 + O(a^7) \tag{4.94}$$

The integrated distance measure of equation 4.73 is for this case

$$\begin{aligned} T_I = d_I^2 &= h(a^2) + h(b^2) + 2h(ab) \quad \text{hence} \\ \frac{\partial T_I}{\partial b} &= 0 \Rightarrow (1-ab)(1-b^2) = 1 \end{aligned} \tag{4.95}$$

whose solution looks like

$$b_I = -a + a^3 - 2a^5 + O(a^7) \tag{4.96}$$

Note also that for $a = 1$, we get $b_I \approx -0.62$. In the case of the intermediate measure d_{INT}^2, the intermediate model is of ARMA(1,2) type, and the minimization target is

$$\begin{aligned} T_{INT} = d_{INT}^2 = \frac{2}{g} - 4 \quad &\text{where} \quad \frac{2 + (a+b)^2 + (ab)^2}{1 + f^2(a+b)^2 + g^2(ab)^2} = \frac{1}{g} \\ \text{and} \quad \left(\frac{1}{f} - 1\right) &= \left(\frac{1}{1+ab}\right)\left(\frac{1}{g} - 1\right) \end{aligned} \tag{4.97}$$

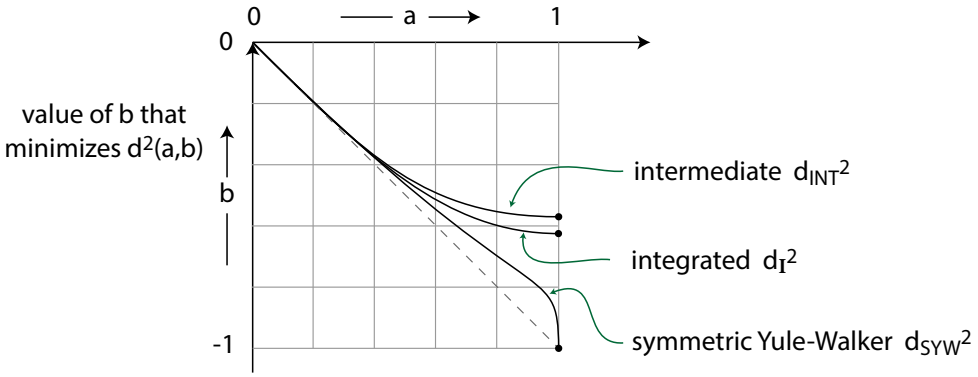


Figure 4.6: Value of b that minimizes $d^2(a, b)$ for the AR(1)[a] / MA(1)[b] system

In this case, when we minimize the target T_{INT} with respect to b, we get that for $a = 1$, $b \approx -0.57$.

These three results are plotted in Figure 4.6. Notice that the curves only differ substantially for $a > 0.5$, reinforcing our earlier observation that for neighboring models, distance measures based on the information matrix yield near-identical classification performance.

4.9 Summary

1. A generalized L_2 distance on a set of nearly-independent events $\{v_1, v_2, \ldots, v_n\}$ is

$$T = \sum_{s=1}^{n} \log^2 \frac{p(v_s, M_1)}{p(v_s, M_2)} \tag{4.98}$$

where $p(v_s, M_k)$ is the theoretical probability of event v_s under model M_k. For large n, that has the second derivative

$$\frac{\partial^2 T}{\partial \theta_i \partial \theta_j} = n \sum_x \frac{1}{p_x} \frac{\partial p_x}{\partial \theta_i} \frac{\partial p_x}{\partial \theta_j} \Delta\theta_j \quad \text{where } p_x = p(v_x, M) \tag{4.99}$$

where the sum is over all possible events, and θ_i and θ_j are parameters belonging to the base model M. That is equivalent to the information matrix of equation 4.21.

2. Applied to an ARMA(p, q) time series model, the generalized L_2 distance of equation 4.98 leads to the information matrix

$$\frac{1}{n}[i(\theta)]_{ij} = \begin{cases} 1/(1-\theta_i\theta_j) \text{ if both are AR or MA} \\ -1/(1-\theta_i\theta_j) \text{ if one is AR, one is MA} \end{cases} \tag{4.100}$$

where θ is the concatenation of the AR and MA roots $\theta = [r_1, \ldots, r_p, s_1, \ldots, s_q]$ (see section 4.3). Figures 4.1 shows how this formula correctly predicts ARMA parameter variance for $n \gg 1$, when the parameters are maximum likelihood estimates derived through the Whittle estimator.

3. Equation 4.100 may be integrated to give the ARMA(p, q) distance measure

$$D_I^2(\theta_1, \theta_2) = \frac{n_1 n_2}{n_1 + n_2} \sum_{k=1}^{\infty} \frac{1}{k^2} \left[\sum_{i=1}^{p_1} r_{1i}^k - \sum_{j=1}^{q_1} s_{1j}^k - \sum_{i=1}^{p_2} r_{2i}^k + \sum_{j=1}^{q_2} s_{2j}^k \right]^2 \tag{4.101}$$

where the models are ARMA$(p_1, q_1)[\vec{r}_1\ \vec{s}_1]$ and ARMA$(p_2, q_2)[\vec{r}_2\ \vec{s}_2]$. Note that the two models may be of differing orders.

4. If two distance measures both reduce to the same quadratic form $(\Delta\theta)^T ni(\theta)(\Delta\theta)$ for $\Delta\theta \approx 0$, then they will have nearly identical type II error rates over the entire range of offsets (see Figure 4.5 and Table 4.3).

Chapter 5

Discrete hidden Markov models

5.1 Preamble

In chapter 3, I developed a theory of distance measures for histograms. A histogram is similar to a vector, but with the restriction that the bin proportions are in the range [0, 1] and sum to one. In chapter 4, we looked at stationary invertible ARMA models. A stationary invertible ARMA model comprises a pair of vectors (the AR and MA coefficients) and two scalars (the long-term mean and innovation variance), with the restriction that the AR roots and MA roots are all less than one in magnitude (see eqn 4.26). A discrete hidden Markov model has two matrices, one for state transition probabilities and one for symbol emission probabilities, both of which are row-stochastic, meaning that their rows sum to one. The purpose of this chapter is to develop enough theory for discrete hidden Markov models to write down a distance measure between two such models, given the amount of data upon which the models are based.

5.2 Discrete hidden Markov model description

A discrete hidden Markov model (HMM) (section 2.5) has a finite set of states, which are not observable, and a finite set of symbols, which are observable (see Figure 5.1). Imagine a system which can exist in one of a finite number of states $\{\omega_1, \ldots \omega_n\}$. The states are not directly observable. When the system is in state ω_i, it will emit one of a finite number of symbols $\{v_1, \ldots v_m\}$. The symbols could be anything, but must be

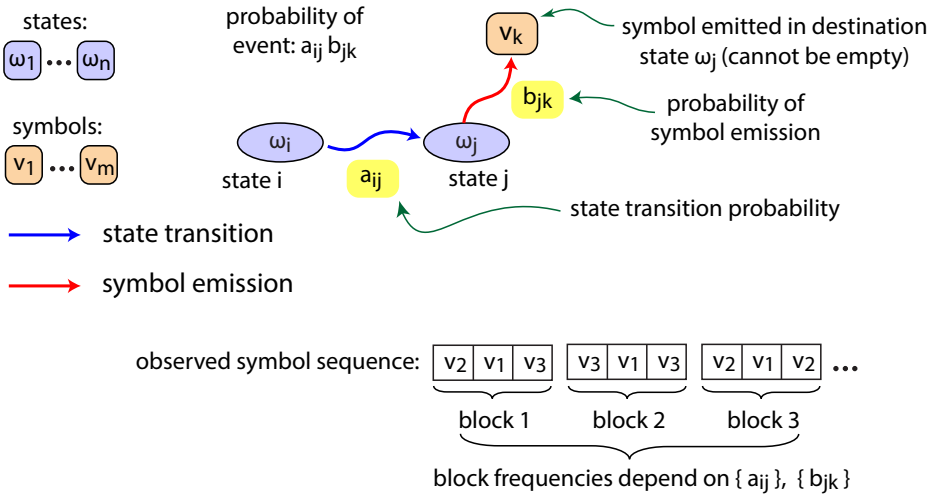


Figure 5.1: Description of variables and illustration of symbol blocks for a discrete HMM

observable. Furthermore, each symbol is observed unambiguously and without any additive or multiplicative noise.

After emitting a symbol in state ω_i, the system will move to another state ω_j (possibly the same as the initial state), with a probability that depends only on the system's current state (that's the Markovian premise). We will denote the state transition matrix as A, with $[A]_{ij} = a_{ij} =$ probability of going from state ω_i to ω_j. The symbol emission probabilities are collected into the matrix B, with $[B]_{jk} = b_{jk} =$ probability of emitting symbol v_k while in state ω_j. Matrices A and B are row-stochastic, meaning that each individual row sums to one. According to these definitions, if the system starts in state ω_i, then the probability of the system going from state ω_i to ω_j and then emitting symbol v_k is $a_{ij}b_{jk}$.

5.3 Ergodicity of the hidden Markov model

The output of such a system is an endless sequence of symbols $\{\ldots, v_{j_1}, \ldots, v_{j_M}, \ldots\}$, and the best we can do is record a subsequence $\{v_{j_1}, v_{j_2}, \ldots, v_{j_M}\}$, which may be very long. In fact, I will always assume that $M \gg 1$. We are going to work with statistics on an observed subsequence $\{v_{j_1} \ldots v_{j_M}\}$, so we need to require that the sequence be ergodic, so

that averages over a large enough subsequence eventually converge to the corresponding ensemble averages.

So I will assume that the state transition matrix is primitive, i.e. that there is some positive integer d such that $[A^d\ ]_{ij} > 0$ for $1 \leq i, j \leq n$, and that $M \gg d$. That means that each state is reachable from every other state by a finite number of transitions [48].

5.4 The stationary distribution for an HMM

If an ensemble of HMMs (all based on the model $\{A, B\}$) has an initial state distribution of $\vec{\pi}$ (i.e. the probability of starting off in state ω_i is π_i), then the state distribution after p transitions will be $\vec{\pi}A^p$ (note here that $\vec{\pi}$ is a stochastic row vector with $\vec{\pi}\vec{1}^T = \vec{\pi}[1 \ldots 1]^T = 1$). If A is primitive, then there is a unique stationary distribution $\vec{a}$ such that $\vec{a}A = \vec{a}$, and $\vec{a}\vec{1}^T = 1$ [49]. That stationary distribution is the row vector $[a_1, \ldots a_n]$ where

$$a_i = m_i / \sum_{j=1}^{n} m_j = [\text{adj}(I - A)^T]_{ii} / \sum_{j=1}^{n} [\text{adj}(I - A)^T]_{jj} \tag{5.1}$$

and m_i is the i-th diagonal minor of $(I_n - A)^T$ [83]. For example, for the case $n = 2$,

$$A = \begin{bmatrix} 1 - a_{12} & a_{12} \\ a_{21} & 1 - a_{21} \end{bmatrix} \quad \Rightarrow \vec{a} = \frac{1}{a_{12} + a_{21}} [a_{21} \quad a_{12}] \tag{5.2}$$

For example if we had $a_{12} = 0.2$ and $a_{21} = 0.4$, then we would get $\vec{a} = [2/3, 1/3]$. Now, a key observation about the minors of $(I_n - A)^T$ is that they are at most linear in the off-diagonal terms of the transition matrix A. That is because minors are determinants, so they do not involve the product of two elements in the same row or column. So, if $a_{ij} = [A]_{ij}$ is an off-diagonal element of A, then

$$\frac{\partial m_k}{\partial a_{ij}} = [m_k]_{a_{ij}=1} - [m_k]_{a_{ij}=0} \tag{5.3}$$

and the gradient of the stationary distribution with respect to the off-diagonal elements of A is

$$\frac{\partial a_k}{\partial a_{ij}} = \left(\frac{\partial m_k}{\partial a_{ij}}\right)\left(\frac{1}{m_1 + \ldots + m_n}\right) - \frac{m_k}{(m_1 + \ldots + m_n)^2}\left(\frac{\partial m_1}{\partial a_{ij}} + \ldots + \frac{\partial m_n}{\partial a_{ij}}\right) \tag{5.4}$$

We will use that formula in section 5.8 later below.

5.5 Expected symbol block frequencies

What is the ensemble probability of the symbol sequence $\{v_{j_1} \dots v_{j_M}\}$? Let $\vec{p}_k$ be a row vector whose i-th element is the probability of observing the sequence $\{v_{j_1} \dots v_{j_k}\}$ and ending up in state ω_i. Then, using the definitions of the transition matrix A and symbol emission matrix B, we get

$$\vec{p}_k = \vec{p}_{k-1} A D(\vec{b}_{j_k}) \text{ where } \vec{b}_j = [b_{1j} \dots b_{nj}]^T, \text{ and } D(\vec{b}_j) = \text{diag}(\vec{b}_j) \tag{5.5}$$

Setting $\vec{p}_0$ to the stationary distribution $\vec{a}$, and making use of its defining equation $\vec{a}A = \vec{a}$, we get

$$p(j_1 \dots j_k) = \text{prob}(\{v_{j_1} \dots v_{j_k}\}) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \dots A D(\vec{b}_{j_k}) \vec{1}^T \tag{5.6}$$

where $\vec{1}^T$ is the all-ones column vector of size n. Thus, for example, the ensemble probability of the string $\{v_j\}$ would be

$$p(j) = \vec{a} D(\vec{b}_j) \vec{1}^T \tag{5.7}$$

and the ensemble probability of the string $\{v_i v_j\}$ would be

$$p(i\, j) = \vec{a} D(\vec{b}_i) A D(\vec{b}_j) \vec{1}^T \tag{5.8}$$

Furthermore, we have that

$$\sum_{j=1}^{m} D(\vec{b}_j) = \text{diag}(\sum_{j=1}^{m} b_{1j}, \dots \sum_{j=1}^{m} b_{nj}) = I_n \tag{5.9}$$

since the system emits a symbol upon every state transition. Consequently,

$$p(i\,\#\,j) = \vec{a} D(\vec{b}_i) A^2 D(\vec{b}_j) \vec{1}^T, \quad p(i\,\#\,\#\,j) = \vec{a} D(\vec{b}_i) A^3 D(\vec{b}_j) \vec{1}^T \tag{5.10}$$

and so forth, where the hash character (#) stands for an arbitrary symbol. Now if A is a primitive stochastic matrix, then its largest eigenvalue is one, and its remaining eigenvalues (after the first) are all less than one in magnitude. In particular, the subdominant eigenvalue λ_2 (the eigenvalue of A that is largest in magnitude, but not one) satisfies $|\lambda_2| < 1$ and A^d converges to $G_1 = \vec{1}^T \vec{a}$ for $d \to \infty$ [68]. Hence

$$\begin{aligned} p(i\#\dots\#j) &\approx \vec{a} D(\vec{b}_i) (\vec{1}^T \vec{a}) D(\vec{b}_j) \vec{1}^T \quad (d\#, d \to \infty) \\ &\approx [\vec{a} D(\vec{b}_i) \vec{1}^T][\vec{a} D(\vec{b}_j) \vec{1}^T] = p(i)p(j) \end{aligned} \tag{5.11}$$

which shows that adjacent symbol blocks get less dependent as they grow in size. For the purpose of this chapter, I will assume that the initial state vector $\vec{p}_0$ is the stationary distribution $\vec{a}$. That way, we can specify a hidden Markov model with just the two matrices $\{A, B\}$ and omit the initial state distribution.

5.6 Variance of empirical symbol block frequencies

From equation 5.6, we have a formula for the expected stationary ensemble proportion of a symbol block of size N, namely

$$p(j_1, \ldots, j_N) = \text{prob}(\{v_{j_1}, \ldots, v_{j_N}\}) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \ldots A D(\vec{b}_{j_N}) \vec{1}^T \tag{5.12}$$

where we have assumed that the ensemble has reached its stationary distribution. We can also measure the empirical covariance of that proportion by observing (or generating) a set of symbol strings of length $M \gg N$, chopping each one into blocks of length N, finding the empirical incidence of all the symbol blocks of size N, and then calculating the empirical covariance

$$\text{cov}(u, v) = \langle \hat{p}(u) \hat{p}(v) \rangle - \langle \hat{p}(u) \rangle \langle \hat{p}(v) \rangle \tag{5.13}$$

over the set of input strings (here, u and v are arbitrary symbol blocks of size N, and $\hat{p}(u)$ and $\hat{p}(v)$ are their measured proportions in the input strings). We would like to be able to predict these second-order statistics, since they will be needed for the convergence argument in section 5.7 below, and also in section 5.9. In Appendix F, I give a derivation for this covariance, and the result is:

$$\begin{aligned} (N/M)\text{cov}(T_u, T_v) &= [p(u)\delta_{uv} - p(u)p(v)] + 2 \sum_{k=1}^{M/N-1} [s_{2,k} - p(u)p(v)] \\ \text{where } s_{2,1} &= \frac{1}{2}[p(uv) + p(vu)] \\ s_{2,2} &= \frac{1}{2}[p(u\#v) + p(v\#u)] \quad \text{etc} \end{aligned} \tag{5.14}$$

In equation 5.14, δ_{uv} is one if $u = v$ and zero otherwise, T_u is the count of symbol block u in a symbol string of length M, $p(u)$ is the expected ensemble proportion of symbol block u, and the hash character (#) represents an arbitrary symbol block of size N. The first term $p(u)\delta_{uv} - p(u)p(v)$ is what we would get if symbol blocks of size N were independent. The remaining terms capture the effect of serial correlation of the emitted symbols.

Let's show that these predictions are reasonable! The methodology is as follows: for a given HMM $\{A, B\}$, generate N_{trial} * N_{batch} strings of length M, where M is large enough so that all symbol blocks will have an expected count greater than 5 [24]. Chop up each generated string into blocks of size N, and calculate the empirical block proportions $\hat{p}(u)$. For each trial of N_{batch} estimates, calculate the empirical covariances $\langle \hat{p}(u) \hat{p}(v) \rangle - \langle \hat{p}(u) \rangle \langle \hat{p}(v) \rangle$. Finally, compute the spread (i.e. confidence intervals) on $\text{cov}(u, v)$ over all

Table 5.1: Covariance of single symbols with $\lambda_2 = 0.44$, showing agreement with predicted values. The underlying HMM is defined in 5.15

symbols	10^6 * cov($\hat{p}(i), \hat{p}(j)$)	
	95% confidence interval	**theoretical**
1,1	[37.02, 37.57]	37.40
1,2	[-26.66, -26.16]	-26.62
1,3	[-10.88, -10.53]	-10.81
2,2	[34.91, 35.50]	35.30
2,3	[-8.97, -8.64]	-8.72
3,3	[19.32, 19.64]	19.50

N_{trial} trials. Compare that to the theoretical result of equation 5.14, where the cutoff value of k is defined by $|\lambda_2|^{kN} < 0.001$, with λ_2 being the subdominant eigenvalue of the transition matrix A.

For Tables 5.1, 5.2, and 5.3, the sequence length is $M = 8192$, and each trial has $N_{\text{batch}} = 1024$ sequences. For Tables 5.1 and 5.3, the number of trials is $N_{\text{trial}} = 128$ and the HMM that generated the sequences is

$$A = \begin{bmatrix} 0.78 & 0.22 \\ 0.34 & 0.66 \end{bmatrix}, \quad B = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \end{bmatrix} \quad (\lambda_2 = 0.44) \tag{5.15}$$

Note that if the subdominant eigenvalue λ_2 is close to zero, then an ensemble of HMMs with parameter set $\{A, B\}$ will achieve stationarity quickly. Conversely, if λ_2 approaches one, then that same ensemble of HMMs will achieve stationarity slowly and the symbol block frequency covariances could be very different than what we would get if the symbol blocks were independent. For Table 5.2, the number of trials is $N_{\text{trial}} = 168$ and the HMM that generated the sequences is

$$A = \begin{bmatrix} 0.88 & 0.12 \\ 0.18 & 0.82 \end{bmatrix}, \quad B = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \end{bmatrix} \quad (\lambda_2 = 0.70) \tag{5.16}$$

What Tables 5.1, 5.2, and 5.3 are showing is that the theoretical predictions of equation 5.14 match empirical results, so our theory of symbol block frequency variance is looking solid so far!

Table 5.2: Variance of single symbols with $\lambda_2 = 0.70$, showing agreement with predicted values. The underlying HMM is defined in 5.16

symbol	10^6 * variance of $\hat{p}(i)$	
	95% confidence interval	**theoretical**
1	[51.77, 52.39]	51.90
2	[49.88, 50.54]	50.00
3	[19.39, 19.66]	19.50

Table 5.3: Variance of symbol pairs with $\lambda_2 = 0.44$, showing agreement with theoretical predictions. The underlying HMM is defined in 5.15

symbol pair ij	10^6 * variance of $\hat{p}(i\,j)$	
	95% confidence interval	**theoretical**
11	[44.48, 45.25]	44.57
12	[28.97, 29.41]	29.13
13	[19.55, 19.85]	19.89
21	[28.86, 29.25]	29.13
22	[33.07, 33.59]	33.25
23	[16.24, 16.46]	16.37
31	[19.69, 19.99]	19.89
32	[16.24, 16.48]	16.37
33	[9.28, 9.43]	9.38

5.7 Convergence of the theoretical symbol block covariance

From equation 5.14, our theoretical formula for symbol block covariance is

$$(N/M)\text{cov}(T_u, T_v) = [p(u)\delta_{uv} - p(u)p(v)] + 2\sum_{k=1}^{M/N-1}[h_{2,k} - p(u)p(v)] \tag{5.17}$$

where $h_{2,k} = [p(u\#\ldots\#v) + p(v\#\ldots\#u)]/2$ and $\#\ldots\#$ represents $(k-1)$ arbitrary symbol blocks, each of length N. In this section, I will show that the convergence of equation 5.17 is geometric, and will establish the conditions under which we may treat symbol blocks as independent for the purposes of calculating a chi-square "goodness of fit" statistic [24]. The first thing to note about equation 5.17 is that the summation is finite, with bounded terms (they are all probabilities or products of probabilities), so it cannot diverge. Now the transition matrix A is a primitive stochastic matrix, but not necessarily symmetric. If it is diagonalizable, then its eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ are semi-simple, meaning that for each eigenvalue, the algebraic and geometric multiplicities are equal. In this case, the modal expansion of the transition matrix in spectral projectors is [84]

$$\begin{aligned} A &= \sum_{i=1}^{s}\lambda_i G_i \quad \text{where } G_iG_j = \delta_{ij}G_i, \sum_{i=1}^{s} G_i = I_n, \text{ and} \\ G_i &= \prod_{j\neq i}^{s}(A - \lambda_j I)/\prod_{j\neq i}^{s}(\lambda_i - \lambda_j) \end{aligned} \tag{5.18}$$

in which case we get

$$A^k = G_1 + \sum_{i=2}^{s}\lambda_i^k G_i = G_1 + O(|\lambda_2|^k) \quad \text{NB } \lambda_1 = 1 \tag{5.19}$$

where $G_1 = \vec{1}^T\vec{a}$ (since $\lambda_1 = 1$ is a simple eigenvalue when A is primitive). If A is not diagonalizable, then the relevant modal expansion is [85]

$$A^k = \sum_{i=1}^{s}\sum_{j=0}^{m_i-1}\binom{k}{j}\lambda_i^{k-j}(A - \lambda_i I)^j G_i \tag{5.20}$$

where m_i is the index of eigenvalue λ_i, and the spectral projectors are still orthogonal and sum to I_n. In this case, we may make use of the approximation

$$\binom{k}{j} \approx \frac{k^j}{j!} \quad \text{for } k \gg j \tag{5.21}$$

to get [100]

$$A^k = G_1 + O(k^{m_2-1}\,|\lambda_2|^k) \tag{5.22}$$

Either way, the convergence of A^k to $G_1 = \vec{1}^T\vec{a}$ is geometric.

The next question is: given the second-order statistics of equation 5.17, what is the correction (if any) that we need to apply to the standard chi-square goodness of fit statistic, when calculated for observed symbol block frequencies in large input strings? If the input string is of length M, and the symbol blocks of interest are of size N and independent, then we would have

$$(\frac{N}{M})\mathrm{cov}(T_u, T_v) = p(u)\delta_{uv} - p(u)p(v) \tag{5.23}$$

whose corresponding chi-square statistic is

$$\chi^2 = \sum_{|u|=N} \frac{(\hat{T}_u - T_u)^2}{(M/N)p(u)} = \left(\frac{M}{N}\right) \sum_{|u|=N} \frac{(\hat{p}(u) - p(u))^2}{p(u)} \tag{5.24}$$

with the number of degrees of freedom being $m^N - 1$, i.e. one less than the number of symbol blocks of size N. Here, m is the number of distinct symbols in the HMM output.

We want to see what additional terms will arise in the chi-square statistic from using equation 5.17 instead of 5.23. First, we'll define a symbol block weighting as

$$W(u) = W(\{v_{j_1}, \ldots v_{j_N}\}) = D(\vec{b}_{j_1})AD(\vec{b}_{j_2}) \ldots AD(\vec{b}_{j_N}) \tag{5.25}$$

where u is the symbol block $\{v_{j_1}, \ldots v_{j_N}\}$. In words: $[W(u)]_{ij}$ is the ensemble probability of starting in state ω_i, making $(N-1)$ state changes, and ending in state ω_j, having emitted the symbols $\{v_{j_1}, \ldots v_{j_N}\}$. The symbol block weighting of an empty string is the identity matrix: $W(\emptyset) = I_n$. Then, from Equations 5.6 and 5.9, we have

$$\begin{aligned} p(u\,v) &= \vec{a}W(u)AW(v)\vec{1}^T \\ p(u\,\#\,v) &= \vec{a}W(u)A^{1+N}W(v)\vec{1}^T \\ p(u\,\#\,\#\,v) &= \vec{a}W(u)A^{1+2N}W(v)\vec{1}^T \quad \text{and so forth} \end{aligned} \tag{5.26}$$

So, our correction to the covariance of the symbol block proportions (equation 5.23) is

$$\begin{aligned} E_{uv} &= 2\sum_{k=1}^{M/N-1} [h_{2,k} - p(u)p(v)] \\ &= \sum_{k=1}^{M/N-1} [\vec{a}W(u)(A^{1+kN} - G_1)W(v)\vec{1}^T + \vec{a}W(v)(A^{1+kN} - G_1)W(u)\vec{1}^T] \end{aligned} \tag{5.27}$$

In the case where the transition matrix A is diagonalizable, with eigenvalues $\{\lambda_1 \dots \lambda_s\}$ and corresponding spectral projectors $\{G_1 \dots G_s\}$, we have

$$A^k = G_1 + \sum_{j=2}^{s} \lambda_j^k G_j = G_1 + (A - G_1)^k \tag{5.28}$$

Setting $A_r = (A - G_1)[I - (A - G_1)^N]^{-1}$, we get in that case

$$E_{uv} = 2 \sum_{k=1}^{M/N-1} [h_{2,k} - p(u)p(v)] \approx \vec{a}[W(u)A_r W(v) + W(v)A_r W(u)]\vec{1}^T \tag{5.29}$$

Clearly, the correction E_{uv} to the multinomial covariance of equation 5.23 is second order with respect to $p(u)$. Furthermore, the correction matrix E has rows and columns that sum to zero:

$$\begin{aligned} \sum_{|u|=N} [\vec{a}W(u)A^k W(v)\vec{1}^T - p(u)p(v)] &= \vec{a}W(v)\vec{1}^T - p(v) = p(v) - p(v) = 0 \\ \sum_{|v|=N} [\vec{a}W(u)A^k W(v)\vec{1}^T - p(u)p(v)] &= \vec{a}W(u)\vec{1}^T - p(u) = p(u) - p(u) = 0 \end{aligned} \tag{5.30}$$

That gives us the following picture of the correction matrix E:

$$\begin{aligned} E_{uv} &= 2 \sum_{k=1}^{M/N-1} [h_{2,k} - p(u)p(v)] \\ &\approx \vec{a}[W(u)A_r W(v) + W(v)A_r W(u)]\vec{1}^T \approx 2\epsilon_{uv}|\lambda_2|p(u)p(v) \end{aligned} \tag{5.31}$$

where the ϵ_{uv} are of order one and have mixed signs. Setting $\Sigma_{uv} = p(u)\delta_{uv} - p(u)p(v)$, we have

$$(\Sigma + E)^{-1} \approx \Sigma^{-1} - \Sigma^{-1} E \Sigma^{-1} \quad \text{NB excludes one bin} \tag{5.32}$$

and the corresponding chi-square goodness of fit statistic is

$$\chi^{2\prime} = \left(\frac{M}{N}\right) \sum_{|u|=N} \frac{[\hat{p}(u) - p(u)]^2}{p(u)} [1 + 2p(u)\epsilon_u|\lambda_2|] \tag{5.33}$$

where the ϵ_u are of order one and have mixed signs. Supposing that $\sum_u \epsilon_u p(u) \approx 0$, the relative change to the chi-square goodness of fit statistic is approximately

$$\frac{\Delta\chi^2}{\chi^2} \approx \sum_{|u|=N} [2p(u)\epsilon_u|\lambda_2|]^2 \approx 4\lambda_2^2/m^N \tag{5.34}$$

Table 5.4: Comparison between chi-square goodness of fit statistic for symbol block frequencies and a true chi-square distribution, for the HMM defined by 5.36

Block size	**degrees of freedom**	**no. of sequences**	**KS statistic**
2	8	4096	1.20
3	26	4096	0.47
4	80	3584	0.84

where m^N is the number of distinct symbol blocks of size N (recall that m is the number of distinct symbols). That correction is typically small. For example, for $\{\lambda_2 = 1/2, m = 3, N = 3\}$, the correction is of relative magnitude

$$\frac{\Delta\chi^2}{\chi^2} \approx 4\left(\frac{1}{27}\right)\left(\frac{1}{2}\right)^2 = \frac{1}{27} \quad (\text{ about } 4\% \,) \tag{5.35}$$

Thus, when $4\lambda_2^2/m^N \ll 1$, we can use the standard chi-square statistic of equation 5.24 as a test for goodness of fit between the theoretical and empirical symbol block frequencies. As an illustration of this idea, I calculated the standard chi-square statistic for the two-state, three-symbol HMM

$$A = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}, \quad B = \begin{bmatrix} 0.67 & 0.20 & 0.13 \\ 0.20 & 0.57 & 0.23 \end{bmatrix} \quad (\lambda_2 = 0.50) \tag{5.36}$$

with sequence length $M = 32400$ and symbol block sizes $N = \{2, 3, 4\}$. For each symbol block size, I construct an empirical CDF for the standard chi-square statistic of equation 5.24, using equation 5.6 for the theoretical symbol block frequencies, and then calculate the one-sample Kolmogorov-Smirnov [42] statistic between the empirical CDF and that expected for a chi-square distribution of the appropriate number of degrees of freedom. The results are in Table 5.4. The 95% confidence limit for the one-sample Kolmogorov-Smirnov statistic is 1.36, so our conclusion is that for this HMM, even for relatively small m and N, we are safe to use the standard chi-square statistic of equation 5.24 as a goodness of fit measure between empirical and predicted symbol block frequencies. We will use this result in section 5.9.

5.8 Baum-Welch algorithm and covariance of the model parameters

In this section, I will look at how to train a hidden Markov model, and how to estimate the covariance of the resulting model parameters. Our starting point is a symbol sequence $\{v_{j_1}, \ldots, v_{j_M}\}$ where $M \gg 1$, which we take as a sample of the model's output. We'll also use the notation $v(t) = v_{j_t}$, meaning the symbol observed at time t where $1 \leq t \leq M$. The Baum-Welch algorithm, which I described in section 2.5, assumes that the model size n is known and that we have an initial model estimate $\{\hat{A}_0, \hat{B}_0\}$. We can derive the algorithm from equation 5.6 as follows. From equation 5.6, the ensemble probability of the observed sequence is

$$p(j_1 \ldots j_M) = \text{prob}(\{v_{j_1} \ldots v_{j_M}\}) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \ldots A D(\vec{b}_{j_M}) \vec{1}^T \tag{5.37}$$

We'll consider all the elements of the transition matrix A and the emission matrix B as independent, but impose the additional constraints $A\vec{1}^T = \vec{1}^T$ and $B\vec{1}^T = \vec{1}^T$ where $\vec{1}^T$ is the all-ones column vector of length n. Then the optimization target is

$$T = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \ldots A D(\vec{b}_{j_M}) \vec{1}^T - \vec{f}(A - I)\vec{1}^T - \vec{g}(B - I)\vec{1}^T \tag{5.38}$$

where $\vec{f} = [f_1 \ldots f_n]$ and $\vec{g} = [g_1 \ldots g_n]$ are row vectors of Lagrange multipliers. Maximizing T as a function of the model parameters $a_{ij} = [A]_{ij}$ and $b_{jk} = [B]_{jk}$ leads to the equations

$$\frac{1}{a_{ij}} \sum_{t=1}^{M} \gamma(i, j, t) = f_i, \quad \frac{1}{b_{jk}} \sum_{j_t = k} \sum_{i=1}^{n} \gamma(i, j, t) = g_j \tag{5.39}$$

where $\gamma(i, j, t) = \alpha(t - 1, i) a_{ij} b_{jk} \beta(t, j)$, the forward probability $\alpha(t - 1, i)$ is the ensemble probability of being in state ω_i at time $t - 1$, having generated the symbol sequence $\{v(1), \ldots, v(t - 1)\}$, and the backward probability $\beta(t, j)$ is the ensemble probability of starting in state ω_j at time t, and generating the symbol sequence $\{v(t + 1), \ldots, v(M)\}$. Here, I have neglected the contribution of the stationary distribution $\vec{a}$ to derivatives of the objective T, since that contribution is of order $1/M$. The forward and backward probabilities are

$$\begin{aligned} \alpha(t - 1, i) &= \vec{a} D(\vec{b}_{v(1)}) A D(\vec{b}_{v(2)}) \ldots A D(\vec{b}_{v(t-1)}) \vec{g}(i)^T \\ \beta(t, j) &= \vec{g}(j) D(\vec{b}_{v(t+1)}) A D(\vec{b}_{v(t+2)}) \ldots A D(\vec{b}_{v(M)}) \vec{1}^T \end{aligned} \tag{5.40}$$

where $\vec{g}(j)$ is a row vector with one in position j and zero elsewhere. The right-hand sides of equation 5.39 are functions of row index only, so consistency requires that

$$\begin{aligned} a_{ij} &= (\text{constant depending only on } i) \sum_{t=1}^{M} \gamma(i,j,t) \\ b_{jk} &= (\text{constant depending only on } j) \sum_{j_t=k} \sum_{i=1}^{n} \gamma(i,j,t) \end{aligned} \tag{5.41}$$

We can satisfy those equations and the row-stochastic constraints by requiring

$$a_{ij} = \sum_{t=1}^{M} \gamma(i,j,t) / \sum_{t=1}^{M} \sum_{j=1}^{n} \gamma(i,j,t), \quad b_{jk} = \sum_{j_t=k} \sum_{i=1}^{n} \gamma(i,j,t) / \sum_{t=1}^{M} \sum_{i=1}^{n} \gamma(i,j,t) \tag{5.42}$$

Although we cannot solve equation 5.42 directly for the model $\{\hat{A}, \hat{B}\}$, we can use that equation as the basis of a successive approximation algorithm:

$$\begin{aligned} [\hat{A}_{s+1}]_{ij} &= \sum_{t=1}^{M} \gamma(i,j,t) / \sum_{t=1}^{M} \sum_{j=1}^{n} \gamma(i,j,t) \\ [\hat{B}_{s+1}]_{jk} &= \sum_{j_t=k} \sum_{i=1}^{n} \gamma(i,j,t) / \sum_{t=1}^{M} \sum_{i=1}^{n} \gamma(i,j,t) \end{aligned} \tag{5.43}$$

where the $\gamma(i,j,t)$ are evaluated with the model $\{\hat{A}_s, \hat{B}_s\}$. That describes the Baum-Welch algorithm, which has linear convergence, meaning that successive updates are approximately equal in size [62]. The Baum-Welch algorithm does give us an empirical hidden Markov model for a fixed model size and symbol sequence $\{v(1) \dots v(M)\}$. However, it does not give us the covariance of the model parameters. Let's fix that!

We could apply equation 4.22 directly if we had a set of nearly-independent events, but the symbol sequence output of an HMM does not give us that; the symbols all have serial correlation. However, a reasonable argument is that for a primitive transition matrix A, the entropy per symbol settles down to a constant in large enough blocks:

$$\lim_{N \to \infty} \left(\frac{1}{N} \right) (\text{ entropy of } N\text{-symbol block}) = \text{ function of model } \{A, B\} \tag{5.44}$$

If that is true, then we can approximate the observed sequence surprisal as

$$(\text{surprisal of output sequence}) \approx \left(\frac{M}{N} \right) (\text{ entropy of } N\text{-symbol block when } N \gg 1) \tag{5.45}$$

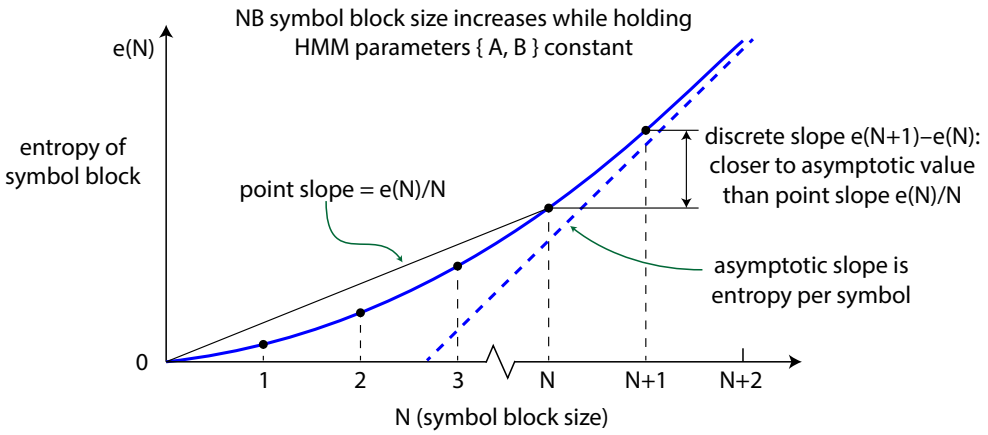


Figure 5.2: Finding the fastest-converging approximation to the asymptotic entropy per symbol for a hidden Markov model

Taking N-symbol blocks as being approximately independent when $N \gg 1$, we would then get

$$\begin{aligned}(\text{entropy per symbol}) \approx \lim_{N\to\infty} [\ &\text{entropy of } (N+1)\text{-symbol block} - \\ &\text{entropy of } N\text{-symbol block }]\end{aligned} \tag{5.46}$$

Figure 5.2 illustrates why equation 5.46 will converge more quickly to the asymptotic entropy per symbol than equation 5.44. Under the assumptions mentioned so far, we can apply equation 4.22 to get

$$\left.\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\right|_{\theta_0} \approx M \lim_{N\to\infty} \left[\sum_{|u|=N} \frac{1}{p(u)} \frac{\partial p(u)}{\partial\theta_i} \frac{\partial p(u)}{\partial\theta_j} \right]_N^{N+1} \tag{5.47}$$

where the summation is over all possible symbol blocks of size N, $p(u)$ is the ensemble probability of the symbol block u, θ_i represents an independent parameter of the model $\{A, B\}$, and θ_0 represents the independent parameters of the model $\{\hat{A}, \hat{B}\}$ that makes

the total surprisal $\mathcal{L}$ stationary. As I mentioned in section 2.5, the Baum-Welch algorithm only guarantees a local minimum of the total surprisal $\mathcal{L}$, not a global one. Setting

$$[Q]_{ij} = \lim_{N\to\infty} \left[\sum_{|u|=N} \frac{1}{p(u)} \frac{\partial p(u)}{\partial \theta_i} \frac{\partial p(u)}{\partial \theta_j} \right]_N^{N+1} \tag{5.48}$$

then by the argument of section 4.1, we would expect the model parameter covariance to be

$$\mathrm{cov}(\hat{\theta}) = \Sigma = (1/M)Q^{-1} \tag{5.49}$$

Of course, in order to verify that prediction, we need to be able to calculate $\partial p(u)/\partial \theta_i$ when θ_i is either a_{ij} (an element of A) or b_{jk} (an element of B). Taking the symbol string u to be $\{v_{j_1} \ldots v_{j_N}\}$, we get

$$\begin{aligned} p(u) &= \mathrm{prob}(\{v_{j_1} \ldots v_{j_N}\}) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \ldots A D(\vec{b}_{j_N}) \vec{1}^T \\ \frac{\partial p(u)}{\partial a_{ij}} &= \frac{\partial \vec{a}}{\partial a_{ij}} W(u) \vec{1}^T + \sum_{p=2}^{N} \vec{a} W(\{v_{j_1} \ldots v_{j_{p-1}}\}) \frac{\partial A}{\partial a_{ij}} W(\{v_{j_p} \ldots v_{j_N}\}) \vec{1}^T \\ \frac{\partial p(u)}{\partial b_{jk}} &= \sum_{p=1}^{N} \vec{a} W(\{v_{j_1} \ldots v_{j_{p-1}}\}) A \frac{\partial D(\vec{b}_{j_p})}{\partial b_{jk}} W(\{v_{j_{p+1}} \ldots v_{j_N}\}) \vec{1}^T \end{aligned} \tag{5.50}$$

Equation 5.4 gives us $\partial \vec{a}/\partial a_{ij}$. As for the other derivatives $\partial A/\partial a_{ij}$ and $\partial D(\vec{b}_l)/\partial b_{jk}$, they depend on which model parameters we consider to be independent. Not all the matrix elements in an HMM parameter set $\{A, B\}$ are independent, since both A and B are row-stochastic. If we take the diagonal elements of A and the last column of B as being dependent, i.e.

$$\begin{aligned} A &= \begin{bmatrix} 1-(a_{12}+\ldots a_{1n}) & a_{12} & \ldots & a_{1n} \\ a_{21} & 1-(a_{21}+\ldots a_{2n}) & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & 1-(a_{n1}+\ldots a_{n,n-1}) \end{bmatrix} \\ B &= \begin{bmatrix} b_{11} & \ldots & b_{1,m-1} & 1-(b_{11}+\ldots b_{1,m-1}) \\ \vdots & & \vdots & \vdots \\ b_{n1} & \ldots & b_{n,m-1} & 1-(b_{n1}+\ldots b_{n,m-1}) \end{bmatrix} \end{aligned} \tag{5.51}$$

then $\partial A/\partial a_{ij}$ is an n by n matrix with

$$\left[\frac{\partial A}{\partial a_{ij}}\right]_{kl} = \delta_{ik}(\delta_{jl} - \delta_{kl}) \tag{5.52}$$

Table 5.5: Agreement between predicted and actual parameter covariances for training done via Baum-Welch where $\lambda_2 = 0.5$. The underlying HMM is defined in equation 5.55

specific	$M\text{var}(a_{12})$	$M\text{var}(a_{21})$	$M\text{var}(b_{11})$	$M\text{var}(b_{12})$	$M\text{var}(b_{21})$	$M\text{var}(b_{22})$
N = 2	72.81	114.2	29.52	16.60	64.20	38.80
N = 3	50.40	75.45	22.04	12.44	47.23	29.13
N = 4	46.29	67.96	20.64	11.67	44.02	27.29
N = 5	45.33	66.12	20.32	11.49	43.26	26.86
N = 6	45.11	65.66	20.25	11.45	43.08	26.76
N = 7	45.05	65.56	20.23	11.44	43.04	26.73
N = 8	45.04	65.53	20.23	11.44	43.03	26.73
mean all trials	45	70	21	12	47	29
95% confidence	±3	±5.5	±1.5	±0.9	±3.8	±2.4

and $\partial D(\vec{b_l})/\partial b_{jk}$ is an n by n diagonal matrix with

$$\left[\frac{\partial D(\vec{b_l})}{\partial b_{jk}}\right]_{pp} = \delta_{pj}(\delta_{kl} - \delta_{ml}) \tag{5.53}$$

We can check this theory as follows. For a given hidden Markov model $\{A, B\}$, generate N_{trial} * N_{batch} strings of length M. For each string, run the Baum-Welch algorithm with the true model as the starting point. Within a batch, calculate the unbiased covariance

$$\text{cov}(\hat{\theta}) = \frac{1}{N_c}\sum(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \tag{5.54}$$

where the summation is over all Baum-Welch runs that converged, N_c is the number of Baum-Welch runs that converged, $\hat{\theta}$ is an estimated column vector of independent model parameters, and θ_0 represents the true values of the independent model parameters. Finally, calculate the spread of $\text{cov}(\hat{\theta})$ over N_{trial} batches and compare with our theoretical predictions.

Table 5.5 shows the results of that program, applied to the two-state, three-symbol HMM

$$A = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}, \quad B = \begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.50 & 0.30 \end{bmatrix} \quad (\lambda_2 = 0.50) \tag{5.55}$$

with $M = 32400, N_{\text{batch}} = 64$, and $N_{\text{trial}} = 58$. Note that in Table 5.5, the rows for different N refer to equation 5.49 evaluated for those values of N. The table shows broad

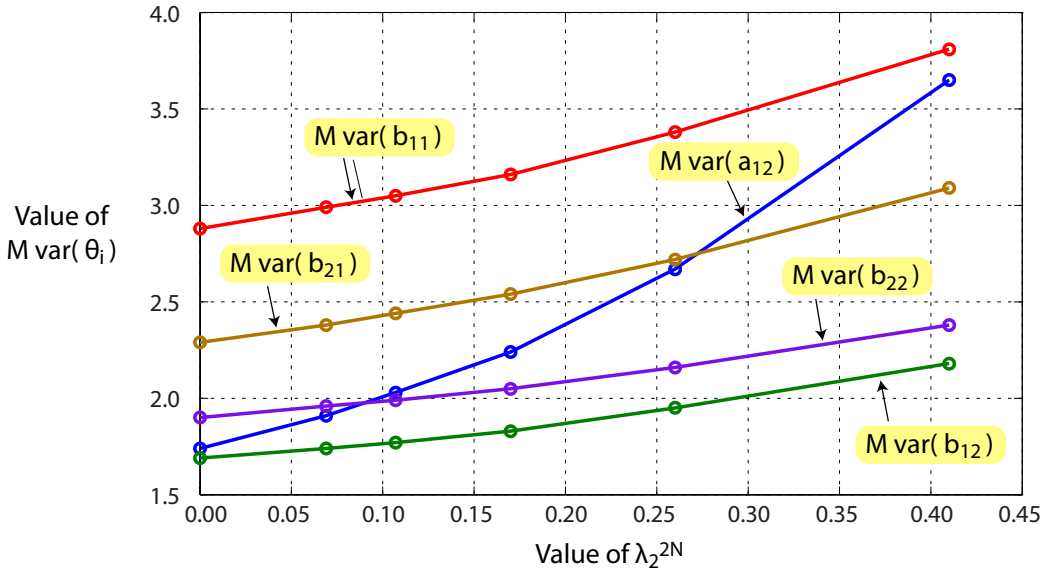


Figure 5.3: Extrapolating the parameter covariance matrix, for the HMM defined in equation 5.56. The extrapolated values are on the left vertical axis, and correspond to the limit $N \to \infty$

agreement with theory, although the empirical covariance does seem to have a bias away from zero, much like what we observed in section 4.3 for ARMA model parameters.

Table 5.5 shows that for $\lambda_2 = 0.5$, equation 5.48 for the independent HMM parameter precision matrix converges pretty rapidly, and we can take the matrix value at $N = 8$ or $N = 9$ as the final one. But what about when the convergence is slower? For example, the hidden Markov model of equation 5.56

$$A = \begin{bmatrix} 0.90 & 0.10 \\ 0.10 & 0.90 \end{bmatrix}, \quad B = \begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.50 & 0.30 \end{bmatrix} \quad (\lambda_2 = 0.80) \tag{5.56}$$

has $\lambda_2 = 0.8$, and would require us to evaluate equation 5.48 for the independent parameter precision matrix for $N \approx 18$, which would involve the gradients of some $3^{18} \approx 4$ x 10^8 symbol block probabilities - not a realistic calculation. By "realistic calculation", I mean a calculation that takes less than one day on a commercial laptop computer. One possibility, however, is to evaluate equation 5.48 up to something reasonable, say $N = 9$, and then extrapolate the covariance matrix out to a larger block size N.

Figure 5.3 and Table 5.6 show the results of that idea, applied to the hidden Markov

Table 5.6: Agreement between predicted and actual parameter variances for training done via Baum-Welch where $\lambda_2 = 0.8$. The underlying HMM is defined in equation 5.56. The predicted values are in the row marked 'extrapolated', and the actual values are in the last row.

specific	$M\text{var}(a_{12})$	$M\text{var}(a_{21})$	$M\text{var}(b_{11})$	$M\text{var}(b_{12})$	$M\text{var}(b_{21})$	$M\text{var}(b_{22})$
N = 3	6.36	6.20	4.78	2.70	3.92	2.89
N = 4	3.65	3.56	3.81	2.18	3.09	2.38
N = 5	2.67	2.61	3.38	1.95	2.72	2.16
N = 6	2.24	2.18	3.16	1.83	2.54	2.05
N = 7	2.03	1.98	3.05	1.77	2.44	1.99
N = 8	1.91	1.87	2.99	1.74	2.38	1.96
extrapolated	[1.71,1.77]	[1.66,1.72]	[2.86,2.90]	[1.67, 1.71]	[2.27,2.31]	[1.88,1.92]
95% confidence	[1.76,1.86]	[1.68,1.79]	[2.83,2.97]	[1.68,1.78]	[2.28,2.41]	[1.86,1.97]

model of equation 5.56 that has $\lambda_2 = 0.8$, with the experimental context $M = 32400$, $N_{\text{batch}} = 64$, and $N_{\text{trial}} = 174$. In Figure 5.3, I extrapolated the values of M var (θ_i), regarded as functions of λ_2^N, to get their final values for large N. Once again we have broad agreement between the observed parameter covariance that we get from the Baum-Welch procedure, and the prediction of equation 5.49.

Suppose now, however, that equation 5.48 is not converging at all, and we need an estimate for parameter covariance. What else could we do? Well, we have equation 5.6 for the likelihood of a model, conditioned on the observed symbol sequence:

$$p(j_1 \dots j_M) = \text{prob}(\{v_{j_1} \dots v_{j_M}\}) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \dots A D(\vec{b}_{j_M}) \vec{1}^T \tag{5.57}$$

and so as a "last resort" we could calculate $p(j_1 \dots j_M)$ for values of the independent parameter vector $\vec{\theta}$ near the Baum-Welch maximum lilelihood solution $\vec{\theta}_0$, and then estimate the Hessian matrix

$$\left.\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\right|_{\theta_0} = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(j_1 \dots j_M)\right]_{\theta_0} \tag{5.58}$$

numerically. I won't describe that procedure any further, it has been well covered by [122] and [78].

5.9 Model fitting via the chi-square goodness of fit criterion

A hidden Markov model manifests itself not only in the overall output sequence probability of equation 5.37, but also in the ensemble symbol block probabilities

$$p(u) = \vec{a}D(\vec{b}_{j_1})AD(\vec{b}_{j_2})\ldots AD(\vec{b}_{j_N})\vec{1}^T \quad \text{where } u = \{v_{j_1}\ldots v_{j_N}\} \tag{5.59}$$

So, we could also estimate a hidden Markov model by minimizing the chi-square goodness-of-fit statistic

$$\chi^2 = \left(\frac{M}{N}\right)\sum_{|u|=N}\frac{[\hat{p}(u) - p(u,\theta)]^2}{p(u,\theta)} \tag{5.60}$$

where $\hat{p}(u)$ is the observed proportion of symbol blocks u in the output string $\{v_{j_1}\ldots v_{j_M}\}$, θ represents the independent model parameters, and $p(u,\theta)$ is the theoretical symbol block occurrence given by equation 5.59.

In order for the goodness-of-fit statistic of equation 5.60 to represent a true chi-square variate, there are two conditions we need to satisfy. First, we need $(M/N)p(u) > 5$ for all those symbol blocks for which $p(u) > 0$. That helps ensure that the sampling distribution of $\hat{p}(u) - p(u)$ is approximately normal [24]. Second, we need $4\lambda_2^2/m^N \ll 1$, where λ_2 is the sub-dominant eigenvalue of the transition matrix A, and m^N is the number of symbol blocks of size N. That also helps ensure that the sampling distribution of $\hat{p}(u) - p(u)$ is approximately normal (see section 5.7 above). More practically, we require that the number of block frequencies measured (m^N) be much greater than the number of independent model parameters. For a hidden Markov model with n states and m distinct symbols, the number of independent parameters is $n(n-1) + n(m-1) = n(n+m-2)$, so we would like

$$m^N \gg n(n+m-2) \quad \text{ideally } m^N > (\,10\,)n(n+m-2) \tag{5.61}$$

Putting all that together: choose the largest possible symbol block size N, subject to $(M/N)p(u) > 5$ where $p(u) > 0$.

Hidden Markov models have one notorious problem with respect to parameter estimation. Because the states are unobserved, their labellings may be permuted without changing the underlying model or its predictions. In order to avoid this issue, I require that the states be ordered according to their self-transition values, i.e. I require that

$$a_{11} \geq a_{22} \geq \ldots \geq a_{nn} \quad \text{state identification} \tag{5.62}$$

The algorithm that I use to minimize equation 5.60 is Matlab's fmincon, with the interior point method. At each step of the search, it uses either a Newton-Raphson step or a conjugate gradient step. Convergence in this case is quadratic, and is much faster that the Baum-Welch algorithm which has linear convergence (see section 5.8).

The goodness-of-fit objective of equation 5.60 is simple enough that we can make some predictions about its use as a maximum likelihood estimator. In particular, it corresponds to taking the total sample surprisal as

$$\mathcal{L}(\theta) = \left(\frac{M}{N}\right) \sum_{|u|=N} \hat{p}(u)[-\log p(u,\theta)] \tag{5.63}$$

At the minimum where $\partial\mathcal{L}/\partial\vec{\theta} = 0$, and assuming we have $p(u,\theta) \approx \hat{p}(u)$ for all symbol blocks, we get for the information matrix

$$[I(\theta)]_{ij} \approx \frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j} \approx n_b \sum_{|u|=N} \frac{1}{p(u)} \frac{\partial p(u)}{\partial\theta_i} \frac{\partial p(u)}{\partial\theta_j} \quad (n_b = M/N) \tag{5.64}$$

Setting $\mathcal{L}'$ to be the entropy per symbol block,

$$\mathcal{L}' = \sum_{|u|=N} p(u,\theta)[-\log p(u,\theta)] \tag{5.65}$$

we can expect the bias of the maximum likelihood estimates to be at most

$$\langle\hat{\theta} - \theta\rangle = \frac{1}{n_b}(\text{ function of third derivatives of } \mathcal{L}') \tag{5.66}$$

and the correction to the parameter covariance to be at most

$$\text{cov}(\hat{\theta}) = I^{-1}(\theta) + \frac{1}{n_b^2}(\text{ function of third derivatives of } \mathcal{L}') \tag{5.67}$$

For further discussion of the higher-order properties of maximum likelihood estimators, see [30]. Our first diagnostic is to look for any possible bias beyond that allowed by equation 5.66. Here, the methodology is to generate $N_{\text{trial}} * N_{\text{batch}}$ output strings of length M. For each one, compute the empirical symbol block proportions $\hat{p}(u)$ for all $|u| = N$. Then, find the independent model parameters $\hat{\theta}$ that minimize equation 5.60 for the goodness of fit. For each batch, compute the mean parameter set $\hat{\theta}$.

Table 5.7: Lack of parameter bias for training done via chi-square goodness-of-fit for the HMM defined in equation 5.68. There should be no bias beyond that allowed by equation 5.66.

parameter	**actual value**	**95% confidence interval**
$\hat{a}_{12}$	0.20	[0.1994, 0.2004]
$\hat{a}_{21}$	0.30	[0.2988, 0.2999]
$\hat{b}_{11}$	0.67	[0.6701, 0.6708]
$\hat{b}_{12}$	0.20	[0.1994, 0.2000]
$\hat{b}_{21}$	0.20	[0.1985, 0.1995]
$\hat{b}_{22}$	0.57	[0.5700, 0.5708]

Table 5.7 shows the 95% confidence intervals for this diagnostic, with $N_{\text{trial}} = 112, N_{\text{batch}} = 64, M = 128000, N = 4$, and based on the two-state, three-symbol hidden Markov model

$$A = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}, \quad B = \begin{bmatrix} 0.67 & 0.20 & 0.13 \\ 0.20 & 0.57 & 0.23 \end{bmatrix} \quad (\lambda_2 = 0.50) \tag{5.68}$$

Note that three of the six independent parameters show no bias under maximum likelihood estimation, and the remaining ones show a bias well within the limits of equation 5.66.

Our next diagnostic is to check equation 5.67 for the predicted parameter variance. Our methodology here is to generate N_{trial} * N_{batch} output strings of length M. For each one, compute the empirical symbol block proportions $\hat{p}(u)$ for all $|u| = N$. Then, find the independent model parameters $\hat{\theta}$ that minimize equation 5.60 for the goodness of fit. For each trial, calculate the unbiased parameter covariance

$$\text{cov}(\hat{\theta}) = \frac{1}{N_{\text{batch}}} \sum_{\text{batch}} (\hat{\theta} - \vec{\theta}_0)(\hat{\theta} - \vec{\theta}_0)^T \tag{5.69}$$

where $\vec{\theta}_0$ represents the true values of the independent parameters. Finally, calculate the spread of $\text{cov}(\hat{\theta})$ over all N_{trial} trials.

Table 5.8 shows the results of that diagnostic applied to the two-state, three-symbol hidden Markov model of equation 5.55, with $M = 260000, N = \{3, 4\}, N_{\text{batch}} = 64$, and $N_{\text{trial}} = 96$. The table includes 95% confidence intervals for the empirical parameter variances, the predictions of equation 5.67 for $N = 3$ and $N = 4$, and also the prediction of equation 5.49 which would be applicable to the Baum-Welch estimation procedure.

Table 5.8: Parameter variance for training done via chi-square goodness-of-fit for the HMM defined by equation 5.55. Figures in square brackets are 95% confidence intervals. The actual values match up with the theoretical values predicted by equation 5.67.

specific	$M\text{var}(a_{12})$	$M\text{var}(a_{21})$	$M\text{var}(b_{11})$	$M\text{var}(b_{12})$	$M\text{var}(b_{21})$	$M\text{var}(b_{22})$
N = 3, theory	208	337	79.7	44.8	188	110
N = 4, theory	115	179	46.8	26.4	107	63.5
N = 8, Baum/Welch	45.04	65.53	20.23	11.44	43.03	26.73
N = 3, experiment	[199,210]	[318,338]	[77,82]	[43,46]	[182,193]	[107,114]
N = 4, experiment	[111,117]	[175,183]	[45,48]	[25.5,27]	[105,111]	[62,66]

What Table 5.8 shows is that we are indeed getting parameter variances as predicted by equation 5.67 for this estimator. Note also that in this case, the parameter variances are substantially greater than what we would get with the Baum-Welch algorithm. That is hardly a surprise. When we count up symbol blocks of size N, we are throwing away all information about symbol correlations beyond that size. That suggests using the goodness-of-fit method as an acceleration step, prior to starting the Baum-Welch algorithm.

5.10 Distance measures for hidden Markov models

Now that we know how to calculate the covariance of the independent model parameters of a hidden Markov model, we need to construct a distance measure that uses that covariance. In particular, if $\hat{\theta}$ is the independent parameter set obtained through maximum likelihood, and θ is the true independent parameter set, then we seek a function that looks like

$$T \approx (\hat{\theta} - \theta)^T (MQ)(\hat{\theta} - \theta) \quad \text{where } Q_{ij} = \lim_{N\to\infty} \left[\sum_{|u|=N} \frac{1}{p(u)} \frac{\partial p(u)}{\partial \theta_i} \frac{\partial p(u)}{\partial \theta_j} \right]_N^{N+1} \tag{5.70}$$

where $\hat{\theta} \approx \theta$, $p(u)$ is the ensemble symbol block probability under the true model θ, and M is the size of the data that the maximum likelihood estimate $\hat{\theta}$ is based on. The existence of the symbol block probability $p(u)$ in that formula recalls the Hellinger and dot-product

measures for histograms (equation 3.40), so we could make a "first stab" with

$$d_H^2 = 8M \lim_{N\to\infty} \left[1 - \sum_{|u|=N} \sqrt{p(u,\hat{\theta})p(u,\theta)} \right]_N^{N+1}$$
$$d_{DP}^2 = 8M \lim_{N\to\infty} \left[-\log \sum_{|u|=N} \sqrt{p(u,\hat{\theta})p(u,\theta)} \right]_N^{N+1} \tag{5.71}$$

both of which reduce to the required quadratic form at $\hat{\theta} \approx \theta$. The corresponding versions for models derived from different samples would be

$$d_H^2 = 8 \left(\frac{M_1 M_2}{M_1 + M_2} \right) \lim_{N\to\infty} \left[1 - \sum_{|u|=N} \sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)} \right]_N^{N+1}$$
$$d_{DP}^2 = 8 \left(\frac{M_1 M_2}{M_1 + M_2} \right) \lim_{N\to\infty} \left[-\log \sum_{|u|=N} \sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)} \right]_N^{N+1} \tag{5.72}$$

where $p_1(u,\hat{\theta}_1)$ and $p_2(u,\hat{\theta}_2)$ are the ensemble probabilities of symbol block u under models $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively. Note that in both cases, the underlying hidden Markov models do not appear directly. Thus, the measures are defined even when the two models involved have differing numbers of states. However, they must have the same number of distinct symbols.

Based on the analysis of section 5.8, we can say that for maximum-likelihood models generated from the same underlying hidden Markov model, d_H^2 and d_{DP}^2 will approximate a chi-square variate, with the number of degrees of freedom equal to the number of independent model parameters (i.e. $n(n+m-2)$ where n is the number of states, and m the number of distinct symbols). We can check this with the following diagnostic: for each of N_{trial} trials, generate two strings of lengths M_1 and M_2 from the same underlying hidden Markov model. Use the Baum-Welch algorithm to get maximum-likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Calculate d_{DP}^2 via equation 5.72, then compare the empirical cumulative distribution to the CDF expected for a chi-square variate with degrees of freedom equal to $n(n+m-2)$.

Figure 5.4 shows the results of that diagnostic procedure applied to the 2-state, 3-symbol hidden Markov model

$$A = \begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}, \quad B = \begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.50 & 0.30 \end{bmatrix} \quad (\lambda_2 = 0.50) \tag{5.73}$$

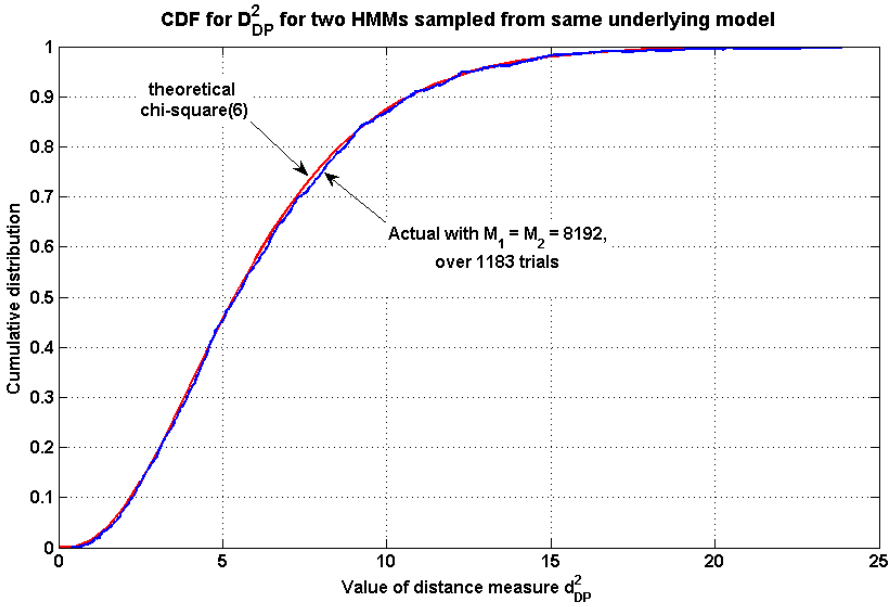


Figure 5.4: Cumulative distribution of d^2_{DP} for same underlying HMM, defined in equation 5.73.

with $M_1 = M_2 = 8192$ and $N_{\text{trial}} = 1183$. In this case, there are $n(n+m-2) = 6$ independent parameters. The one-sample Kolmogorov-Smirnov statistic for this case is 0.69, well within the 95% confidence limit of 1.35.

Our next task is to estimate the distribution of d^2_H and d^2_{DP} in the case where the underlying hidden Markov models are materially different. To that end, let θ_1 be the vector of independent parameters for model 1, and let θ_2 be the independent parameters for model 2. Note that θ_1 and θ_2 could have different lengths. Then, whether for d^2_H or for d^2_{DP}, we need the approximate mean and variance of

$$T_N = \left[\sum_{|u|=N} \sqrt{p_1(u, \hat{\theta}_1) p_2(u, \hat{\theta}_2)} \right]_N^{N+1} \tag{5.74}$$

with respect to variation over $\Delta\theta_1$ and $\Delta\theta_2$ where $\hat{\theta}_1 = \theta_1 + \Delta\theta_1$ and $\hat{\theta}_2 = \theta_2 + \Delta\theta_2$. I assume here that the models $\hat{\theta}_1$ and $\hat{\theta}_2$ are maximum likelihood estimates, based on data of sizes M_1 and M_2 respectively, and whose variance is given by equation 5.49. Expanding

p_1 and p_2 around their values at θ_1 and θ_2 yields

$$\sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)} \approx \sqrt{p_1(u,\theta_1)p_2(u,\theta_2)}\sqrt{1+\frac{1}{p_1}\sum_i \frac{\partial p_1}{\partial \theta_i}\Delta\theta_{1,i}+\ldots} \bullet \sqrt{1+\frac{1}{p_2}\sum_j \frac{\partial p_2}{\partial \theta_j}\Delta\theta_{2,j}+\ldots} \tag{5.75}$$

We can ignore any cross terms involving $\Delta\theta_{1,i}\Delta\theta_{2,j}$. Summing over the symbol blocks and taking a mean value over $\hat{\theta}_1$ and $\hat{\theta}_2$ yields

$$\langle T_N\rangle \approx \left[\sum_{|u|=N}\sqrt{p_1(u,\theta_1)p_2(u,\theta_2)}\right]_N^{N+1} - \frac{1}{8}\left\langle \sum_{|u|=N}\sqrt{\frac{p_2}{p_1}}\frac{1}{p_1}\sum_{i,j}\frac{\partial p_1}{\partial \theta_{1,i}}\frac{\partial p_1}{\partial \theta_{1,j}}\Delta\theta_{1,i}\Delta\theta_{1,j}\right\rangle_N^{N+1} - \frac{1}{8}\left\langle \sum_{|u|=N}\sqrt{\frac{p_1}{p_2}}\frac{1}{p_2}\sum_{i,j}\frac{\partial p_2}{\partial \theta_{2,i}}\frac{\partial p_2}{\partial \theta_{2,j}}\Delta\theta_{2,i}\Delta\theta_{2,j}\right\rangle_N^{N+1} \tag{5.76}$$

As a first-order approximation, we will take $\sqrt{p_1/p_2} \approx 1$, in which case equation 5.76 simplifies for large N to

$$\langle T_N\rangle \approx \left[\sum_{|u|=N}\sqrt{p_1(u,\theta_1)p_2(u,\theta_2)}\right]_N^{N+1} - \frac{1}{8}\left(\frac{|\theta_1|}{M_1}+\frac{|\theta_2|}{M_2}\right) \tag{5.77}$$

where $|\theta_1|$ and $|\theta_2|$ are the number of independent parameters in θ_1 and θ_2 respectively.

As for the approximate variance of T_N in equation 5.74, that is expressible in terms of the known variances of $\hat{\theta}_1$ and $\hat{\theta}_2$:

$$\begin{aligned} \mathrm{var}(T_N) &\approx \left(\frac{\partial T_N}{\partial \theta_1}\right)^T \mathrm{var}(\hat{\theta}_1)\left(\frac{\partial T_N}{\partial \theta_1}\right) + \left(\frac{\partial T_N}{\partial \theta_2}\right)^T \mathrm{var}(\hat{\theta}_2)\left(\frac{\partial T_N}{\partial \theta_2}\right) \\ \mathrm{var}(\hat{\theta}_k) &= \frac{1}{M_k}Q_k^{-1}, \quad [Q_k]_{ij} = \left[\sum_{|u|=N}\frac{1}{p(u)}\frac{\partial p(u)}{\partial \theta_{k,i}}\frac{\partial p(u)}{\partial \theta_{k,j}}\right]_N^{N+1} \end{aligned} \tag{5.78}$$

where

$$\frac{\partial T_N}{\partial \theta_1}(\theta_1,\theta_2) = \sum_i \frac{1}{2}\left[\sum_{|u|=N}\sqrt{\frac{p(u,\theta_2)}{p(u,\theta_1)}}\frac{\partial p(u,\theta_1)}{\partial \theta_{1,i}}\right]_N^{N+1}\hat{\imath}$$
$$\frac{\partial T_N}{\partial \theta_2}(\theta_1,\theta_2) = \sum_j \frac{1}{2}\left[\sum_{|u|=N}\sqrt{\frac{p(u,\theta_1)}{p(u,\theta_2)}}\frac{\partial p(u,\theta_2)}{\partial \theta_{2,j}}\right]_N^{N+1}\hat{\jmath} \tag{5.79}$$

Next, we will summarize those predictions for the dot-product measure d^2_{DP} of equation 5.72 and compare with empirical results. The predictions for mean are:

$$\begin{aligned} T &= \lim_{N\to\infty}\left[-\log\sum_{|u|=N}\sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)}\right]_N^{N+1} \\ &= \lim_{N\to\infty}\left[-\log S_N(\hat{\theta}_1,\hat{\theta}_2)\right]_N^{N+1} \text{ where } S_N(\theta_1,\theta_2) = \sum_{|u|=N}\sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)} \\ \langle T\rangle &\approx \lim_{N\to\infty}\left[-\log\sum_{|u|=N}\sqrt{p_1(u,\theta_1)p_2(u,\theta_2)}\right]_N^{N+1} + \frac{1}{8}\left(\frac{|\theta_1|}{M_1}+\frac{|\theta_2|}{M_2}\right) \end{aligned} \tag{5.80}$$

and the predictions for variance are:

$$\begin{aligned} \mathrm{var}(T) &\approx \left(\frac{\partial T}{\partial \theta_1}\right)^T \mathrm{var}(\hat{\theta}_1)\left(\frac{\partial T}{\partial \theta_1}\right) + \left(\frac{\partial T}{\partial \theta_2}\right)^T \mathrm{var}(\hat{\theta}_2)\left(\frac{\partial T}{\partial \theta_2}\right) \\ \mathrm{var}(\hat{\theta}_k) &= \frac{1}{M_k}Q_k^{-1}, \quad [Q_k]_{ij} = \lim_{N\to\infty}\left[\sum_{|u|=N}\frac{1}{p(u)}\frac{\partial p(u)}{\partial \theta_{k,i}}\frac{\partial p(u)}{\partial \theta_{k,j}}\right]_N^{N+1} \end{aligned} \tag{5.81}$$

where

$$\frac{\partial T}{\partial \theta_1} = \sum_i \frac{1}{2}\lim_{N\to\infty}\left[\frac{1}{S_N(\theta_1,\theta_2)}\sum_{|u|=N}\sqrt{\frac{p(u,\theta_2)}{p(u,\theta_1)}}\frac{\partial p(u,\theta_1)}{\partial \theta_{1,i}}\right]_N^{N+1}\hat{\imath}$$
$$\frac{\partial T}{\partial \theta_2} = \sum_j \frac{1}{2}\lim_{N\to\infty}\left[\frac{1}{S_N(\theta_1,\theta_2)}\sum_{|u|=N}\sqrt{\frac{p(u,\theta_1)}{p(u,\theta_2)}}\frac{\partial p(u,\theta_2)}{\partial \theta_{2,j}}\right]_N^{N+1}\hat{\jmath} \tag{5.82}$$

Table 5.9: Representative model pairs for HMM distance measure d^2_{DP}

Case	A_1	B_1	M_1	A_2	B_2	M_2
1	$\begin{bmatrix} 0.85 & 0.15 \\ 0.35 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.50 & 0.30 \end{bmatrix}$	8192	$\begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$	(Same as B_1)	8192
2	[1]	[0.5 0.3 0.2]	8192	$\begin{bmatrix} 0.80 & 0.20 \\ 0.30 & 0.70 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.50 & 0.30 \end{bmatrix}$	8192
3	$\begin{bmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{bmatrix}$	$\begin{bmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{bmatrix}$	8192	$\begin{bmatrix} 0.20 & 0.80 \\ 0.80 & 0.20 \end{bmatrix}$	(Same as B_1)	8192

I made that comparison for three representative cases, which are detailed in Table 5.9. In the first case, the hidden Markov models are the same size, and slightly different in their transition probabilities. In the second case, the models are of different size, but still similar in their symbol block probabilities. In the third case, the models are the same size, but very different in their symbol block probabilities. The comparison methodology is as follows: for each of N_{trial} trials, generate output strings of length M_1 for model $\{A_1, B_1\}$ and of length M_2 for model $\{A_2, B_2\}$. Run the Baum/Welch algorithm, starting with the true model parameters, to get the maximum-likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Calculate the target

$$T = \lim_{N\to\infty} \left[-\log \sum_{|u|=N} \sqrt{p_1(u,\hat{\theta}_1) p_2(u,\hat{\theta}_2)} \right]_N^{N+1} \tag{5.83}$$

using $N_{\max} = 9$. Compute the mean and variance of the target T over all trials and compare to the predictions of equation 5.80. Those predictions involved some simplifying assumptions, so we are only looking for broad agreement, not exact agreement. Table 5.10 shows the results that I got for these three cases, along with predictions for the correponding type II error rates. The table does indeed show broad agreement between the predicted and actual spread of the hidden Markov model distance measure for the three cases. However, we should not get too complacent about those type II error rates. They depend very heavily on the data sizes M_1 and M_2. For example, if in Case 1 we had $M_1 = M_2 = 2048$ instead of 8192, then $\text{var}(\hat{\theta}_1)$ and $\text{var}(\hat{\theta}_2)$ (and hence $\text{var}(T)$) would quadruple by equation 5.80, and the resulting type II error rate would increase to approximately 0.136.

Table 5.10: Predicted and actual spread for HMM distance measure d^2_{DP}, for cases outlined in Table 5.9

Case	N_{trials}	$\langle T \rangle$		var(T)		type II error
		predicted	actual	predicted	actual	predicted
1	921	30.2 x 10^{-4}	28.0 x 10^{-4}	33 x 10^{-8}	35 x 10^{-8}	2.3 x 10^{-5}
2	973	29.5 x 10^{-4}	29.4 x 10^{-4}	28 x 10^{-8}	28 x 10^{-8}	6.5 x 10^{-7}
3	1024	212 x 10^{-4}	212 x 10^{-4}	238 x 10^{-8}	242 x 10^{-8}	2.6 x 10^{-41}

5.11 Summary

1. For a stationary discrete hidden Markov model with primitive transition probability matrix A and symbol emission probability matrix B, both of which are row-stochastic, the ensemble probability of a symbol block $\{v_{j_1}, \ldots, v_{j_N}\}$ is

$$p(j_1 \ldots j_k) = \vec{a} D(\vec{b}_{j_1}) A D(\vec{b}_{j_2}) \ldots A D(\vec{b}_{j_k}) \vec{1}^T \tag{5.84}$$

   where $\vec{a}$ is the stationary distribution (section 5.4) and $D(\vec{b}_j)$ is a diagonal matrix containing the j-th column of B (section 5.5).

2. If we take a long output string from a discrete hidden Markov model, say $\{v_{j_1}, \ldots, v_{j_M}\}$ where $M \gg 1$, divide it into adjoining blocks of size N, and count the number of times that each distinct block appears, then the resulting counts will have the covariance

$$\begin{gathered}(N/M)\text{cov}(T_u, T_v) = [p(u)\delta_{uv} - p(u)p(v)] + 2 \sum_{k=1}^{M/N-1} [s_{2,k} - p(u)p(v)] \\ \text{where } s_{2,1} = \frac{1}{2}[p(uv) + p(vu)],\ s_{2,2} = \frac{1}{2}[p(u\#v) + p(v\#u)] \quad \text{etc}\end{gathered} \tag{5.85}$$

   where $T(u)$ is the count for word u, $p(u)$ is the ensemble probability for word u, and $\#$ represents an arbitray word of length N.

3. The information matrix for a discrete hidden Markov model derived by maximum likelihood estimation from an output string of length M, is

$$M[i(\theta)]_{ij} = M \lim_{N\to\infty} \left[\sum_{|u|=N} \frac{1}{p(u)} \frac{\partial p(u)}{\partial \theta_i} \frac{\partial p(u)}{\partial \theta_j} \right]_N^{N+1} \tag{5.86}$$

where $p(u)$ is the ensemble probability of word u, and θ_i, θ_j are members of the discrete HMM independent parameter set θ. That formula correctly predicts the parameter variance for models estimated via the Baum-Welch algorithm (see section 5.8).

4. There are several distance measures that reduce to $(\Delta\theta)^T Mi(\theta)(\Delta\theta)$ for $\Delta\theta \approx 0$, where θ represents the independent parameter set of a discrete hidden Markov model (see equation 5.73). In my calculated examples (section 5.10) I used

$$d_{DP}^2 = 8\left(\frac{M_1 M_2}{M_1 + M_2}\right) \lim_{N\to\infty} \left[-\log \sum_{|u|=N} \sqrt{p_1(u,\hat{\theta}_1) p_2(u,\hat{\theta}_2)} \right]_N^{N+1} \tag{5.87}$$

Chapter 6

Results and further research directions

What follows is a summary of the results of this thesis that I consider to be contributions to knowledge. Those marked with an asterisk contribute directly to the overall thesis goal, as described in Chapter 1. Those without an asterisk support the overall thesis goal indirectly.

1. For a cluster of parameter sets $\hat{\theta}$ derived by maximum likelihood estimation from large datasets generated by an underlying model $\vec{\theta_0}$: if we hold the mean of a distance function $f(\hat{\theta}, \vec{\theta_0})$ constant, then the minimum variance is achieved when $f(\hat{\theta}, \vec{\theta_0})$ is a function of $|\hat{\theta} - \vec{\theta_0}|$ only (see section 3.3).

2. * For clusters C_0 and C_1 of maximum likelihood estimates, based on underlying models $\vec{\theta_0}$ and $\vec{\theta_1}$ respectively, the type II error rate for an L_k distance measure other than L_2 may be lowest when the cluster separation $\vec{\theta_1} - \vec{\theta_0}$ is in a particular direction, but when averaged over all directions, the type II error rate is lowest for the L_2 distance based on the Fisher information matrix (see section 3.4 and Appendix G).

3. * Similarly, for clusters C_0 and C_1 of maximum likelihood estimates, based on underlying models $\vec{\theta_0}$ and $\vec{\theta_1}$ respectively, the type II error rate for a weighted L_2 distance may be lowest when the cluster separation $\vec{\theta_1} - \vec{\theta_0}$ is in one of the boosted directions, but when averaged over all directions, the type II error rate is lowest for the unweighted $L_2 = |\vec{z}\,|$ where $\vec{z} = L\triangle\theta$, $L^T L = ni(\theta_0)$, and n is the sample size (see

section 3.5 and Appendix G). A boosted direction is one weighted more than the average.

4. For histograms, the Cramer von Mises distance is an unevenly weighted L_2 distance, and the Earth Mover Distance is an unevenly weighted L_1 distance. As such, their directionally-averaged type II error will exceed that of an unweighted L_2 distance based on the information matrix (see Section 3.5).

5. When comparing a histogram to the true probability distribution $p(x)$ via the Hellinger metric, the optimum number of equally-sized bins is proportional to $N^{1/3}$ (N being the sample size) and to the membrane energy of $\sqrt{p(x)}$ (see section 3.6).

6. * A generalized L_2 distance measure between maximum likelihood models θ_1 and θ_2, based on samples of size n is:

$$d^2 = \sum_s \log^2 \frac{p(v_s, \theta_1)}{p(v_s, \theta_2)} \tag{6.1}$$

where the summation is over the independent (or nearly independent) events $\{v_s\}$ contained in a sample (see section 4.1). By construction, this measure matches the ideal L_2 quadratic form when $\theta_1 \approx \theta_2$.

7. * Using equation 6.1, I derived the information matrix for an ARMA(p, q) time series model, namely

$$[\frac{1}{n} i(\vec{\theta})]_{ij} = \begin{cases} 1/(1 - \theta_i \theta_j) \text{ if both are AR or MA} \\ -1/(1 - \theta_i \theta_j) \text{ if one is AR, one is MA} \end{cases} \tag{6.2}$$

where $\vec{\theta}$ is the concatenation of the autoregressive and moving average roots of the ARMA(p, q) model. Note: the contributions to knowledge are the proofs in Appendix D and Appendix E, not the formula itself.

8. * From the information matrix of equation 6.2, I derived the integrated L_2 distance

$$d_I^2(\theta_1, \theta_2) = \frac{n_1 n_2}{n_1 + n_2} \sum_{k=1}^{\infty} \frac{1}{k^2} \left[\sum_{i=1}^{p_1} r_{1i}^k - \sum_{j=1}^{q_1} s_{1j}^k - \sum_{i=1}^{p_2} r_{2i}^k + \sum_{j=1}^{q_2} s_{2j}^k \right]^2 \tag{6.3}$$

for ARMA(p, q) models with ARMA root representations $\{[r_{11}, \ldots, r_{1,p_1}], [s_{11}, \ldots, s_{1,q_1}], n_1\}$ and $\{[r_{21}, \ldots, r_{2,p_2}], [s_{21}, \ldots, s_{2,q_2}], n_2\}$.

9. For a discrete hidden Markov model with state transition probability matrix A and symbol emission probability matrix B, the covariance of the counts T_u and T_v of two symbol blocks u and v (of equal length N) in an output string of length M is

$$(N/M)\mathrm{cov}(T_u, T_v) = [p(u)\delta_{uv} - p(u)p(v)] + 2\sum_{k=1}^{M/N-1}[s_{2,k} - p(u)p(v)]$$
$$\text{where } s_{2,k} = \frac{1}{2}[p(u\#\ldots\#v) + p(v\#\ldots\#u)] \quad (k-1 \text{ arbitrary symbol blocks }) \tag{6.4}$$

Here, $p(u)$ is the probability of word u under the model $\{A, B\}$, and $\#$ represents an arbitrary word of length N (see section 5.6 and Appendix F).

10. The chi-square "goodness of fit" quantity

$$\chi^2 = \frac{M}{N}\sum_{|u|=N}\frac{[\hat{p}(u) - p(u)]^2}{p(u)} \tag{6.5}$$

where $p(u)$ is the expected occurrence rate of word u under a hidden Markov model $\{A, B\}$ and $\hat{p}(u)$ is its actual occurrence rate in an output string of length M, will be close to a true chi-square variate on $m^N - 1$ degrees of freedom when $4\lambda_2^2/m^N \ll 1$, where λ_2 is the subdominant eigenvalue of the transition probability matrix A, m is the number of distinct output symbols, and N is the length of word u (see section 5.7).

11. * The information matrix for a hidden Markov model with state transition probability matrix A and symbol emission probability matrix B is

$$\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j} = M\lim_{N\to\infty}\left[\sum_{|u|=N}\frac{1}{p(u)}\frac{\partial p(u)}{\partial\theta_i}\frac{\partial p(u)}{\partial\theta_j}\right]_N^{N+1} \tag{6.6}$$

where θ collects together the independent components of A and B, M is the length of the output string that the model is based on, and $p(u)$ is the occurrence rate of a word u under the model $\{A, B\}$ (see section 5.8).

12. * The Bhattacharya distance measure corresponding to equation 6.6 is

$$d_{DP}^2 = 8\left(\frac{M_1 M_2}{M_1 + M_2}\right)\lim_{N\to\infty}\left[-\log\sum_{|u|=N}\sqrt{p_1(u,\hat{\theta}_1)p_2(u,\hat{\theta}_2)}\right]_N^{N+1} \tag{6.7}$$

where $p_1(u, \hat{\theta}_1)$ and $p_2(u, \hat{\theta}_2)$ are the occurrence probabilities of the word u (of length N) under the two hidden Markov models $\hat{\theta}_1$ and $\hat{\theta}_2$. Equation 5.80 also gives an approximation for the spread of d^2_{DP} when the hidden Markov models are materially different.

13. In Appendix G, I develop approximations for the distribution functions of the central L_k distance and the non-central L_k distance, as linear combinations of the distribution functions for chi-square variates and for non-central chi-square variates respectively. The approximation for the central L_k distribution predicts the T_1 thresholds for $T_1 = 0.05$ to 0.3% accuracy for $\{m \leq 3, k \leq 4\}$ where m is the number of dimensions involved (see Appendix G, in particular equations G.40 and G.45).

My main question in this thesis was: how can the principles of information theory and statistics guide us in formulating distance measures for probabilistic models obtained through maximum likelihood estimation? We have seen that information theory and statistics can definitely guide us. In particular, the L_2 distance based on the information matrix of equation 2.3 will have the lowest spatially-averaged type II error for the T_1 threshold that I have been using in this thesis ($T_1 = 0.05$). However, information theory does not answer all questions, in particular questions about model or feature relevance. Those require human judgment.

This work has suggested many directions for future research, and here are my favorites:

1. Equation G.45 needs a much better approximation to the non-central distribution for $d_1 = \sum_{i=1}^{m} |z_i - v_i|^k$ for a specific offset vector $\vec{v}$, for $\vec{z} \sim N(0, I_m)$. That requires a model for the variation of $|z_i - v_i|^k$ over the secondary angles, not just over the principal angle from the all-ones axis. The goal is to develop an approximation for the offset magnitude $\lambda = v^2[1 + h(m)\epsilon^2]$.

2. If we add a feature to a feature vector, or add a bin to a histogram, then we push out both the same-cluster distance distribution and the different-cluster distribution (see section 3.7, in particular Figure 3.11). If there is no corresponding increase in offset, then the type II error will increase. There must be some minimum increase in the offset to justify the inclusion of another feature or bin. The goal is to calculate that minimum increase in offset, as a function of the cluster centres and sample sizes.

3. In chapter 4 on ARMA(p, q) time series models, I converted a GARCH(1,1) model into an ARMA(1,1) model for the purposes of further analysis. Can we express the likelihood of a dataset directly in terms of the parameters of a GARCH(p, q) model,

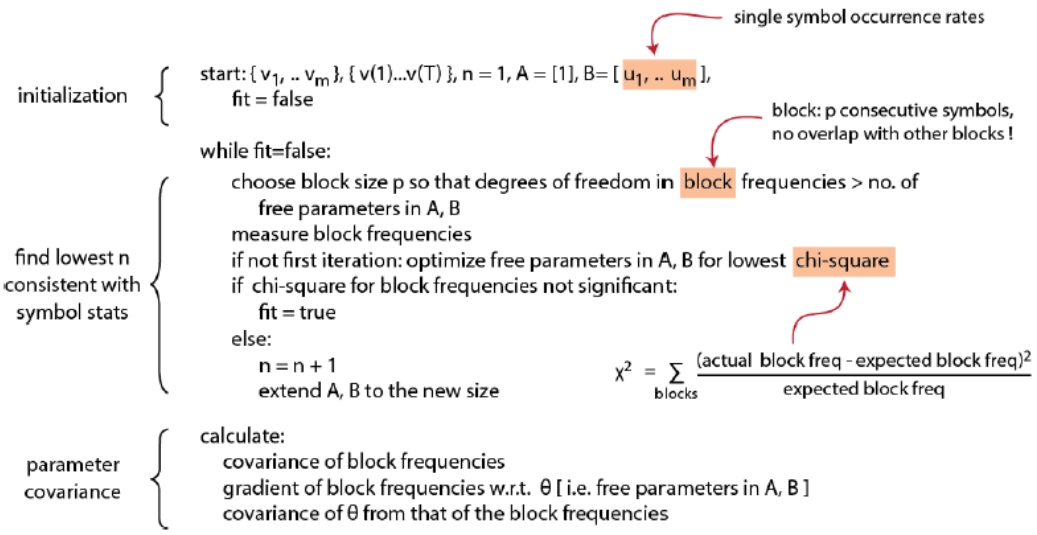


Figure 6.1: Proposed algorithm for finding the smallest number of states for a hidden Markov model

without doing a conversion? For further information on this problem, see Lumsdaine [76].

4. In chapter 5 on hidden Markov models, I gave an expression for the information matrix involving a limit for symbol block size (equation 6.6) and pointed out that in certain situations, it is impractical to calculate. In particular, when the subdominant eigenvalue of the transition probability matrix is close to one, then we would have to calculate an exponentially large number of gradients. The goal is to estimate that sum by doing a random sampling, or by estimating the spread of those gradient values in some other way.

5. Chapter 5 on hidden Markov models does not address the issue of model size selection. In particular, how do we know when the number of states n is "large enough"? A reasonable starting point is: the number of states n is sufficient when the chi-square goodness-of-fit statistic of equation 6.5 is below our required significance threshold for any symbol block size. That suggests an algorithm for determining the number of states, which is described in Figure 6.1. The research goal would be to test the efficacy of the algorithm, and determine the circumstances under which it overestimates or underestimates n.

6. In section 3.3 I showed the importance of minimizing the variance of the distance function $f(\hat{\theta}_{11}, \hat{\theta}_{12})$ when we constrain its mean (here, $\hat{\theta}_{11}$ and $\hat{\theta}_{12}$ are maximum-likelihood estimates from a cluster based on model θ_1). That theory would suggest doing the following for a non-Gaussian cluster: (A) Find the transform (rotation and scaling) that makes the covariance as close as possible to I_m where m is the number of dimensions (B) Estimate the remaining cluster shape as $1 + \sum Y_{\vec{j}}(\vec{\phi})$ for suitable spherical harmonics defined on the spherical angles $\vec{\phi}$. (C) The distance measure is then $(\triangle\hat{\theta})^T \Sigma^{-1} [1 + \sum Y_{\vec{j}}(\vec{\phi})]^{-1} (\triangle\hat{\theta})$ where Σ is the cluster covariance, and $\vec{\phi}$ are the spherical angles associated with $\triangle\hat{\theta}$. That allows a tighter T_1 threshold, and possibly a lower type II error when averaged over direction. For an example of this style of algorithm, see [111].

7. Similarity between two-dimensional silhouettes comes into play when we are segmenting images into semantic objects. The goal here is to define a distance measure between two closed two-dimensional curves, derived from electronic images. There are several issues involved here: (A) Discovering an object boundary is in itself a difficult algorithm. A boundary involves contiguous pixels that represent a change of color or a change of texture. However, we also want object boundaries with minimum curvature, so as to be as simple as possible. See Liu [75] for further discussion of this issue. (B) Given an object silhouette, there is a fundamental problem of contour representation. We would like a single, canonical definition of a closed curve in two dimensions, but there is no obvious candidate. We could consider a contour as a collection of contiguous line segments [58]. We could represent a contour as a shock graph [101], or as a set of C^1 splines. I favor representing a contour as a graph of curvature vs. arc length, with primitive elements consisting of delta functions (for sharp corners) and quadratic ranges (for everything else). (C) Given a silhouette representation, we need to find the independent events, or nearly independent events, whose total likelihood depends upon the parameters of the representation. An obvious candidate would be the pixels of the original boundary, along with their discrete slope and curvature.

References

[1] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*, pages 944–945. Dover Publications, 1965.

[2] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*, pages 940–944. Dover Publications, 1965.

[3] T. W. Anderson and D. A. Darling. Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[4] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[5] J. K. Baker. Trainable grammars for speech recognition. *Journal of the Acoustical Society of America*, 65, 1979.

[6] Michele Basseville. Divergence measures for statistical data processing - an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

[7] Leonard Baum, Ted Petrie, George Soulos, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[8] Johannes Bausch. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *Journal of Physics A: Mathematical and Theoretical*, 46(50):1–19, 2013.

[9] Morten Bay. Social media ethics: A Rawlsian approach to hypertargeting and psychometrics in political and commercial campaigns. *ACM Transactions Social Computing*, 1(4), 2018.

[10] Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization*, pages 65–67. Athena Scientific and Dynamic Ideas, 1997.

[11] L. M. Blumenthal. *Theory and Applications of Distance Geometry*, pages 7–12. Chelsea Publishing Company, 1970. semimetric.

[12] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[13] Jean Bourgain, Van Vu, and Philip Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, (258):559–603, 2010.

[14] George Box, Gwilym Jenkins, and Gregory Reinsel. *Time Series Analysis: Forecasting and Control 4th edition*, pages 2–7. John Wiley and Sons, 2008.

[15] George Box, Gwilym Jenkins, and Gregory Reinsel. *Time Series Analysis: Forecasting and Control 4th edition*, pages 79–86. John Wiley and Sons, 2008.

[16] George Box, Gwilym Jenkins, and Gregory Reinsel. *Time Series Analysis: Forecasting and Control 4th edition*, pages 35–43. John Wiley and Sons, 2008.

[17] George Box, Gwilym Jenkins, and Gregory Reinsel. *Time Series Analysis: Forecasting and Control 4th edition*, pages 264–267. John Wiley and Sons, 2008.

[18] Peter Brockwell and Richard Davis. *Introduction to Time Series and Forecasting*, pages 73–96. Springer, 2016.

[19] Jorge Caiado, Nuno Crato, and Daniel Pena. A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, 50:2668–2684, 2006.

[20] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*, pages 621–625. Springer-Verlag, 1986.

[21] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[22] Taolue Chen and Stefan Kiefer. On the total variation distance of labelled Markov chains. *Computer Science Logic and Logic in Computer Science 2014*, 33, 2014.

[23] Andrzej Cichocki and Shunichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Theory of Probability and its Applications*, (12):1532–1568, 2010.

[24] William Cochran. The chi-square test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345, 1952.

[25] Marcella Corduas and Domenico Piccolo. Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, 52:1860–1872, 2008.

[26] Corinna Cortes, Mehryar Mohri, and Ashish Rastogi. Lp distance and equivalence of probabilistic automata. *International Journal of Foundations of Computer Science*, 18(4):761–779, 2007.

[27] Corinna Cortes, Mehryar Mohri, Ashish Rastogi, and Michael Riley. On the computation of the relative entropy of probabilistic automata. *International Journal of Foundations of Computer Science*, 19(1):219–242, 2008.

[28] Thomas Cover and Joy Thomas. *Elements of Information Theory 2nd Edition*, pages 25–30. Wiley Interscience, 2006.

[29] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*, pages 279–311. Chapman and Hall, 1974.

[30] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*, pages 309–311. Chapman and Hall, 1974.

[31] D. A. Darling. The Kolmogorov-Smirnov, Cramer-von Mises tests. *Annals of Mathematical Statistics*, 28(4):823–838, 1957.

[32] Yoram Gdalyahu David Jacobs, Daphna Weinshall. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

[33] P. L. de Micheaux. *Package CompQuadForm*. CRAN Repository, 2017.

[34] Yuxin Deng and Wenjie Du. The Kantorovich metric in computer science: A survey. *Electronic Notes in Theoretical Computer Science*, 252:73–82, 2009.

[35] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[36] Minh N. Do. Fast approximation of Kullback Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 10(4):115–118, 2003.

[37] Richard Duda, Peter Hart, and David Stork. *Pattern Classification 2nd Edition*, pages 458–461. Wiley Interscience, 2001.

[38] Richard Duda, Peter Hart, and David Stork. *Pattern Classification 2nd Edition*, pages 7–8. Wiley Interscience, 2001.

[39] Richard Duda, Peter Hart, and David Stork. *Pattern Classification 2nd Edition*, pages 24–26. Wiley Interscience, 2001.

[40] Richard Duda, Peter Hart, and David Stork. *Pattern Classification 2nd Edition*, pages 128–138. Wiley Interscience, 2001.

[41] P. Dupont, F. Denis, and Y. Esposito. Links between probabilistic automata and hidden Markov models: probability distributions, learning models and induction algorithms. *Pattern Recognition*, 38:1349–1371, 2005.

[42] James Durbin. *Distribution theory for tests based on the sample distribution function*, pages 10–14. Society for Industrial and Applied Mathematics, 1973.

[43] James Durbin and M. Knott. Components of Cramer-von Mises statistics I. *Journal of the Royal Statistical Society Series B*, 34(2):290–307, 1972.

[44] Costas Efthimiou and Christopher Frye. *Spherical Harmonics in p Dimensions*, pages 39–62. World Scientific, 2014.

[45] Elsayed Elamir and Allan Seheult. Trimmed l-moments. *Computational Statistics and Data Analysis*, 43:299–314, 2003.

[46] Scott Eliason. *Maximum Likelihood Estimation: Logic and Practice*, pages 7–18. Sage Publications, 1993.

[47] Robert Elliott, Lakhdar Aggoun, and John Moore. *Hidden Markov Models: Estimation and Control*, pages 247–289. Springer-Verlag, 1995.

[48] Yariv Ephraim and Neri Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569, 2002.

[49] Yariv Ephraim and Neri Merhav. Hidden markov processes. *IEEE Transactions on Information Theory*, 48(6):1531, 2002.

[50] R. Fagin and L. Stockmeyer. Relaxing the triangle inequality in pattern matching. *International Journal of Computer Vision*, 30(3):219–231, 1998.

[51] Markus Falkhausen, Herbert Reininger, and Dietrich Wolf. Calculation of distance measures between hidden Markov models. *Eurospeech 95*, pages 1487–1490, 1995.

[52] R. A. Fisher. Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society Series 2*, 30(1):199–238, 1930.

[53] Georgios Giannakis and Sanyogita Shamsunder. Information theoretic criteria for non-Gaussian ARMA order determination and parameter estimation. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages IV996 – IV999, 1993.

[54] C. W. J. Granger and M. J. Morris. Time series modelling and interpretation. *Journal of the Royal Statistical Society Series A (General)*, 139(2):246–257, 1976.

[55] Ulf Grenander and Michael Miller. *Pattern Theory: From Representation to Inference*, pages 154–173. Oxford University Press, 2007.

[56] Ulf Grenander and Michael Miller. *Pattern Theory: From Representation to Inference*, pages 1–2. Oxford University Press, 2007.

[57] Ulf Grenander and Michael Miller. *Pattern Theory: From Representation to Inference*, pages 92–94. Oxford University Press, 2007.

[58] Ulf Grenander and Michael Miller. *Pattern Theory: From Representation to Inference*, pages 183–190. Oxford University Press, 2007.

[59] E. J. Hannan and L. Kavalieris. A method for autoregressive-moving average estimation. *Biometrika*, 72(2):273–280, 1984.

[60] Clifford Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[61] J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426, 1961.

[62] Mortaza Jamshidian and Robert Jennrich. Acceleration of the EM algorithm by using quasi-Newton methods. *Journal of the Royal Statistical Society Series B*, 59(3):569–587, 1997.

[63] Eric Jondeau, Michael Rockinger, and Ser-Huang Poon. *Financial Modeling Under Non-Gaussian Distributions*, pages 94–108. Springer Finance, 2007.

[64] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408, 1985.

[65] Dan Kalman. A matrix proof of Newton's identities. *Mathematics Magazine*, 73(4):313–315, 2000.

[66] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta. Distance measures for effective clustering of ARIMA time series. *Proceedings 2001 IEEE International Conference on Data Mining*, pages 273–280, 2001.

[67] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, page 44. Prentice Hall, 1993.

[68] Steve Kirkland. On the sequence of powers of a stochastic matrix with large exponent. *Linear Algebra and its Applications*, 310:109–122, 2000.

[69] Timo Koski. *Hidden Markov Models for Bioinformatics*, pages 231–262. Kluwer Academic Publishers, 2001.

[70] Joseph Kruskal. An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Review*, 25(2):201–237, 1983.

[71] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.

[72] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, 1983.

[73] T. Warren Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38:1857–1874, 2005.

[74] Jiianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[75] Dorothy Liu, Christian Scharfenberger, Khalil Fergani, Alexander Wong, and David Clausi. Enhanced decoupled active contour using structural and textural variation energy functionals. *IEEE Transactions on Image Processing*, 23(2):855–869, 2014.

[76] Robin Lumsdaine. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica*, 64(3):575–596, 1996.

[77] Helmut Lutkepohl. *New Introduction to Multiple Time Series Analysis*, pages 28–31. Springer, 2005.

[78] Theodore Lystig and James Hughes. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computatinal and Graphical Statistics*, 11(3):678689, 2002.

[79] David J. Marchette. *Random Graphs for Statistical Pattern Recognition*, pages 7–17. Wiley Interscience, 2004.

[80] Richard Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48(4):1164–1170, 2000.

[81] P. Massart. The tight constant in the Dvoretsky Kiefer Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[82] A. M. Mathai and S. B. Provost. *Quadratic forms in random variables: theory and applications*, pages 15–18. Dekker, 1992.

[83] Carl Meyer. *Matrix Analysis and Applied Linear Algebra*, page 703. Society for Industrial and Applied Mathematics, 2000.

[84] Carl Meyer. *Matrix Analysis and Applied Linear Algebra*, page 517. Society for Industrial and Applied Mathematics, 2000.

[85] Carl Meyer. *Matrix Analysis and Applied Linear Algebra*, page 612. Society for Industrial and Applied Mathematics, 2000.

[86] Alan Oppenheim and Ronald Schafer. *Digital Signal Processing*, pages 500–511. Prentice-Hall International, 1975.

[87] Alan Oppenheim and Ronald Schafer. *Digital Signal Processing*, pages 88–120. Prentice-Hall International, 1975.

[88] Edoardo Otranto. Clustering heteroskedastic time series by model-based procedures. *Computational Statistics and Data Analysis*, 52:4685–4698, 2008.

[89] Marc Paolella. *Linear Models and Time Series Analysis*, pages 287–294. Wiley, 2019.

[90] Marc Paolella. *Linear Models and Time Series Analysis*, pages 525–574. Wiley, 2019.

[91] Azaria Paz. *Introduction to Probabilistic Automata*, pages 116–126. Academic Press, 1971.

[92] Domenico Piccolo. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11(2):153–164, 1990.

[93] Boaz Porat and Benjamin Friedlander. Computation of the exact information matrix of Gaussian time series with stationary random components. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(1):118–130, 1986.

[94] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[95] Svetlozar Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Entropy*, 29(4):647–676, 1985.

[96] Svetlozar Rachev, Lev Klebanov, Stoyan Stoyanov, and Frank Fabozzi. *The Methods of Distances in the Theory of Probability and Statistics*, page 119. Springer, 2013.

[97] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[98] Elena Ruiz, J. Ramirez, J. M. Gorriz, and J. Casillas. Alzheimer's disease computer-aided diagnosis: Histogram-based analysis of regional MRI volumes for feature selection and classification. *Journal of Alzheimer's Disease*, 65(3):819–842, 2018.

[99] David Ruppert. *Statistics and Data Analysis for Financial Engineering*, pages 483–491. Springer Science and Business Media, 2011.

[100] E. Seneta. *Non-Negative Matrices: An Introduction to Theory and Applications*, page 7. George Allen and Unwin Ltd, 1973.

[101] Kaleem Siddiqui, Ali Shokoufandeh, Sven Dickinson, and Steven Zucker. Shock graphs and shape matching. *International Journal of Computer Vision*, 35(1):13–32, 1999.

[102] Roland Siegwart, Illah R. Nourbakhsh, and Davide Scaramuzza. *Introduction to Autonomous Mobile Robots 2nd Edition*, pages 145–180. MIT Press, 2011.

[103] Jorge Silva and Shrikanth Narayanan. Average divergence distance as a statistical discrimination measure for hidden Markov models. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):890–906, 2006.

[104] Daniel Straumann. *Estimation in Conditionally Heteroscedastic Time Series Models*, pages 1–2. Springer, 2005. GARCH introduction.

[105] Alan Stuart and J. Keith Ord. *Kendall's Advanced Theory of Statistics 5th Edition*, pages 649–679. Oxford University Press, 1987.

[106] Alan Stuart and J. Keith Ord. *Kendall's Advanced Theory of Statistics 5th Edition*, pages 385–410. Oxford University Press, 1987.

[107] Andrew Stumer. *The presumption of innocence: evidential and human rights perspectives*, pages 27–40. Hart Publishing, 2010.

[108] P. V. Sukhatme. Moments and product moments of moment-statistics for samples of the finite and infinite populations. *Sankhya The Indian Journal of Statistics*, 6(4):363–382, 1943.

[109] Wilson A. Sutherland. *Introduction to Metric and Topological Spaces 2nd Edition*, pages 39–57. Oxford University Press, 2009.

[110] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[111] Jigang Wang, Predrag Neskovic, and Leon Cooper. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, 28(2):207–213, 2007.

[112] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[113] P. Whittle. Estimation and information in stationary time series. *Arkiv for Matematik*, 2(5):423–434, 1953.

[114] P. Whittle. Curve and periodogram smoothing. *Journal of the Royal Statistical Society Series B*, 19(1):38–63, 1957.

[115] Samuel Wilks. *Mathematical Statistics*, pages 133–136. John Wiley and Sons, 1962.

[116] Samuel Wilks. *Mathematical Statistics*, pages 138–139. John Wiley and Sons, 1962.

[117] A. Wong and M. You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(5):599–609, 1985.

[118] Li Xie, Valery Ugrinovskii, and Ian Petersen. Probabilistic distances between finite-state finite-alphabet hidden Markov models. *IEEE Transactions on Automatic Control*, 50(4):505–511, 2005.

[119] J. M. Yearsley, A. Barque-Duran, E. Scerrati, J. A. Hampton, and E. M. Pothos. The triangle inequality constraint in similarity judgments. *Progress in Biophysics and Molecular Biology*, 130A:26–32, 2017.

[120] Xian-Da Zhang. *Matrix Analysis and Applications*, pages 123–126. Cambridge University Press, 2017.

[121] Walter Zucchini, Iain MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R (Second Edition)*, pages 30–41. CRC Press Taylor and Francis Group, 2016.

[122] Walter Zucchini, Iain MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R (Second Edition)*, pages 56–58. CRC Press Taylor and Francis Group, 2016.

Appendix A

Logarithm of dot product

If $\vec{p} = \{p_1, \ldots, p_c\}$ and $\vec{q} = \{q_1, \ldots, q_c\}$ are two normalized histograms based on sample sizes of n_1 and n_2 respectively, and satisfy

$$\sum_{i=1}^{c} p_i = \sum_{i=1}^{c} q_i = 1 \tag{A.1}$$

then the log dot product distance measure (also known as the Bhattacharya distance [21]) is

$$d_{DP}^2 = 8\left(\frac{n_1 n_2}{n_1 + n_2}\right)\left[-\log \sum_{i=1}^{c} \sqrt{p_i q_i}\right] \tag{A.2}$$

We want to show that this does not satisfy the triangle inequality. We'll position a third normalized histogram $\vec{t} = \{t_1, \ldots, t_c\}$ just off the great circle joining $\vec{p}$ and $\vec{q}$ in square root space. In Figure A.1, the point marked T is the point on that great circle that is closest to $\{\sqrt{t_1}, \ldots, \sqrt{t_c}\}$. At that point, we will construct a coordinate system with three directions: x is along the great circle from $\{\sqrt{p_1}, \ldots, \sqrt{p_c}\}$ to $\{\sqrt{q_1}, \ldots, \sqrt{q_c}\}$ at T, y is along the great circle from T to $\{\sqrt{t_1}, \ldots, \sqrt{t_c}\}$, and z is along the normal from the origin O through T.

Figure A.1 shows the coordinates of $\{\sqrt{p_1}, \ldots, \sqrt{p_c}\}$, $\{\sqrt{q_1}, \ldots, \sqrt{q_c}\}$ and $\{\sqrt{t_1}, \ldots, \sqrt{t_c}\}$ in this coordinate system. Arcs are labelled with the angles they subtend. We want to show that there are solutions of

$$\sqrt{-\log\cos\theta_p} + \sqrt{-\log\cos\theta_q} < \sqrt{-\log\cos 2\theta_0} \tag{A.3}$$

First, assume that the angles involved are small, i.e. $\theta_0, \phi \ll 1$. For small angles θ, we

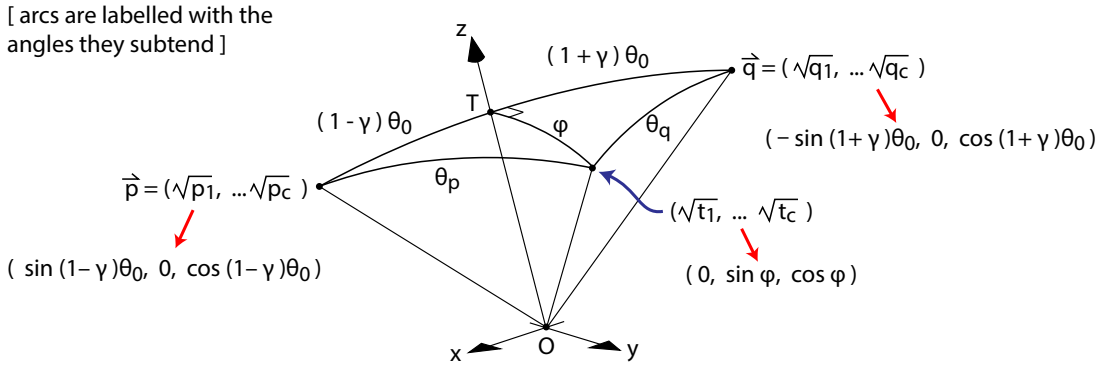


Figure A.1: Angles involved in the log dot product distance measure

have

$$\begin{aligned} -\log(\cos\theta) &\approx \frac{1}{2}\theta^2\left(1+\frac{1}{6}\theta^2\right),\ \sqrt{-\log(\cos\theta)} \approx \frac{1}{\sqrt{2}}\theta\left(1+\frac{1}{12}\theta^2\right), \\ \sqrt{-\log(\cos 2\theta_0)} &\approx \sqrt{2}\theta_0\left(1+\frac{1}{3}\theta_0^2\right) \end{aligned} \tag{A.4}$$

From Figure A.1, we also see that

$$\begin{aligned} -\log(\cos\theta_p) &= -\log(\cos\phi\cos(1-\gamma)\theta_0) \\ &= -\log\cos\phi - \log\cos(1-\gamma)\theta_0 \\ &\approx \frac{1}{2}\phi^2 + \frac{1}{12}\phi^4 + \frac{1}{2}(1-\gamma)^2\theta_0^2 + \frac{1}{12}(1-\gamma)^4\theta_0^4 \end{aligned} \tag{A.5}$$

Similarly,

$$-\log(\cos\theta_q) \approx \frac{1}{2}\phi^2 + \frac{1}{12}\phi^4 + \frac{1}{2}(1+\gamma)^2\theta_0^2 + \frac{1}{12}(1+\gamma)^4\theta_0^4 \tag{A.6}$$

Suppose now that $\phi \approx a\theta_0^2$. Then we seek a solution of

$$\begin{aligned}
&\frac{1}{\sqrt{2}}(1-\gamma)\theta_0\sqrt{1+\frac{2}{(1-\gamma)^2\theta_0^2}\left[\frac{1}{2}a^2\theta_0^4+\frac{1}{12}(1-\gamma)^4\theta_0^4\right]}+ \\
&\frac{1}{\sqrt{2}}(1+\gamma)\theta_0\sqrt{1+\frac{2}{(1+\gamma)^2\theta_0^2}\left[\frac{1}{2}a^2\theta_0^4+\frac{1}{12}(1+\gamma)^4\theta_0^4\right]}< \\
&\qquad\sqrt{2}\theta_0+\frac{\sqrt{2}}{3}\theta_0^3
\end{aligned} \tag{A.7}$$

From that equation, we get

$$\begin{aligned}
&\frac{a^2}{2\sqrt{2}}\left(\frac{1}{1-\gamma}+\frac{1}{1+\gamma}\right)+\frac{1}{12\sqrt{2}}\left[(1-\gamma)^3+(1+\gamma)^3\right]<\frac{\sqrt{2}}{3} \\
&\qquad\Rightarrow |a|<\frac{1}{\sqrt{2}}(1-\gamma^2),\quad |\phi|<\frac{1}{\sqrt{2}}(1-\gamma^2)\theta_0^2
\end{aligned} \tag{A.8}$$

So, between $\{\sqrt{p_1},\ldots,\sqrt{p_c}\}$ and $\{\sqrt{q_1},\ldots,\sqrt{q_c}\}$, there is a narrow band along the great circle between them where the triangle inequality does not hold. Figure A.2 illustrates this situation.

For example, if $\theta_0 = 0.2$, then the total band allowed at $\gamma = 0$ is approximately 0.06. Take $\phi = 0.02$ at $\gamma = 0$. Then

$$\begin{aligned}
\cos\theta_p = \cos\theta_q &= \cos\phi\cos\theta_0 = 0.9799 \\
\cos 2\theta_0 &= 0.9211 \\
2\sqrt{-\log(0.9799)} &< \sqrt{-\log(0.9211)} \text{ is this true?} \\
0.2850 &< 0.2867 \text{ verified, yes it's true!}
\end{aligned} \tag{A.9}$$

which shows that indeed, there are sets of three histograms which do not satisfy the triangle inequality for the log dot product distance measure.

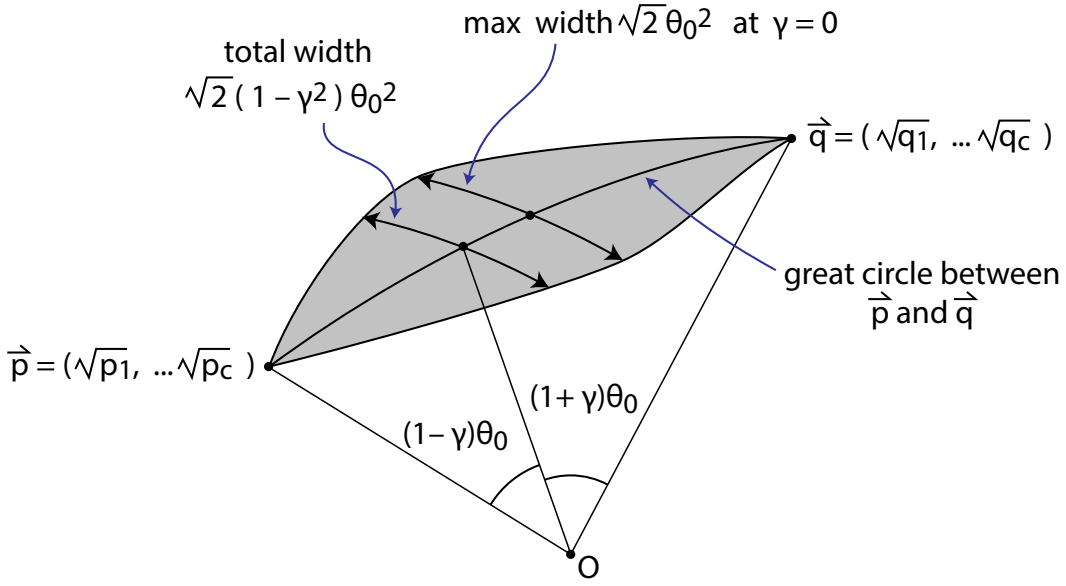


Figure A.2: Region in which the log dot product does not satisfy the triangle inequality. It is a band along the great circle joining $\vec{p}$ and $\vec{q}$, of varying width.

Appendix B

Covariance of sample moments

Here, our aim is to derive the statistics of sample moments, given the sample size and the parent population characteristics. Let a sample of size n be $\{x_1, \ldots, x_n\}$, and let the parent population be $\{X_1, \ldots, X_N\}$ where $N > n$. We'll make the definitions

$$\begin{aligned} s_r &= \sum_{i=1}^{n} x_i^r, \quad S_r = \sum_{i=1}^{N} X_i^r \quad (r \geq 1) \\ m_{r_1 \ldots r_p} &= \sum_{i_1=1}^{n} \ldots \sum_{i_p=1}^{n} x_{i_1}^{r_1} \ldots x_{i_p}^{r_p}, \quad M_{r_1 \ldots r_p} = \sum_{i_1=1}^{N} \ldots \sum_{i_p=1}^{N} X_{i_1}^{r_1} \ldots X_{i_p}^{r_p} \end{aligned} \tag{B.1}$$

where $i_1 \neq i_2 \ldots \neq i_p$ and $r_1 \geq r_2 \ldots \geq r_p$. That is, the partitions $\{r_1, r_2, \ldots r_p\}$ are in descending order. Note that as defined here, the symmetric products $m_{r_1 \ldots r_p}$ allow repeated terms. For example, if a sample is $s = \{x_1, x_2, x_3\}$, then

$$\begin{aligned} m_{11} &= x_1x_2 + x_1x_3 + x_2x_1 + x_2x_3 + x_3x_1 + x_3x_2 \\ &= 2(x_1x_2 + x_1x_3 + x_2x_3) \end{aligned} \tag{B.2}$$

We would like a way of relating products of the form $s_{r_1} \ldots s_{r_p}$ to the corresponding symmetric products $m_{r_1 \ldots r_p}$. The fundamental rules for that are:

$$\begin{aligned} m_r &= s_r \\ m_{r_1 \ldots r_p} s_r &= m_{r_1+r, \ldots r_p} + m_{r_1, r_2+r, \ldots r_p} + \ldots + m_{r_1, \ldots r_p+r} + m_{r_1, \ldots r_p, r} \end{aligned} \tag{B.3}$$

For example,

$$\begin{aligned}[\sum_{i\neq j} x_i x_j][\sum x_k] &= [\sum_{i\neq j} x_i x_j](x_i + x_j + \sum_{k\neq i,j} x_k) \\ &= \sum_{i\neq j} x_i^2 x_j + \sum_{i\neq j} x_i x_j^2 + \sum_{i\neq j\neq k} x_i x_j x_k \\ \Rightarrow m_{11}s_1 &= 2m_{21} + m_{111}\end{aligned} \tag{B.4}$$

For each integer n, we may form the partitions $\{r_1 \dots r_p\}$ of n, and use the preceding rule to find the relations between the $s_{r_1} \dots s_{r_p}$ terms and the $m_{r_1 \dots r_p}$ symmetric products. For example, for $n = 3$ we have

$$\begin{aligned}&s_3 = m_3, \quad s_2 s_1 = m_2 s_1 = m_3 + m_2 m_1 \\ &s_1^3 = (m_1 s_1)s_1 = (m_2 + m_{11})s_1 = m_3 + 3m_{21} + m_{111}\end{aligned} \tag{B.5}$$

and for $n = 4$ we have

$$\begin{bmatrix} s_4 \\ s_3 s_1 \\ s_2^2 \\ s_2 s_1^2 \\ s_1^4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 \\ 1 & 4 & 3 & 6 & 1 \end{bmatrix} \begin{bmatrix} m_4 \\ m_{31} \\ m_{22} \\ m_{211} \\ m_{1111} \end{bmatrix} \tag{B.6}$$

Of course, those relations may be inverted to give the symmetric products in terms of the $s_{r_1} \dots s_{r_p}$:

$$\begin{aligned}\begin{bmatrix} m_3 \\ m_{21} \\ m_{111} \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -3 & 1 \end{bmatrix} \begin{bmatrix} s_3 \\ s_2 s_1 \\ s_1^3 \end{bmatrix} \\ \begin{bmatrix} m_4 \\ m_{31} \\ m_{22} \\ m_{211} \\ m_{1111} \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 2 & -2 & -1 & 1 & 0 \\ -6 & 8 & 3 & -6 & 1 \end{bmatrix} \begin{bmatrix} s_4 \\ s_3 s_1 \\ s_2^2 \\ s_2 s_1^2 \\ s_1^4 \end{bmatrix}\end{aligned} \tag{B.7}$$

The heart of these calculations is the question: if I have a term $m_{r_1 \dots r_p}$ relating to a sample, what is $\sum m_{r_1 \dots r_p}$ over all possible samples from the parent population? Take $m_1 = \sum_s x_i$, for example, where $\sum_s$ means the sum over one possible sample. In the sum

$\sum_{\text{all s}} \sum_s x_i$, if we choose one of the x_i to be a specific X_j, then we can make the rest of the sample in $\binom{N-1}{n-1}$ ways, hence

$$\sum_{\text{all samples}} m_1 = \binom{N-1}{n-1} M_1 \tag{B.8}$$

Similarly, in the sum $\sum_{\text{all samples}} \sum_{i \neq j} x_i x_j$, if we choose a specific (X_i, X_j) pair, then we may choose the rest of the sample in $\binom{N-2}{n-2}$ ways, so that

$$\sum_{\text{all samples}} m_{11} = \binom{N-2}{n-2} M_{11} \tag{B.9}$$

Note that this argument only depends on the number of indices of the symmetric product, so in general we have

$$\sum_{\text{all samples}} m_{r_1 \dots r_p} = \binom{N-p}{n-p} M_{r_1 \dots r_p} \tag{B.10}$$

and when we take the average over all possible $\binom{N}{n}$ samples, we get

$$\langle m_{r_1 \dots r_p} \rangle = \frac{n(n-1)\dots(n-p+1)}{N(N-1)\dots(N-p+1)} M_{r_1 \dots r_p} \tag{B.11}$$

Let's give some examples of how to use these equations. For a sample of size n, the sample third moment is

$$\begin{aligned} \hat{\mu}_3 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \frac{s_1}{n})^3 = \frac{1}{n} s_3 - \frac{3}{n^2} s_2 s_1 + \frac{2}{n^3} s_1^3 \\ &= \frac{1}{n} m_3 - \frac{3}{n^2}(m_3 + m_{21}) + \frac{2}{n^3}(m_3 + 3m_{21} + m_{111}) \end{aligned} \tag{B.12}$$

and averaging over all possible samples gives

$$\begin{aligned}
\langle\hat{\mu}_3\rangle &= \left(\frac{1}{n} - \frac{3}{n^2} + \frac{2}{n^3}\right)\frac{n}{N}M_3 + \left(-\frac{3}{n^2} + \frac{6}{n^3}\right)\frac{n(n-1)}{N(N-1)}M_{21} + \frac{2}{n^3}\frac{n(n-1)(n-2)}{N(N-1)(N-2)}M_{111} \\
&= \frac{(n-1)(n-2)}{n^3}\frac{n}{N}S_3 - 3\frac{n-2}{n^3}\frac{n(n-1)}{N(N-1)}(-S_3 + S_2S_1) \\
&\quad + \frac{2}{n^3}\frac{n(n-1)(n-2)}{N(N-1)(N-2)}(2S_3 - 3S_2S_1 + S_1^3) \\
&= \frac{(n-1)(n-2)}{n^2}\frac{N^2}{(N-1)(N-2)}\left[\frac{S_3}{N} - 3\frac{S_2S_1}{N^2} + 2\left(\frac{S_1}{N}\right)^3\right] \\
&= \frac{(n-1)(n-2)}{n^2}\frac{N^2}{(N-1)(N-2)}\mu_3
\end{aligned} \tag{B.13}$$

where μ_3 is the third central moment of the parent population. If $N \gg n$, then we get

$$\langle\hat{\mu}_3\rangle = \frac{(n-1)(n-2)}{n^2}\mu_3 \tag{B.14}$$

and there is a bias toward zero of magnitude $(3/n)\mu_3$ for $n \gg 1$. Similarly, the ensemble average of the sample variance is

$$\langle\hat{\mu}_2\rangle = \langle\frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{s_1}{n})^2\rangle = \frac{n-1}{n}\mu_2 \tag{B.15}$$

where μ_2 is the population variance and $N \gg n$. The ensemble variance of the sample variance is then

$$\begin{aligned}
\mathrm{var}(\hat{\mu}_2) &= \langle\hat{\mu}_2^2\rangle - \langle\hat{\mu}_2\rangle^2 = \langle\left[\frac{s_2}{n} - \left(\frac{s_1}{n}\right)^2\right]^2\rangle - \langle\hat{\mu}_2\rangle^2 \\
&= \langle\frac{1}{n^2}s_2^2 - \frac{2}{n^3}s_2s_1^2 + \frac{1}{n^4}s_1^4\rangle - \langle\hat{\mu}_2\rangle^2
\end{aligned} \tag{B.16}$$

Making the assumption that $N \gg n$, and using the equations developed so far, we get

$$\mathrm{var}(\hat{\mu}_2) = \frac{(n-1)^2}{n^3}(\mu_4 - \frac{n-3}{n-1}\mu_2^2) \approx \frac{1}{n}(\mu_4 - \mu_2^2) \quad \text{if } n \gg 1 \tag{B.17}$$

where μ_4 is the fourth central moment of the parent population. For more information on the statistics of the sample moments, see [108] and [106].

Appendix C

Sampling the Cramer von Mises distribution for unequal classes

The analysis of empirical distribution functions, and of distance measures based on them, rests upon the theory of the Brownian bridge. In particular, if we transform the domain of a sample's empirical CDF by the corresponding theoretical CDF, then the result has the same statistics as an ordered sample of the uniform $U(0,1)$ distribution [3]. Also, for large sample size n, the resulting empirical CDF has the statistics of a Brownian random walk between (0,0) and (1,1). The Brownian bridge is the difference between a Brownian walk from (0,0) to (1,c), and the straight line joining (0,0) and (1,c) [42]. We can generate a Brownian bridge from $N(0,1)$ variates as follows:

$$b(t) = \lim_{T\to\infty}\left[\sum_{j=0}^{tT}\frac{z_j}{\sqrt{T}} - t\sum_{j=0}^{T}\frac{z_j}{\sqrt{T}}\right], \quad z_j \sim N(0,1), t \in [0,1] \tag{C.1}$$

From that definition, we get

$$\mathrm{cov}(b(t), b(s)) = \langle b(t)b(s)\rangle = \min(s,t) - st, \quad \mathrm{var}(b(t)) = t(1-t) \tag{C.2}$$

Furthermore, we also get

$$\langle\int_0^1 b^2(t)\ \mathrm{d}t\rangle = \langle\lim_{T\to\infty}\frac{1}{T}\sum_{j=0}^{T} b^2(\frac{j}{T})\rangle = \lim_{T\to\infty}\frac{1}{T}\left[\sum_{j=0}^{T}\left(\frac{j}{T} - \frac{j^2}{T^2}\right)\right] = \lim_{T\to\infty}\frac{1}{6}\left(1 - \frac{1}{T^2}\right) = \frac{1}{6} \tag{C.3}$$

Similar calculations [3] show that

$$\mathrm{var}(\int_0^1 b^2(t)\ \mathrm{d}t) = \frac{1}{45} \tag{C.4}$$

The relationship between $b(t)$ and a sample's empirical distribution is that the sequence $y_j = \{\sqrt{n}[F_n(x_j) - F(x_j)]\}$, where $F_n(x)$ is the empirical CDF and $F(x)$ is the theoretical CDF, becomes statistically equivalent to $b(t)$ where $t_j = F(x_j)$ and $n \gg 1$.

Now, we will examine the two-sample case, in which

$$y_j = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[\hat{F}_{1,n_1}(x_j) - \hat{F}_{2,n_2}(x_j)\right], \tag{C.5}$$

the $\{x_j\}$ are from the joint sample, and the subscripts remind us that $\hat{F}_1(x)$ and $\hat{F}_2(x)$ are based on samples of sizes n_1 and n_2 respectively (i.e. not a joint sample). If there is an actual difference between the true cumulatives $F_1(x)$ and $F_2(x)$, then we can set

$$\begin{aligned} y_j &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[[\hat{F}_1(x_j) - F_1(x_j) - \hat{F}_2(x_j) + F_2(x_j)] + [F_1(x_j) - F_2(x_j)]\right] \\ &= b(x_j) + u(x_j) \quad \text{where } u(x_j) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} [F_1(x_j) - F_2(x_j)] \end{aligned} \tag{C.6}$$

Now $b(x_j)$ is asymptotically a Brownian bridge as $n_1, n_2 \to \infty$, but $u(x_j)$ is not random at all. Let the Cramer von Mises statistic be

$$d^2_{CM} = \lim_{T\to\infty} \frac{1}{T} \sum [b(t/T) + u(t/T)]^2 = \int_0^1 [b(t) + u(t)]^2\ \mathrm{d}t \tag{C.7}$$

We have right away that

$$\begin{aligned} \langle \lim_{T\to\infty} \frac{1}{T} \sum [b(t/T) + u(t/T)]^2 \rangle &= \langle \int_0^1 b^2(t)\ \mathrm{d}t \rangle + \int_0^1 u^2(t)\ \mathrm{d}t \\ &= \frac{1}{6} + \int_0^1 u^2(t)\ \mathrm{d}t = \frac{1}{6} + k \end{aligned} \tag{C.8}$$

where k is the total energy of $u(t)$:

$$k = \frac{n_1 n_2}{n_1 + n_2} \int_0^1 [F_1(x) - F_2(x)]^2\ \mathrm{d}F(x) \tag{C.9}$$

As for the variance of d^2_{CM}, observe that with $b_t = b(t/T)$ and $u_t = u(t/T)$, the only non-zero terms of

$$\langle d^4_{CM} \rangle = \lim_{T\to\infty} \frac{1}{T^2} \langle \sum_{t=0}^{T} \sum_{s=0}^{T} (b_t + u_t)^2 (b_s + u_s)^2 \rangle \tag{C.10}$$

that are not contained in

$$\langle d^2_{CM} \rangle^2 = \lim_{T\to\infty} \frac{1}{T^2} \langle \sum_{t=0}^{T} (b_t + u_t)^2 \rangle \langle \sum_{s=0}^{T} (b_s + u_s)^2 \rangle \tag{C.11}$$

are

$$\begin{aligned} \mathrm{var}(d^2_{CM}) &= \lim_{T\to\infty} \left[\mathrm{var}(\frac{1}{T} \sum_{t=0}^{T} b_t^2) + \frac{4}{T^2} \sum_{t=0}^{T} \sum_{s=0}^{T} \langle b_s, b_t \rangle u_s u_t \right] \\ &= \frac{1}{45} + \lim_{T\to\infty} \frac{4}{T^2} \sum_{t=0}^{T} \sum_{s=0}^{T} \langle b_s, b_t \rangle u_s u_t \end{aligned} \tag{C.12}$$

However, we know that

$$\langle b(s), b(t) \rangle = \min(s,t) - st \tag{C.13}$$

is a Mercer kernel with expansion

$$\begin{aligned} \min(s,t) - st = \sum_{j=1}^{\infty} \lambda_j f_j(s) f_j(t) \quad &\text{where} \\ \lambda_j = \frac{1}{\pi^2 j^2} \quad \text{and} \quad f_j(t) &= \sqrt(2) \sin(\pi j t) \end{aligned} \tag{C.14}$$

For further details, see [43]. Thus, we also have

$$\begin{aligned} \mathrm{var}(d^2_{CM}) &= \frac{1}{45} + \lim_{T\to\infty} \frac{4}{T^2} \sum_{t=0}^{T} \sum_{s=0}^{T} \sum_{j=1}^{\infty} \lambda_j f_j(s/T) f_j(t/T) u_s u_t \\ &= \frac{1}{45} + 4 \sum_{j=1}^{\infty} \lambda_j \left[\int_0^1 f_j(t) u(t) \, \mathrm{d}t \right]^2 = \frac{1}{45} + 4 \sum_{j=1}^{\infty} \lambda_j a_j^2 \end{aligned} \tag{C.15}$$

where the $\{a_j\}$ form the Fourier sine transform of $u(t)$ over $[0,1]$. In particular,

$$\begin{aligned} a_j &= \int_0^1 u(t)(\sqrt{2}\sin(j\pi t))\ \mathrm{d}t \quad \text{and} \\ k &= \int_0^1 u^2(t)\ \mathrm{d}t = \sum_{j=1}^{\infty} a_j^2 \quad \text{by Parseval's theorem} \\ \Rightarrow \mathrm{var}(d_{CM}^2) &= \frac{1}{45} + 4\left(\frac{\sum \lambda_j a_j^2}{\sum a_j^2}\right) k \end{aligned} \tag{C.16}$$

As for the actual distribution of d_{CM}^2 when the samples are from different distributions, we need the principal components of

$$W_n^2 = \frac{1}{n}\sum_{j=1}^{n} y_j^2 \quad \text{where } n = n_1 + n_2 \tag{C.17}$$

The covariance $\langle b(s), b(t)\rangle = \min(s,t) - st$, when rotated to its principal axes, gives

$$\begin{aligned} W_n^2 &= \sum_{j=1}^{\infty} \lambda_j z_{nj}^2 \quad \text{where} \quad \lambda_j = \frac{1}{j^2\pi^2} \quad \text{and} \\ z_{nj} &= (j\pi)\sqrt{\frac{n_1 n_2}{n_1+n_2}} \int_0^1 [\hat{F}_1(t) - \hat{F}_2(t)](\sqrt{2}\sin j\pi t)\ \mathrm{d}t \end{aligned} \tag{C.18}$$

See Durbin and Knott [43] for a fuller development. The z_{nj} are $IID(a_j/\sqrt{\lambda_j}, 1)$ and for large sample sizes n_1, n_2 approach $N(a_j/\sqrt{\lambda_j}, 1)$. Separating $\hat{F}_1(x) - \hat{F}_2(x)$ into its random and fixed parts yields

$$d_{CM}^2 = \sum_{j=1}^{\infty} \lambda_j (x_j + \frac{a_j}{\sqrt{\lambda_j}})^2 \tag{C.19}$$

where the $\{x_j\}$ are asymptotically $N(0,1)$, and

$$a_j = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \int_0^1 [F_1(t) - F_2(t)](\sqrt{2}\sin j\pi t)\ \mathrm{d}t \tag{C.20}$$

Thus, d_{CM}^2 is a weighted combination of non-central chi-square variables. We may sample this distribution by calculating the Fourier coefficients $\{a_j\}$ for a specific cumulative distribution pair $\{F_1, F_2\}$, and by choosing different sample sets $\{x_1 \ldots x_p\} \sim N(0, I_p)$. In Figure C.1, I have done this for the same H_0 and H_1 distribution pairs that are illustrated

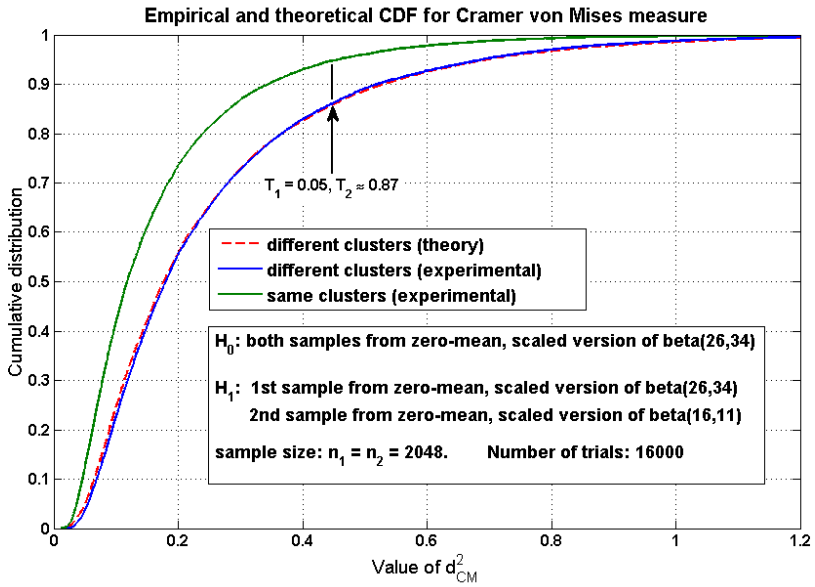


Figure C.1: Prediction of type II error for Cramer von Mises measure on two neighboring distribution classes

in Figure 3.13. The two empirical PDFs for the case of unequal, neighboring clusters (one generated by calculating the Cramer von Mises measure for sample pairs, and one generated via equation C.19 with 40 terms) have a Kolmogorov-Smirnov statistic of 1.04, well within the 95% confidence limit of 1.36. So this theory looks very solid.

I will give an abbreviated summary of how to sample the Anderson Darling distance measure. The interested reader will find more in [43] and [4]. The Anderson Darling measure on empirical CDFs from two samples is

$$d^2_{AD} = \frac{n_1 n_2}{n_1 + n_2} \int_0^1 \frac{[F_{n_1}(x) - G_{n_2}(x)]^2}{H_{n_1+n_2}(x)[1 - H_{n_1+n_2}(x)]} \, \mathrm{d}H_{n_1+n_2}(x) \tag{C.21}$$

where $F_{n_1}(x)$ and $G_{n_2}(x)$ are the empirical CDFs of the two samples, and $H_{n_1+n_2}(x)$ is the empirical CDF of the joint sample. Supposing $F(x) - G(x)$ to be the actual difference in CDFs for the two samples, the total Anderson Darling energy is

$$\begin{aligned} k &= \frac{n_1 n_2}{n_1 + n_2} \int_0^1 \frac{[F(x) - G(x)]^2}{H(x)[1 - H(x)]} \, \mathrm{d}H(x) \quad \text{where} \\ H(x) &= \frac{n_1}{n_1 + n_2} F(x) + \frac{n_2}{n_1 + n_2} G(x) \end{aligned} \tag{C.22}$$

The corresponding Mercer kernel expansion is

$$\begin{aligned} \frac{\min(s,t)-st}{\sqrt{s(1-s)t(1-t)}} &= \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t) \\ \text{where } \phi_j(t) &= \sqrt{\frac{2j+1}{j(j+1)}} P_j^1(2t-1) \quad t \in [0,1] \text{ and } \lambda_j = \frac{1}{j(j+1)} \end{aligned} \tag{C.23}$$

and $P_j^1(x)$ is the first associated Legendre function

$$P_j^m(x) = \frac{(-1)^m}{2^j j!}(1-x^2)^{m/2}\left(\frac{d}{dx}\right)^{j+m}(x^2-1)^j \tag{C.24}$$

In a manner similar to the one developed above for the Cramer von Mises measure, the Anderson Darling statistic may be sampled as

$$\begin{aligned} d_{AD}^2 &\approx \sum_{j=1}^{p} \lambda_j \left(x_j + \frac{a_j}{\sqrt{\lambda_j}}\right)^2 \quad x_j \sim N(0,1), \quad p > 20 \\ a_j &= \int_0^1 \frac{F(x)-G(x)}{\sqrt{H(x)[1-H(x)]}} \sqrt{\frac{2j+1}{j(j+1)}} P_j^1(2H(x)-1)\, \mathrm{d}H(x) \end{aligned} \tag{C.25}$$

Figure C.2 shows empirical and theoretical distributions for d_{AD}^2, for the same H_0 and H_1 distribution pairs illustrated in Figure 3.13. The corresponding Kolmogorov-Smirnov statistic is 1.21, well within the 95% confidence level of 1.36.

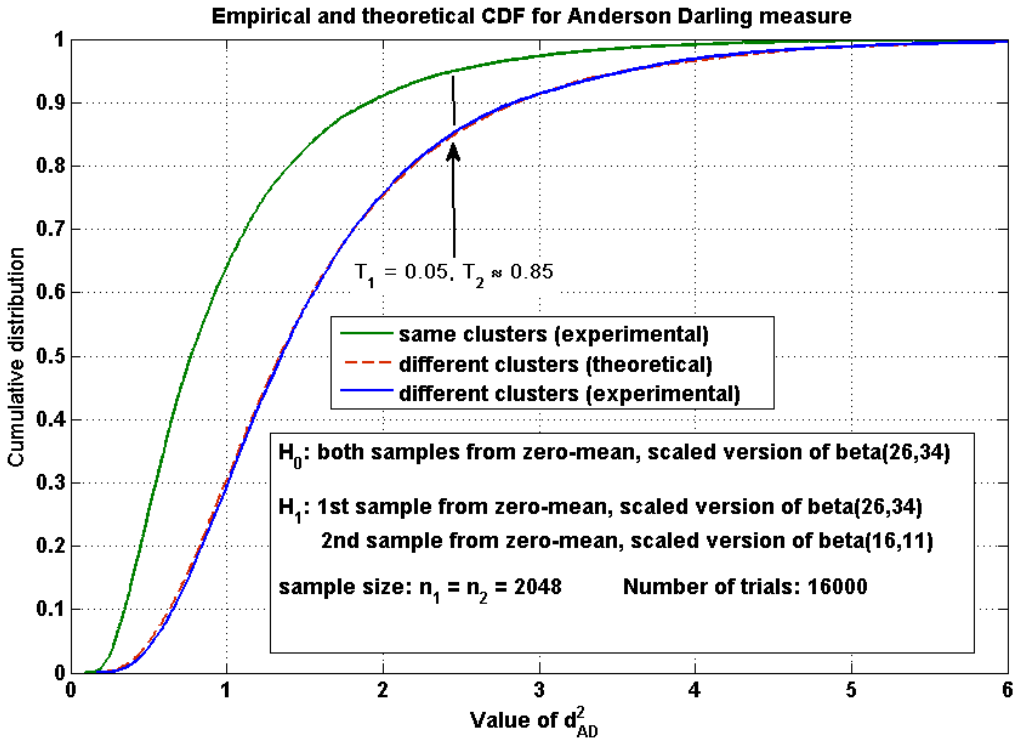


Figure C.2: Prediction of type II error for Anderson Darling measure on two neighboring distribution classes

Appendix D

Curvature of total sample surprisal for an ARMA process - from innovations

The purpose of this appendix is to derive the expected value of the Hessian second derivative of the total sample surprisal (see equation 2.3), where the Hessian is with respect to the roots of a stationary, invertible ARMA time series process (see section 2.4). This derivation will look at the total surprisal of the series innovations, here assumed to be normally distributed. In Appendix E, we will relax that assumption.

Our first step will be to establish preliminary results for upper and lower triangular Toeplitz matrices. A Toeplitz matrix is one in which cells on a descending diagonal have equal value [120]. Let's define

$$\mathrm{ltt}(\vec{a}, n) = \begin{bmatrix} a_1 & 0 & & \\ a_2 & a_1 & 0 & \\ & \ddots & \ddots & 0 \\ \dots & a_3 & a_2 & a_1 \end{bmatrix} \quad (n \text{ rows }) \tag{D.1}$$

where $\vec{a} = [a_1, a_2, \dots a_n]$. Furthermore, set $\mathrm{ltt}(\vec{a}) = \mathrm{ltt}(\vec{a}, n)$ where $n \gg 1$, meaning: n is large enough for end effects to be negligible, but n is still finite. Then we also have

$$\begin{aligned} \mathrm{ltt}(\vec{a})\mathrm{ltt}(\vec{b}) = \mathrm{ltt}(\vec{c}) \quad \text{where } c_p = \sum_{j=1}^{p} b_j a_{p+1-j}, \text{ and} \\ \mathrm{ltt}^{-1}([1 \quad -r]) = \mathrm{ltt}([1 \quad r \quad r^2 \dots]) \quad \text{where } |r| < 1 \end{aligned} \tag{D.2}$$

Those last two results show that lower triangular Toeplitz matrices are closed under addition, multiplication, and inversion, and commute with each other. Clearly, the same goes for upper triangular Toeplitz matrices. However, lower and upper triangular Toeplitz matrices do not necessarily commute. Finally, we have

$$\begin{aligned} \mathrm{tr}\left(\mathrm{ltt}(\vec{a})\mathrm{ltt}^T(\vec{b})\right) &= na_1b_1 + (n-1)a_2b_2 + (n-2)a_3b_3 + \dots \\ &\approx n(\vec{a} \bullet \vec{b}) \quad \text{if } a_j \text{ and } b_j \text{ decrease exponentially} \end{aligned} \tag{D.3}$$

Next, we'll look at expressions of the form

$$\langle \vec{e}(B^T)^j \mathrm{ltt}^T(\vec{a})\mathrm{ltt}(\vec{b})B^k\vec{e}\rangle \tag{D.4}$$

where $\vec{e} \sim IID(0, vI_n)$, and B is the backward-shift operator in matrix form, i.e.

$$B = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ & \ddots & \ddots & \end{bmatrix}, \quad B[x_1 \quad x_2 \dots x_n]^T = [0 \quad x_1 \quad x_2 \dots x_{n-1}]^T \tag{D.5}$$

Note that $B\vec{e}$ and $\vec{e}$ have the same statistics to order $O(1/n)$. Also, we will define

$$\mathrm{tr}(A,k) = \sum_{j=1}^{n-k} A_{j,j+k} \quad (A \text{ is } n \text{ by } n), \quad \mathrm{tr}(A) = \mathrm{tr}(A,0) \tag{D.6}$$

Thus, $\mathrm{tr}(A,+k)$ is the sum of elements along the k-th super-diagonal of A, while $\mathrm{tr}(A,-k)$ is the sum of elements along the k-th subdiagonal of A. From the independence of the elements of $\vec{e}$, we get

$$\begin{aligned} \langle \vec{e}(B^T)^k A B^l \vec{e}\rangle &= \langle (B^k\vec{e})^T A (B^l\vec{e})\,\rangle \\ &= \langle \sum_{i,j=1}^{n} A_{ij}[B^k\vec{e}]_i[B^l\vec{e}]_j\rangle = \sum_{i,j=1}^{n} A_{ij}\langle [\vec{e}]_{i-k}[\vec{e}]_{j-l}\rangle \\ &= v\sum_{i,j=1}^{n} A_{ij}\delta_{i-k,j-l} = v\sum A_{i,i-k+l} = v\,\mathrm{tr}(A, l-k) \end{aligned} \tag{D.7}$$

An ARMA(p,q) zero-mean process (see section 2.4) is a time series model with defining equation

$$x_j - a_1x_{j-1} \dots - a_px_{j-p} = e_j - b_1e_{j-1} \dots - b_qe_{j-q} \tag{D.8}$$

where $\vec{x}, \vec{e}, \vec{a}$, and $\vec{b}$ are real, e_i is independent of x_i, and $e_i \sim IID(0, v)$. Here I am neglecting end effects of order $O(1/n)$. If we were to take these into account, the first $\max(p, q)$ values of $\vec{x}$ would have different defining equations. The index j in the defining equation is a time marker for equal time increments. The ARMA model itself is the triple $\{\vec{a}, \vec{b}, v\}$, and the usual task is to estimate the model given an observed sample $\{x_1, x_2, \ldots x_n\}$, with the goal of predicting future values $\{x_{n+1}, x_{n+2}, \ldots\}$ with the minimum possible variance. I only consider the case where $n \gg 1$.

Let B represent the backshift operator again, this time meaning the operation of going back one time period (i.e. $Bx_j = x_{j-1}, Be_j = e_{j-1}$). Then

$$(1 - a_1 B - a_2 B^2 \ldots - a_p B^p) x_j = (1 - b_1 B - b_2 B^2 \ldots - b_q B^q) e_j \tag{D.9}$$

If the process is stationary and invertible, then equation D.9 can be put in the form

$$(1 - r_p B) \ldots (1 - r_1 B) x_j = (1 - s_q B) \ldots (1 - s_1 B) e_j \tag{D.10}$$

where the AR roots $\{r_j\}$ and MA roots $\{s_j\}$ are less than one in magnitude, and are either real or occur in conjugate pairs. Now let $\vec{x}$ and $\vec{e}$ be column vectors representing the observations and innovations respectively. The matrix equivalent of equation D.10 is

$$L(r_p) \ldots L(r_1) \vec{x} = L(s_q) \ldots L(s_1) \vec{e} \tag{D.11}$$

where $L(r) = \mathrm{ltt}([1 \quad -r])$ and we are neglecting end effects of order $1/n$ where n is the sample size. Assuming for the moment that the innovations are normally distributed, i.e. $e_j \sim N(0, v)$, the probability of the sample is

$$p(\vec{e}) = \left[\frac{1}{\sqrt{2\pi v}}\right]^n \exp -\frac{1}{2v}(\vec{e} \bullet \vec{e}) \tag{D.12}$$

and the total sample surprisal is

$$\mathcal{L} = -\log p(\vec{e}) = \frac{n}{2} \log 2\pi v + \frac{1}{2v}(\vec{e} \bullet \vec{e}) \tag{D.13}$$

However, under the assumption of invertibility,

$$\vec{e} = L^{-1}(s_1) \ldots L^{-1}(s_q) L(r_p) \ldots L(r_1) \vec{x} \tag{D.14}$$

Note that all these matrix multipliers are lower triangular Toeplitz, so they all commute. We want the curvature of $Q = \vec{e} \bullet \vec{e}$ with respect to $\{\vec{r}, \vec{s}\}$. There are five cases of $\partial^2 Q / \partial\theta_j \partial\theta_k$: the two parameters θ_j and θ_k represent the same AR root, different AR

roots, one AR root and one MA root, different MA roots, or the same MA root. Here, I'll work through two cases and state results for the other three cases.

The first case is when the parameters θ_j and θ_k represent different AR roots. Let $L_i = L(r_i), L_j = L(r_j)$, and let O be the product $L^{-1}(s_1)\dots L^{-1}(s_q)L(r_p)\dots L(r_1)$ with $L(r_i)$ and $L(r_j)$ removed. Then

$$\vec{e} = OL_jL_i\vec{x}, \quad Q = \vec{e}^T\vec{e} = \vec{x}^T L_i^T L_j^T O^T O L_j L_i \vec{x} \tag{D.15}$$

Note that

$$\frac{\partial L(r_i)}{\partial r_i} = -B, \quad \frac{\partial^2 L(r_i)}{\partial r_i^2} = 0 \tag{D.16}$$

where B is the backshift operator in matrix form. Using those equations, we get

$$\begin{aligned}\frac{\partial^2 Q}{\partial r_i \partial r_j} &= \vec{x}^T O^T (B^2)^T O L_j L_i \vec{x} + \vec{x}^T L_j^T O^T B^T B O L_i \vec{x} + \\ &\quad \vec{x}^T L_i^T O^T B^T B O L_j \vec{x} + \vec{x}^T L_i^T L_j^T O^T B^2 O \vec{x} \\ &= \vec{e}^T (B^2)^T L_j^{-T} L_i^{-T} \vec{e} + \vec{e}^T B^T L_i^{-T} L_j^{-1} B\vec{e} + \vec{e}^T B^T L_j^{-T} L_i^{-1} B\vec{e} + \\ &\quad \vec{e}^T L_i^{-1} L_j^{-1} B^2 \vec{e}\end{aligned} \tag{D.17}$$

and taking the expectation over $\vec{e}$ gives

$$\langle \frac{\partial^2 Q}{\partial r_i \partial r_j} \rangle = v\mathrm{tr}(L_j^{-T} L_i^{-T}, -2) + v\mathrm{tr}(L_i^{-T} L_j^{-1}) + v\mathrm{tr}(L_j^{-T} L_i^{-1}) + v\mathrm{tr}(L_i^{-1} L_j^{-1}, 2) \tag{D.18}$$

The first term is the sum over a subdiagonal of an upper triangular matrix, so that's zero. A similar argument applies to the last term. Using equation D.3, the remaining two terms give

$$\begin{aligned}&\langle \frac{\partial^2 Q}{\partial r_i \partial r_j} \rangle \approx 2nv(1 + r_i r_j + r_i^2 r_j^2 + \dots) = 2nv\left(\frac{1}{1 - r_i r_j}\right) \quad \text{to order } 1/n \\ \Rightarrow &\langle \frac{\partial^2 \mathcal{L}}{\partial r_i \partial r_j} \rangle = n\left(\frac{1}{1 - r_i r_j}\right) \quad \text{to order } 1/n\end{aligned} \tag{D.19}$$

where "to order $1/n$" means that the calculated value and the true value differ by at most $O(1/n)$.

The second case is when the parameters θ_j and θ_k represent an AR root and an MA root. Once again, let $L_i = L(r_i)$. Since $L_i L_i^{-1} = I$, then

$$\frac{\partial}{\partial r_i}(L_i L_i^{-1}) = \frac{\partial L_i}{\partial r_i} L_i^{-1} + L_i \frac{\partial L_i^{-1}}{\partial r_i} = 0 \quad \Rightarrow \frac{\partial L_i^{-1}}{\partial r_i} = -L_i^{-1} \frac{\partial L_i}{\partial r_i} L_i^{-1} \tag{D.20}$$

As for the previous case, let $L_a = L(r_a)$, $L_c = L(s_c)$, and let O be the product $L^{-1}(s_1)\ldots L^{-1}(s_q)\bullet$ $L(r_p)\ldots L(r_1)$ with $L(r_a)$ and $L^{-1}(s_c)$ removed. Then

$$\vec{e} = OL_c^{-1}L_a\vec{x}, Q = \vec{e}^T\vec{e} = \vec{x}^T L_a^T L_c^{-T} O^T O L_c^{-1} L_a \vec{x} \tag{D.21}$$

Using the same arguments as in the first case, we get

$$\begin{aligned} \frac{\partial^2 Q}{\partial r_a \partial s_c} &= -\vec{x}^T L_c^{-T} O^T B^T B O L_c^{-1} L_c^{-1} L_a \vec{x} - \vec{x}^T L_a^T L_c^{-T} L_c^{-T} O^T B^T B O L_c^{-1} \vec{x} \\ &= \vec{e}^T B^T L_a^{-T} L_c^{-1} B\vec{e} + \vec{e}^T B^T L_c^{-T} L_a^{-1} B\vec{e} \end{aligned} \tag{D.22}$$

Finally, using equation D.3, we get

$$\begin{aligned} \langle \frac{\partial^2 Q}{\partial r_a \partial s_c} \rangle &= -nv[\mathrm{tr}(L_a^{-T} L_c^{-1}) + \mathrm{tr}(L_a^{-1} L_c^{-T})] \\ &\approx -2nv(1 + r_a s_c + r_a^2 s_c^2 + \ldots) = -2nv\left(\frac{1}{1 - r_a s_c}\right) \quad \text{to order } 1/n \\ \Rightarrow \langle \frac{\partial^2 \mathcal{L}}{\partial r_a \partial s_c} \rangle &= -n\left(\frac{1}{1 - r_a s_c}\right) \quad \text{to order } 1/n \end{aligned} \tag{D.23}$$

The other three cases are calculated in a similar way. The net result is that

$$\frac{1}{n}\langle \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k} \rangle = \begin{cases} 1/(1 - \theta_j\theta_k) \text{ if both are AR or MA} \\ -1/(1 - \theta_j\theta_k) \text{ if one is AR and one is MA} \end{cases} \tag{D.24}$$

Appendix E

Curvature of total sample surprisal for an ARMA process - from periodogram

In this appendix, I will derive the expected value of the Hessian second derivative (see equation 2.3) of the total sample surprisal of a zero-mean, stationary, invertible ARMA time series process (see section 2.4), this time assuming only that the innovations are $IID(0, v_e)$ and have a finite fourth moment. As mentioned in Appendix D in equation D.10, the defining equation of the ARMA(p, q) process can be expressed as

$$(1 - r_p B) \dots (1 - r_1 B) x_j = (1 - s_q B) \dots (1 - s_1 B) e_j \tag{E.1}$$

where B is the backshift operator that represents going back one time period (i.e. $Bx_j = x_{j-1}, Be_j = e_{j-1}$), $\{x_1, \dots, x_n\}$ are the observations, $\{e_1, \dots, e_n\}$ are the IID$(0, v_e)$ innovations, and the AR roots $\{r_j\}$ and MA roots $\{s_j\}$ are less than one in magnitude, and are either real or occur in conjugate pairs. We have assumed invertibility here, so there is also a causal MA(∞) representation

$$x_j = \frac{(1 - s_q B) \dots (1 - s_1 B)}{(1 - r_p B) \dots (1 - r_1 B)} e_j = e_j + m_1 e_{j-1} + m_2 e_{j-2} + \dots \tag{E.2}$$

Setting the Fourier transform matrix F to be [87]

$$F_{kl} = \frac{1}{\sqrt{n}} w^{kl} \quad \text{where } w = \exp\left(-j \frac{2\pi}{n}\right), \text{ and } k, l \in [0, 1, \dots n-1] \tag{E.3}$$

and n is the sample size (here assumed even), the k-th Fourier transform variate for a sample $\{x_1, \ldots x_n\}$ is

$$\begin{aligned}\sqrt{n}v_k &= x_1 + w^k x_2 + \ldots + w^{(n-1)k} x_n \\ &= [e_1 + w^k e_2 + \ldots][1 + w^k m_1 + \ldots] \approx \sqrt{n}[F\vec{e}]_k \sqrt{n}[F\vec{m}]_k\end{aligned} \tag{E.4}$$

where $\vec{e} = [e_1 \ldots e_n]^T$ and $\vec{m} = [1\ m_1 \ldots m_{n-1}]^T$. So, for each discrete frequency index k, we get

$$v_k = [F\vec{e}]_k \sqrt{n}[F\vec{m}]_k \quad \vec{m} = [1\ m_1 \ldots m_{n-1}]^T \tag{E.5}$$

The first part, $[F\vec{e}]_k$, is a random variate with the same total variance as the innovations. However, by the central limit theorem, it is normally distributed to order $1/n$ as long as the innovation distribution has a finite fourth moment. The second part, $\sqrt{n}[F\vec{m}]_k$, is not random at all; it measures the frequency content of the MA(∞) representation.

The periodogram is $f_k = |v_k|^2$, and is approximately uncorrelated. The correlation of distinct periodogram variates f_k and f_l is

$$\mathrm{corr}(f_k, f_l) = -\frac{1}{n}\frac{\kappa_4}{v_e^2} \quad (k \neq l) \tag{E.6}$$

where κ_4 is the fourth cumulant of the innovations [114]. Hence, as far as the statistical properties of the periodogram are concerned, we may replace $[F\vec{e}]_k$ by $z_k\sqrt{v_e}$, where z_k is a complex N(0,1) variate (i.e. with 2 degrees of freedom). Note that since the observation sequence $\{x_1, \ldots x_n\}$ is real, there are just $(n/2)$ independent v_k variates. Their magnitudes are

$$|v_k| = \sqrt{v_e}|z_k|\sqrt{n}|F\vec{m}|_k = \sqrt{v_e}|z_k|\sqrt{\mathrm{PSD}_k} \tag{E.7}$$

where $\mathrm{PSD}_k = n|F\vec{m}|_k^2$ is the squared magnitude of the transfer function at frequency index k. The periodogram is

$$f_k = |v_k|^2 = (v_e \mathrm{PSD}_k)|z_k|^2 \tag{E.8}$$

Thus, $|v_k|^2/(v_e\mathrm{PSD}_k)$ is a standard chi-square(2) variate with order of approximation $(1/n)$. Its probability is

$$p(v_k) = \frac{1}{v_e \mathrm{PSD}_k} \exp\left(-\frac{|v_k|^2}{v_e \mathrm{PSD}_k}\right) \tag{E.9}$$

so the overall probability of the sequence $\{v_0, v_1, \ldots v_{n/2-1}\}$ is

$$p(v_0, \ldots v_{n/2-1}) = \prod_{k=0}^{n/2-1} \frac{1}{v_e \mathrm{PSD}_k} \exp\left(-\frac{|v_k|^2}{v_e \mathrm{PSD}_k}\right) \tag{E.10}$$

and the total sample surprisal is

$$\mathcal{L} = -\log p(v_0, \dots v_{n/2-1}) = \sum_{k=0}^{n/2-1} \left[\log\left(v_e \text{PSD}_k\right) + \frac{|v_k|^2}{v_e \text{PSD}_k} \right] \tag{E.11}$$

What we'll do next is take the variation of that total surprisal with respect to the model parameters, square it, take the expectation over all periodograms, and find the precision matrix entries, as suggested by equation 4.14. For example, for the innovation variance v_e, the squared variation is

$$\begin{aligned} \Delta^2\mathcal{L} &= \sum_{k=0}^{n/2-1} \left[\frac{\Delta v_e}{v_e} - \frac{\Delta v_e}{v_e^2} \frac{|v_k|^2}{\text{PSD}_k} \right]^2 \quad (v_e \text{ varying }) \\ \Rightarrow \langle \Delta^2\mathcal{L} \rangle &= \left(\frac{\Delta v_e}{v_e} \right)^2 \sum_{k=0}^{n/2-1} \langle \left(1 - \frac{|v_k|^2}{v_e \text{PSD}_k} \right)^2 \rangle = \frac{n}{2v_e^2} (\Delta v_e)^2 \end{aligned} \tag{E.12}$$

and thus the precision entry for v_e is $n/(2v_e^2)$.

As for the ARMA(p, q) model, let's take the AR roots $\{r_1, \dots r_p\}$ and $\{s_1, \dots s_q\}$ as being the quantities of interest. The squared variation with respect to the ARMA roots is:

$$\Delta^2\mathcal{L} = \sum_k \left[\frac{\Delta \text{PSD}_k}{\text{PSD}_k} - \frac{\Delta \text{PSD}_k}{\text{PSD}_k^2} \frac{|v_k|^2}{v_e} \right]^2 = \sum_k \left(\frac{\Delta \text{PSD}_k}{\text{PSD}_k} \right)^2 \left(1 - \frac{|v_k|^2}{v_e \text{PSD}_k} \right)^2 \tag{E.13}$$

Taking the expectation over all periodograms yields

$$\begin{aligned} \langle \Delta^2\mathcal{L} \rangle &= \sum_k \left(\frac{\Delta \text{PSD}_k}{\text{PSD}_k} \right)^2 \langle \left(1 - \frac{|v_k|^2}{v_e \text{PSD}_k} \right)^2 \rangle \\ &= \sum_k \left(\frac{\Delta \text{PSD}_k}{\text{PSD}_k} \right)^2 \approx n \int_0^\pi [\Delta \log \text{PSD}(\omega)]^2 \frac{\mathrm{d}\omega}{2\pi} \end{aligned} \tag{E.14}$$

where I have approximated the discrete sum by a continuous integral over the discrete frequency $\omega = 2\pi k/n$. For an ARMA(p, q) process, the squared magnitude of the transfer function is

$$\text{PSD}(\omega) = \frac{g(s_q, \omega) \dots g(s_1, \omega)}{g(r_p, \omega) \dots g(r_1, \omega)} \tag{E.15}$$

where

$$g(a, \omega) = (1 - ae^{j\omega})(1 - ae^{-j\omega}) = 1 + a^2 - 2a\cos(\omega) \tag{E.16}$$

Note that

$$\begin{aligned} \log g(a,\omega) &= -2(a\cos\omega + \frac{1}{2}a^2\cos 2\omega + \frac{1}{3}a^3\cos 3\omega + \ldots) \\ \Rightarrow \Delta\log g(a,\omega) &= -2(\cos\omega + a\cos 2\omega + a^2\cos 3\omega + \ldots)(\Delta a) \end{aligned} \tag{E.17}$$

Hence the expectation of the total squared variation of the sample surprisal is

$$\begin{aligned} \langle\Delta^2\mathcal{L}\rangle &\approx n\int_0^{\pi}[\Delta\log \mathrm{PSD}(\omega)]^2\frac{\mathrm{d}\omega}{2\pi} \\ &= 2n\int_0^{2\pi}[-\sum_{j=1}^{q}\Delta s_j(\cos\omega + s_j\cos 2\omega + \ldots) + \sum_{i=1}^{p}\Delta r_i(\cos\omega + r_i\cos 2\omega + \ldots)]^2\frac{\mathrm{d}\omega}{2\pi} \\ &= n\left[\sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\Delta s_i\Delta s_j}{1-s_is_j} + \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\Delta r_i\Delta r_j}{1-r_ir_j} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\Delta r_i\Delta s_j}{1-r_is_j}\right] \end{aligned} \tag{E.18}$$

That shows that the precision matrix entries for the ARMA roots are

$$\frac{1}{n}\langle\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\rangle = \begin{cases} 1/(1-\theta_i\theta_j) \text{ if both are AR or MA} \\ -1/(1-\theta_i\theta_j) \text{ if one is AR and one is MA} \end{cases} \tag{E.19}$$

Note that our only assumptions on the innovations are that they are $\mathrm{IID}(0, v_e)$ and that their fourth moment is finite. From the expression for $\log g(a,\omega)$ (equation E.17), we can see that

$$0 = \int_0^{2\pi}\log g(a,\omega)\ \mathrm{d}\omega = \int_0^{2\pi}\log \mathrm{PSD}(\omega)\ \mathrm{d}\omega = \int_0^{2\pi}\log\frac{\mathrm{PSD}_1(\omega)}{\mathrm{PSD}_2(\omega)}\ \mathrm{d}\omega \tag{E.20}$$

If two ARMA models with power spectral densities $\mathrm{PSD}_1(\omega)$ and $\mathrm{PSD}_2(\omega)$ are close together, so that $\mathrm{PSD}_1(\omega)/\mathrm{PSD}_2(\omega) \approx 1$ for all ω, then we can use the approximation

$$x = e^{\log x} \approx 1 + \log x + \frac{1}{2}\log^2 x \quad (x \approx 1) \tag{E.21}$$

to deduce

$$\frac{1}{2}\int_0^{2\pi}\log^2\frac{\mathrm{PSD}_1(\omega)}{\mathrm{PSD}_2(\omega)}\frac{\mathrm{d}\omega}{2\pi} \approx \int_0^{2\pi}\frac{\mathrm{PSD}_1(\omega)}{\mathrm{PSD}_2(\omega)}\frac{\mathrm{d}\omega}{2\pi} - 1 \tag{E.22}$$

Furthermore, if we define the sample autocovariance as

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h}x_jx_{j+h} \tag{E.23}$$

and the sample periodogram as

$$\hat{I}(\omega) = \sum_{-n+1}^{n-1} \hat{\gamma}(h) e^{-jh\omega} \tag{E.24}$$

then it can be shown that the residual variance

$$T = \int_0^{2\pi} \frac{\hat{I}(\omega)}{\mathrm{PSD}(\omega)} \frac{\mathrm{d}\omega}{2\pi} \tag{E.25}$$

achieves the same curvature at its minimum, i.e. where $\partial T/\partial r_i = \partial T/\partial s_j = 0$, we have

$$\frac{1}{2T}\langle \frac{\partial^2 T}{\partial \theta_i \partial \theta_j} \rangle = \begin{cases} 1/(1-\theta_i\theta_j) \text{ if both are AR or MA} \\ -1/(1-\theta_i\theta_j) \text{ if one is AR and one is MA} \end{cases} \tag{E.26}$$

although the derivation is lengthy and I will leave it out.

Appendix F

Covariance of symbol block counts for a serially correlated symbol sequence

The purpose of this appendix is to calculate the theoretical covariance between symbol block counts in a sample string of length $M \gg 1$, when the symbols are serially correlated. For the purposes of this appendix, I will assume that the symbols are integers in the range $\{1, 2, \ldots m\}$. I assume that the ensemble probability of a symbol block $\{j_1 \ j_2 \ldots j_N\}$, where N is the symbol block size, is calculable; see for example Equation 5.6 for the case of symbol sequences generated by a hidden Markov model. We will let $p(j_1 \ldots j_N)$ represent the ensemble probability of the symbol sequence $\{j_1 \ j_2 \ldots j_N\}$. Also, the hash sign (#) will be a "wild card" representing an arbitrary symbol.

Consider a symbol sequence $\{j_1 \ j_2 \ldots j_k \ldots j_M\}$ where the j_k's are individual symbols. Suppose, for the moment, that symbol blocks of size N are not correlated, i.e. that

$$p(j_1 \ldots j_M) = p(j_1 \ldots j_N)p(j_{N+1} \ldots j_{2N}) \ldots p(j_{(k-1)N+1} \ldots j_{kN}) \tag{F.1}$$

where $k = M/N$ (an integer). We will obtain results for finite N, and eventually take the limit as N becomes very large. In this scenario, the overall sequence probability is

$$p(j_1 \ldots j_M) = \prod_{|u|=N} p(u)^{n(u)} \tag{F.2}$$

where u is a symbol block of size N, $p(u)$ is its ensemble probability, and $n(u)$ records how many times the symbol block appears in the block sequence $\{u_1, u_2, \ldots, u_k\}$ in which $u_1 =$

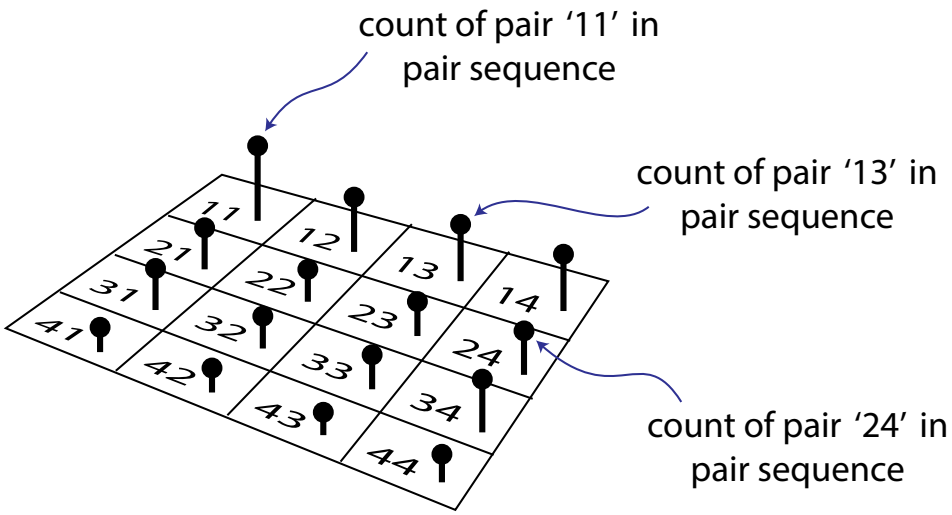


Figure F.1: Counting the occurrence of symbol pairs in a sequence

$\{j_1 \dots j_N\}$, $u_2 = \{j_{N+1} \dots j_{2N}\}$, and so forth. Under our assumptions so far, the $\{n(u)\}$ will have a multinomial distribution. In particular, in an ensemble of such sequences, the multinomial frequencies and covariances are

$$\langle n(u) \rangle = \left(\frac{M}{N}\right) p(u), \quad \text{cov}(n(u), n(v)) = \left(\frac{M}{N}\right) (p(u)\delta_{uv} - p(u)p(v)) \tag{F.3}$$

where u and v are symbol blocks of size N, and δ_{uv} is one if $u = v$, and zero otherwise. See section 3.1 for further details. We can now form linear combinations of the individual block counts and determine their expectations and covariances. In particular, any non-overlapping partition of the bins for symbol blocks of size N will also have a multinomial distribution.

We will look first at the simplest possible case, namely $N = 2$, and then see what happens as N increases indefinitely. We will take sequences of length M and divide them up into non-overlapping but contiguous symbol pairs, and count the number of occurrences of each symbol pair. Figure F.1 illustrates the case $m = 4$; in this case, there are $m^2 = 16$

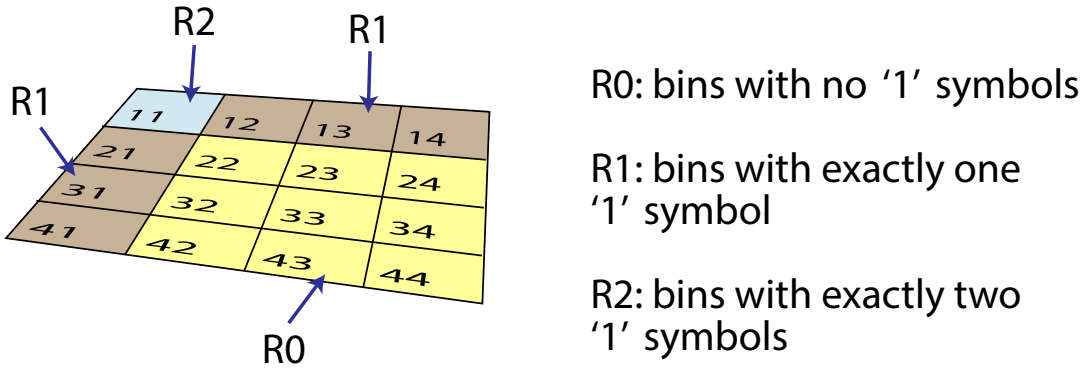


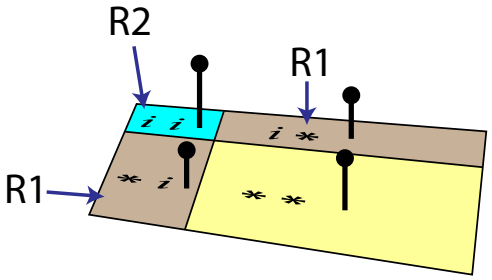
Figure F.2: Partition of symbol pair bins with respect to the symbol ‘1’

possible symbol pairs. Suppose that we are interested in the variance of the number of ‘1’ symbols that we see in output sequences. Figure F.2 illustrates how we can partition the bins in this case. Let $\hat{r}_2$ be the proportion of pairs falling into $R2$, and let $\hat{r}_1$ be the proportion of pairs falling into $R1$. Then their expectations are

$$\begin{aligned} r_2 &= p(11) = p_{11} \\ r_1 &= p(12) + p(13) + p(14) + p(21) + p(31) + p(41) \\ &= p(1\#) - p(11) + p(\#1) - p(11) \\ &= 2[p(1) - p(11)] = 2(p_1 - p_{11}) \quad \text{where } p_1 = p(1) \end{aligned} \tag{F.4}$$

Since $R1$ and $R2$ are part of a multinomial partition, we also have

$$\left(\frac{M}{2}\right) \operatorname{cov}(\hat{r}_1, \hat{r}_2) = \begin{bmatrix} r_1(1-r_1) & -r_1 r_2 \\ -r_1 r_2 & r_2(1-r_2) \end{bmatrix} \tag{F.5}$$



R1: bins containing exactly one '*i*' symbol
R2: bins containing two '*i*' symbols

r_1: expected proportion for R1
r_2: expected proportion for R2

p_1: expected proportion for symbol '*i*'
p_2: expected proportion for symbol pair '*ii*'

Figure F.3: Partition of symbol pair bins with respect to the symbol i

Let $\hat{T}_1$ be the number of '1' symbols in the sequence. Then

$$\begin{aligned} \hat{T}_1 &= \left(\frac{M}{2}\right)(\hat{r}_1 + 2\hat{r}_2) \\ \Rightarrow \left(\frac{2}{M}\right)\mathrm{var}(\hat{T}_1) &= \left(\frac{M}{2}\right)[1 \quad 2]\mathrm{cov}(\hat{r}_1, \hat{r}_2)[1 \quad 2]^T \\ &= (r_1 + 4r_2) - (r_1 + 2r_2)^2 \end{aligned} \tag{F.6}$$

Substituting $r_2 = p_{11}$ and $r_1 = 2(p_1 - p_{11})$ yields

$$\left(\frac{1}{M}\right)\mathrm{var}(\hat{T}_1) = (p_1 - p_1^2) - (p_{11} - p_1^2) \tag{F.7}$$

Of course, there is nothing special about symbol '1'. Figure F.3 shows how we would partition the symbol pair bins if we were interested in symbol i. In this case, $R2$ represents the bin for symbol pair (ii), and $R1$ represents the bins for symbol pairs $(*i)$ and $(i*)$ that contain exactly one i. In this case, we would have

$$\begin{aligned} r_2 &= p(ii) = p_{ii} \\ r_1 &= p(i\#) - p(ii) + p(\#i) - p(ii) \\ &= 2[p(i) - p(ii)] = 2(p_i - p_{ii}) \quad \text{where } p_i = p(i) \end{aligned} \tag{F.8}$$

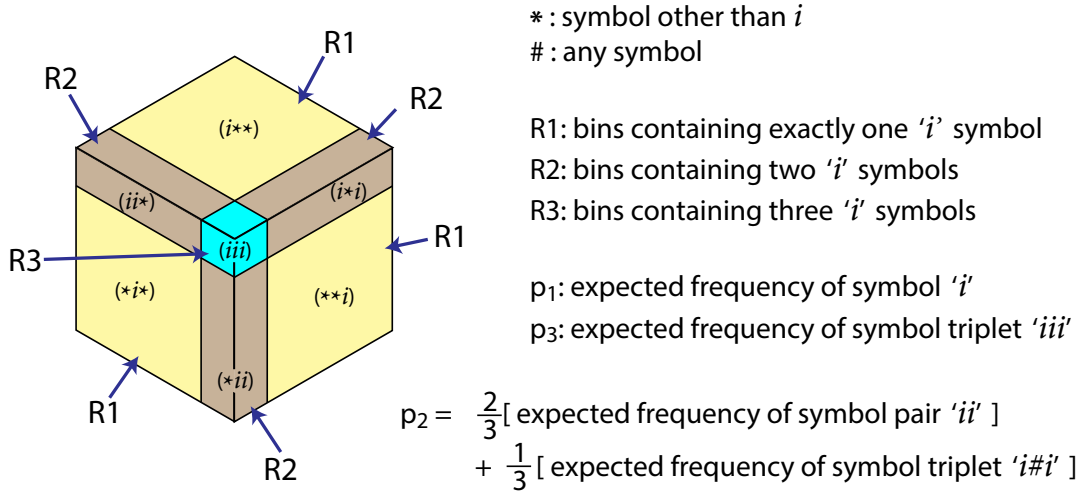


Figure F.4: Partition of symbol triplet bins with respect to a specific symbol i

Let T_i be the total count of i symbols in the sequence. Then, by a similar argument,

$$\begin{aligned}
\left(\frac{2}{M}\right)\langle T_i\rangle &= r_1 + 2r_2 = 2p_i \quad \text{where } p_i = p(i) \\
\left(\frac{2}{M}\right)\mathrm{var}(T_i) &= [1 \quad 2]\begin{bmatrix} r_1(1-r_1) & -r_1r_2 \\ -r_1r_2 & r_2(1-r_2)\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix} \\
&= 2(p_i - p_i^2) + 2(p_{ii} - p_i^2) \\
\left(\frac{1}{M}\right)\mathrm{var}(T_i) &= (p_i - p_i^2) + (p_{ii} - p_i^2)
\end{aligned} \tag{F.9}$$

where $p_{ii} = p(ii)$ and $p_i = p(i)$. The first term $p_i - p_i^2$ is a multinomial approximation. The second term is due to correlation between symbols in a pair. If there were no such correlation, then we would have $p_{ii} = p_i^2$ and $(1/M)\mathrm{var}(T_i) = p_i(1-p_i)$.

Suppose now that $N = 3$. We will divide sequences of length M into blocks of three symbols (triplets), and count the number of occurrences of each possible symbol block. For a specific symbol i, $R3$ will represent the bin for triplet (iii), $R2$ will represent the triplet bins with exactly 2 i symbols, and $R1$ will represent the triplet bins with exactly one i

symbol (see figure F.4). Their expectations, as proportions of triplets, are:

$$\begin{aligned} r_3 &= p(iii) = p_3 \\ r_2 &= p(ii*) + p(*ii) + p(i*i) \quad [*\text{: symbol other than } i] \\ &= 2p(ii) + p(i\#i) - 3p(iii) \\ &= 3(p_2 - p_3) \quad \text{where } 3p_2 = 2p(ii) + p(i\#i) \\ r_1 &= p(i**) + p(**i) + p(*i*) \\ &= 3p(i) - [4p(ii) + 2p(i\#i)] + 3p(iii) \\ &= 3p_1 - 6p_2 + 3p_3 \quad \text{where } p_1 = p(i) \end{aligned} \tag{F.10}$$

We are interested in the total count of i symbols in the sequence, $T_i = (M/3)(r_1+2r_2+3r_3)$. Its expectation and variance are:

$$\begin{aligned} \left(\frac{3}{M}\right)\langle T_i\rangle &= 3r_3 + 2r_2 + r_1 = 3p_1 \\ \left(\frac{3}{M}\right)\mathrm{var}(T_i) &= [3 \quad 2 \quad 1] \begin{bmatrix} r_3 - r_3^2 & -r_3 r_2 & -r_1 r_3 \\ -r_2 r_3 & r_2 - r_2^2 & -r_1 r_2 \\ -r_1 r_3 & -r_1 r_2 & r_1 - r_1^2 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \\ &= [3 \quad 2 \quad 1][\mathrm{diag}(\vec{r}\,) - \vec{r}\,\vec{r}^{\,T}][3 \quad 2 \quad 1]^T \end{aligned} \tag{F.11}$$

where $\vec{r} = [r_3 \;\; r_2 \;\; r_1]^T$. That simplifies to

$$\left(\frac{3}{M}\right)\mathrm{var}(T_i) = (9r_3 + 4r_2 + r_1) - (3p_1)^2 \tag{F.12}$$

which eventually works out to

$$\left(\frac{3}{M}\right)\mathrm{var}(T_i) = 3(p_1 - p_1^2) + 6(p_2 - p_1^2) \tag{F.13}$$

where $3p_2 = 2p(ii) + p(i\#i)$. So, we have

$$\left(\frac{1}{M}\right)\mathrm{var}(T_i) = (p_1 - p_1^2) + \frac{4}{3}[p(ii) - p_1^2] + \frac{1}{3}[p(i\#i) - p_1^2] \tag{F.14}$$

where $p(ii)$ is the expected proportion of symbol pairs (ii), and $p(i\#i)$ is the expected proportion of triples of the form $(i\#i)$.

The case $N = 4$ is harder to visualize, being 4-dimensional. Blocks of four symbols will be 'quadruplets'. We will partition the bins so that $R4$ represents the quadruplet $(iiii)$,

$R3$ represents the quadruplets with exactly three i symbols, $R2$ represents the quadruplets with exactly two i symbols, and $R1$ represents the quadruplets with exactly one i symbol. Following the same kind of analysis as for $N = 3$, we get the corresponding expectations

$$\begin{aligned} r_4 &= p(iiii) = p_4 \\ r_3 &= 4(p_3 - p_4) \quad \text{where } 4p_3 = 2p(iii) + p(i\#ii) + p(ii\#i) \\ r_2 &= 6(p_2 - 2p_3 + p_4) \quad \text{where } 6p_2 = 3p(ii) + 2p(i\#i) + p(i\#\#i) \\ r_1 &= 4(p_1 - 3p_2 + 3p_3 - p_4) \quad \text{where } p_1 = p(i) \end{aligned} \tag{F.15}$$

and setting $T_i = (M/4)(r_1 + 2r_2 + 3r_3 + 4r_4)$, we get

$$\begin{aligned} \left(\frac{1}{M}\right) \langle T_i \rangle &= p_1 \\ \left(\frac{1}{M}\right) \text{var}(T_i) &= (p_1 - p_1^2) + \frac{3}{2}[p(ii) - p_1^2] + [p(i\#i) - p_1^2] + \frac{1}{2}[p(i\#\#i) - p_1^2] \end{aligned} \tag{F.16}$$

where $p(i\#\#i)$ is the expected frequency of symbol blocks of length 4 with symbol i in the first and last positions.

From those last two cases with $N = 3$ and $N = 4$, we can set up the general case of $\text{var}(T_i)$ for arbitrary N, where T_i is the total count of i symbols in the sequence. The split into non-overlapping regions is as follows:

$$\begin{aligned} \text{exactly } N \text{ i: } r_N &= \binom{N}{0}\left[\binom{0}{0}p_N\right] \\ \text{exactly } (N-1) \text{ i: } r_{N-1} &= \binom{N}{1}\left[\binom{1}{1}p_{N-1} - \binom{1}{0}p_N\right] \\ \text{exactly } (N-2) \text{ i: } r_{N-2} &= \binom{N}{2}\left[\binom{2}{2}p_{N-2} - \binom{2}{1}p_{N-1} + \binom{2}{0}p_N\right] \\ \ldots \text{exactly 2 i: } r_2 &= \binom{N}{N-2}\left[\binom{N-2}{N-2}p_2 - \binom{N-2}{N-3}p_3 \ldots \pm \binom{N-2}{0}p_N\right] \\ \text{exactly 1 i: } r_1 &= \binom{N}{N-1}\left[\binom{N-1}{N-1}p_1 - \binom{N-1}{N-2}p_2 \ldots \pm \binom{N-1}{0}p_N\right] \end{aligned} \tag{F.17}$$

where p_1 is the ensemble probability of symbol (i), p_2 is the average ensemble probability of at least two symbols in a block of size N, and so forth. The variance of T_i is

$$\left(\frac{N}{M}\right) \text{var}(T_i) = N^2 r_N + (N-1)^2 r_{N-1} + \ldots + r_1 - (Np_1)^2 \tag{F.18}$$

Assuming for now that only the terms in p_2 and p_1 survive in equation F.18, we get

$$\begin{aligned}(\frac{N}{M})\text{var}(T_i) &= 4\binom{N}{N-2}[p_2] + \binom{N}{N-1}\left[p_1 - \binom{N-1}{N-2}p_2\right] - N^2p_1^2 \\ &= N(N-1)p_2 + Np_1 - N^2p_1^2\end{aligned} \tag{F.19}$$

By "average ensemble probability", I mean specifically:

$$\frac{N(N-2)}{2}p_2 = (N-1)p_{2,1} + (N-2)p_{2,2} + \ldots + p_{2,N-1} \tag{F.20}$$

where $p_{2,1} = p(ii), p_{2,2} = p(i\#i), p_{2,3} = p(i\#\#i)$, and so forth. That works out to

$$(\frac{1}{M})\text{var}(T_i) = (p_1 - p_1^2) + 2\left[\frac{N-1}{N}(p_{2,1} - p_1^2) + \frac{N-2}{N}(p_{2,2} - p_1^2) + \ldots \frac{1}{N}(p_{2,N-1} - p_1^2)\right] \tag{F.21}$$

and taking the limit as the block size N increases indefinitely, we get

$$(\frac{1}{M})\text{var}(T_i) = (p_1 - p_1^2) + 2\left[(p_{2,1} - p_1^2) + (p_{2,2} - p_1^2) + \ldots\right] \tag{F.22}$$

which is my final expression for $\text{var}(T_i)$.

As for the assumption that only p_1 and p_2 survive, we can prove that as follows. The coefficient of p_J in equation F.18 is

$$\begin{aligned}\text{coeff of } p_J &= \sum_{k=0}^{J-1}(-1)^k(J-k)^2\binom{N}{N-J+k}\binom{N-J+k}{N-J} \\ &= N(N-1)\ldots(N-J+1)\left[\frac{J}{(J-1)!} - \frac{J-1}{(J-2)!1!} + \frac{J-2}{(J-3)!2!}\cdots\frac{(-1)^{J-1}}{(J-1)!}\right]\end{aligned} \tag{F.23}$$

In order to evaluate that sum, note that

$$\begin{aligned}p(p-1)^{J-1} &= \binom{J-1}{J-1}p^J - \binom{J-1}{J-2}p^{J-1} \pm \ldots + \binom{J-1}{0}(-1)^{(J-1)}p \\ \rightarrow \frac{\partial}{\partial p}\left[p(p-1)^{J-1}\right] &= J\binom{J-1}{J-1}p^{J-1} - (J-1)\binom{J-1}{J-2}p^{J-2}\ldots\binom{J-1}{0}(-1)^{J-1} \\ \rightarrow \frac{\partial}{\partial p}\left[p(p-1)^{J-1}\right]_{p=1} &= (J-1)!\left[\frac{J}{(J-1)!} - \frac{J-1}{(J-2)!1!} + \frac{J-2}{(J-3)!2!}\cdots\frac{(-1)^{J-1}}{(J-1)!}\right]\end{aligned} \tag{F.24}$$

However, $(\partial/\partial p)[p(p-1)^{J-1}]_{p=1}$ has the value 1 for $J = 1, 2$ and zero for all integers $J > 2$. So, the only coefficients of p_J that survive in equation F.18 are those for p_1 and p_2.

What about the covariance between T_i and T_j, where T_i is the count of i symbols in a sequence of length M, and T_j is the count of j symbols in that sequence? The calculations are similar to those just described for $\mathrm{var}(T_i)$, and the result is

$$\begin{aligned}(\frac{1}{M})\mathrm{cov}(T_i, T_j) &= (\delta_{ij}p_i - p_i p_j) + 2\sum_{k=1}^{M-1}(s_{2,k} - p_i p_j) \quad \text{where} \\ s_{2,1} &= \frac{1}{2}\left[p(ij) + p(ji)\right], \quad s_{2,2} = \frac{1}{2}\left[p(i\#j) + p(j\#i)\right] \quad \text{and so forth}\end{aligned} \tag{F.25}$$

Finally, note that the arguments leading to equations F.19 and F.25 remain valid if symbol i is replaced by a symbol block u. In this way, we can calculate the covariance between sample counts for symbol blocks u and v, each of which are of size N:

$$\begin{aligned}(\frac{N}{M})\mathrm{cov}(T_u, T_v) &= (\delta_{uv}p(u) - p(u)p(v)) + 2\sum_{k=1}^{(M/N)-1}[s_{2,k} - p(u)p(v)] \quad \text{where} \\ s_{2,1} &= \frac{1}{2}\left[p(uv) + p(vu)\right], \quad s_{2,2} = \frac{1}{2}\left[p(u\#v) + p(v\#u)\right] \quad \text{and so forth}\end{aligned} \tag{F.26}$$

where # now stands for any symbol block with the same size as u and v. I give numerical examples of the validity of this result in Section 5.6, and use it to estimate the magnitude of the corrections to the chi-square goodness-of-fit statistic for the empirical symbol block frequencies $\{p(u)\}$ in Section 5.7.

Appendix G

Positive definite quadratic forms in normal variables

We are interested in positive definite quadratic forms in normal variables, i.e. $z = (\vec{x} - \vec{v})^T Q(\vec{x} - \vec{v})$ where $\vec{x} \sim N(0, I_m)$ and Q is positive definite. The subject is large - see for example [82] - but here I am interested in just one proposition, namely that when we hold the trace of Q constant at m, then the lowest type II error for offsets exceeding the type I threshold, is achieved when $Q = I_m$. In the context of cluster separation, $\vec{v}$ represents the "offset vector", namely the difference between cluster centres.

With reference to Figure G.1, λ_0 is the value of $z_0 = \vec{x}^T Q\vec{x}$ at which the cumulative distribution function (cdf) reaches $1 - T_1$, where T_1 is the type I error threshold, here taken to be 0.05. For a given offset vector $\vec{v}$, T_2 is the type II error, namely the value of the cumulative distribution function for $z_1 = (\vec{x} - \vec{v})^T Q(\vec{x} - \vec{v})$ at $z_1 = \lambda_0$. We want to see what happens to the spatial average of T_2 as $\vec{v}$ varies in direction.

First of all, rotate and scale Q so that it is diagonal, with diagonal entries $\{1+\delta_1, \ldots, 1+\delta_m\}$. We'll keep the trace of Q constant at m, in which case the $\{\delta_i\}$ add up to zero. We then have

$$z_0 = \sum_{i=1}^{m}(1 + \delta_i)x_i^2 \quad x_i \sim N(0, 1), \quad \sum_{i=1}^{m} \delta_i = 0 \tag{G.1}$$

We will examine the case where the eigenvalue perturbations are small, i.e. $|\delta_i| \ll 1$. Clearly, the mean of z_0 is m and its variance is $2(m + s)$ where $s = \sum \delta_i^2$. Relative to the probability distribution function (pdf) of a standard chi-square(m) variate, the pdf of z_0 flattens and widens, with a corresponding increase in λ_0 (see Figure G.2). In that figure,

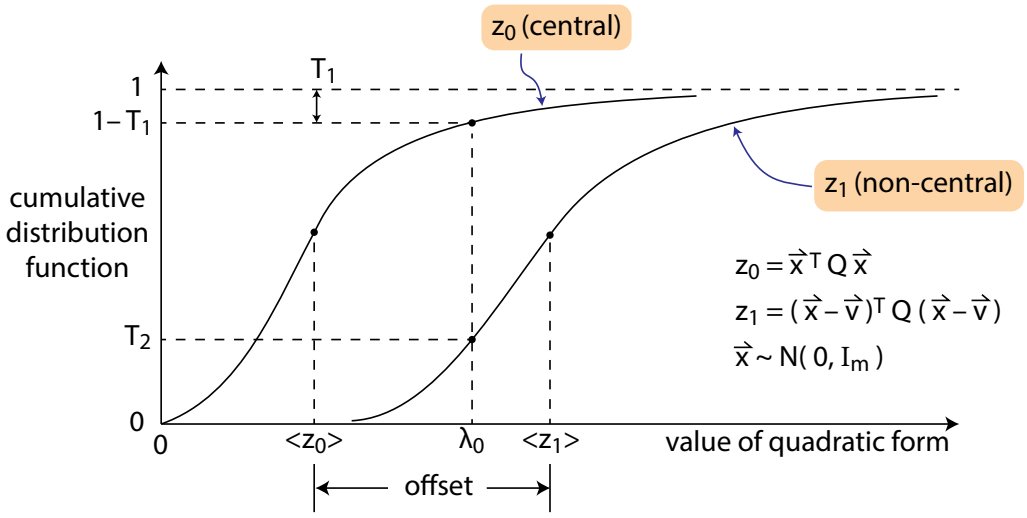


Figure G.1: Terminology for central and non-central quadratic forms

$q_0(z|m)$ is the cdf for a standard chi-square(m) variate, whereas $q_s(z|m)$ represents the cdf for z_0 when $s > 0$.

We can estimate the increase in λ_0 as follows. The case $m = 2$ has a known pdf (see Bausch [8] for its derivation). For $z_0 = (1+\delta)x_1^2 + (1-\delta)x_2^2$ where $x_1, x_2 \sim N(0,1)$, the exact pdf is

$$p(z,\delta) = \frac{1}{2\sqrt{1-\delta^2}} I_0\left(\frac{z\delta}{2(1-\delta^2)}\right) \exp\left(-\frac{1}{2(1-\delta^2)}z\right) \tag{G.2}$$

where I_0 is a modified Bessel function of the first kind. To first order in δ^2, that is

$$p(z,\delta) \approx \frac{1}{2}(1+\frac{1}{2}\delta^2)(1+\frac{1}{16}\delta^2 z^2)\exp\left(-\frac{1}{2}(1+\delta^2)z\right) \tag{G.3}$$

which suggests the approximation

$$p_s(z|m) = cz^{m/2-1}(1+bz^2)\exp\left(-az\right) \tag{G.4}$$

where a, b, and c are all linear functions of $s = \sum \delta_i^2$, and the formula is meant to be valid for all m. We can determine a, b, and c by requiring that $p_s(z|m)$ be normalized, and that

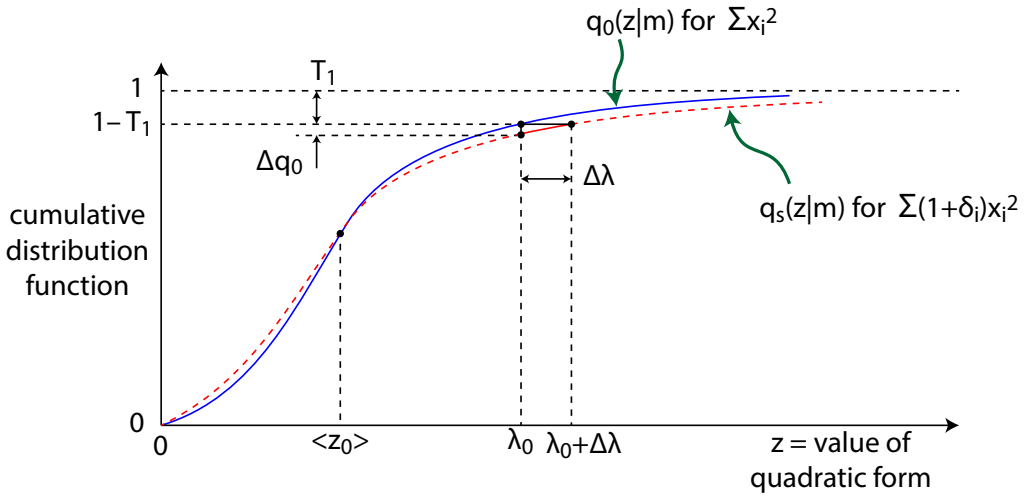


Figure G.2: Finding the increase in λ_0 when $s > 0$

its mean and variance be m and $2(m+s)$ respectively. Those requirements are that

$$\begin{aligned} &\int_0^\infty p_s(z|m)\mathrm{d}z = 1, \quad \int_0^\infty z p_s(z|m)\mathrm{d}z = m, \quad \text{and} \\ &\int_0^\infty z^2 p_s(z|m)\mathrm{d}z = m^2 + 2(m+s) \end{aligned} \tag{G.5}$$

Solving that set of equations to first order in s yields

$$p_s(z|m) = \frac{1}{2^{m/2}\Gamma(m/2)}\left(1+\frac{s}{4}\right) z^{m/2-1}\left[1+\frac{sz^2}{4m(m+2)}\right]\exp\left[-\frac{1}{2}\left(1+\frac{s}{m}\right)z\right] \tag{G.6}$$

In order to estimate the increase in λ_0 due to a non-zero s, we need $\partial p_s(z|m)/\partial s$, and working that out yields

$$\begin{aligned} \left.\frac{\partial p_s(z|m)}{\partial s}\right|_{s=0} &= \frac{1}{4}p_0(z|m) - \frac{1}{2}p_0(z|m+2) + \frac{1}{4}p_0(z|m+4) \quad \text{where} \\ p_0(z|m) &= \frac{1}{2^{m/2}\Gamma(m/2)} z^{m/2-1}\exp\left(-\frac{1}{2}z\right) \end{aligned} \tag{G.7}$$

Consequently, using the definition of the cumulative distributions

$$q_0(z|m) = \int_0^z p_0(u|m)\mathrm{d}u, \quad q_s(z|m) = \int_0^z p_s(u|m)\mathrm{d}u \tag{G.8}$$

we get

$$\left.\frac{\partial q_s(z|m)}{\partial s}\right|_{s=0} = \frac{1}{4}q_0(z|m) - \frac{1}{2}q_0(z|m+2) + \frac{1}{4}q_0(z|m+4) \tag{G.9}$$

That last expression is actually a special case of a very useful formula. The non-central chi-square pdf has the expansion [2]

$$p_0(z|m,\lambda) = e^{-\lambda/2}\sum_{j=0}^{\infty}\frac{(\lambda/2)^j}{j!}p_0(z|m+2j) \tag{G.10}$$

where λ is the non-centrality. This yields the derivatives

$$\begin{aligned}\frac{\partial p_0(z|m,\lambda)}{\partial\lambda} &= -\frac{1}{2}p_0(z|m,\lambda) + \frac{1}{2}p_0(z|m+2,\lambda)\\ \frac{\partial^2 p_0(z|m,\lambda)}{\partial\lambda^2} &= \frac{1}{4}p_0(z|m,\lambda) - \frac{1}{2}p_0(z|m+2,\lambda) + \frac{1}{4}p_0(z|m+4,\lambda)\end{aligned} \tag{G.11}$$

Using the cumulative distribution

$$q_0(z|m,\lambda) = \int_0^z p_0(u|m,\lambda)\mathrm{d}u \tag{G.12}$$

we get

$$\frac{\partial^2 q_0(z|m,\lambda)}{\partial\lambda^2} = \frac{1}{4}q_0(z|m,\lambda) - \frac{1}{2}q_0(z|m+2,\lambda) + \frac{1}{4}q_0(z|m+4,\lambda) = g(z|m,\lambda) \tag{G.13}$$

and so our earlier formula (equation G.9) becomes

$$\left.\frac{\partial q_s(z|m)}{\partial s}\right|_{s=0} = \frac{1}{4}q_0(z|m) - \frac{1}{2}q_0(z|m+2) + \frac{1}{4}q_0(z|m+4) = g(z|m,0) \tag{G.14}$$

With that derivative, we can now estimate the increase in λ_0 due to a non-zero s. In particular, with reference to Figure G.2, we have

$$\Delta\lambda \approx -\frac{1}{p_0(\lambda_0|m)}\Delta q_0 \approx -\frac{1}{p_0(\lambda_0|m)}g(\lambda_0|m,0)s \tag{G.15}$$

Note that $g(\lambda_0|m,0)$ is always negative when λ_0 is the 95% threshold of $q_0(z|m)$. Table G.1 shows representative values of the ratio $-g(\lambda_0|m,0)/p_0(\lambda_0|m)$. The important observation here is that $\Delta\lambda$ is always positive when $s = \sum\delta_i^2$ is greater than zero.

Table G.1: Values for multiplier that gives increase in λ_0 for s. It is always positive.

m	λ_0	$\Delta\lambda/s$
2	5.991	0.752
3	7.815	0.732
4	9.488	0.689
5	11.07	0.642

Table G.2: Approximate agreement of exact and estimated values for $\Delta\lambda$

m	δ_1	δ_2	δ_3	s	λ_0	Actual $\Delta\lambda$	Expected $\Delta\lambda$
2	0.0	0.0	n/a	0.0	5.9915	0.0	0.0
2	0.1	-0.1	n/a	0.02	6.0065	0.0150	0.0150
2	0.2	-0.2	n/a	0.08	6.0530	0.0615	0.0602
2	0.3	-0.3	n/a	0.18	6.1349	0.1434	0.1354
3	0.0	0.0	0.0	0.0	7.8147	0.0	0.0
3	-0.0577	0.1154	-0.0577	0.02	7.8291	0.0144	0.0146
3	-0.1	0.0	0.1	0.02	7.8294	0.0147	0.0146
3	-0.1154	0.2308	-0.1154	0.08	7.8711	0.0564	0.0586
3	-0.2	0.0	0.2	0.08	7.8740	0.0593	0.0586

I checked the approximation of equation G.15 against exact values of the cdf that I calculated using the Imhof method [61], as implemented in the R package CompQuadForm. In Table G.2, the column 'Expected' shows the result of equation G.15, while the column for $\Delta\lambda$ shows the exact values. Table G.2 shows that the approximation of equation G.15 is good to 5% , even at $|\delta_i| \approx 0.3$; recall that we started with the assumption that $|\delta_i| \ll 1$. The important point here is that $\Delta\lambda$ is always positive.

Next, we will look at the cumulative distribution function for $z_1 = \sum(1+\delta_i)(x_i + v\cos\theta_i)^2$ where the $\{\cos\theta_i\}$ are direction cosines for the offset vector $\vec{v}$, and v is its magnitude. With $\lambda = v^2$, the pdf for z_1 when $s = 0$ is just the non-central chi-square density $p_0(z|m,\lambda)$. Equation G.10 expresses $p_0(z|m,\lambda)$ as an affine linear combination of chi-square densities, so we can make the approximation

$$p_s(z|m,\lambda) \approx e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} p_s(z|m+2j) \tag{G.16}$$

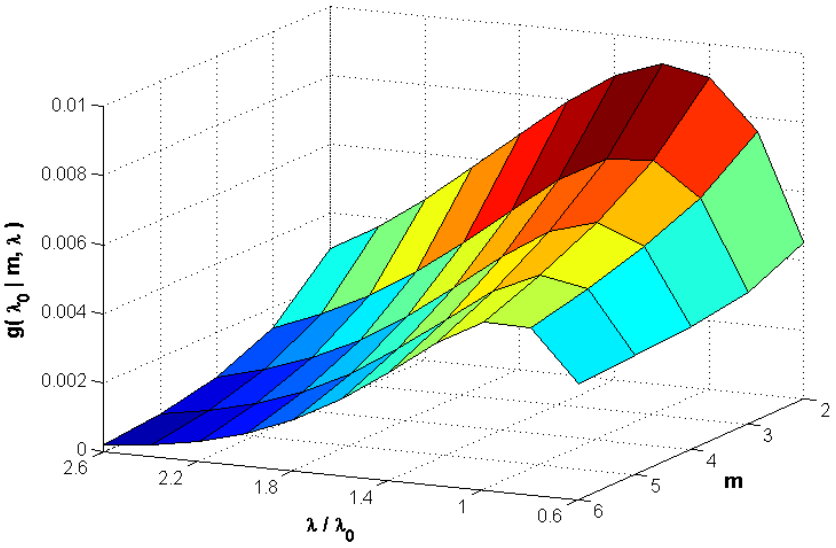


Figure G.3: Gradient of $q_s(z|m,\lambda)$ with respect to s when $z = \lambda_0$. It is always positive.

with corresponding cumulative distribution function

$$q_s(z|m,\lambda) = \int_0^\infty p_s(u|m,\lambda)\mathrm{d}u \tag{G.17}$$

and consequently get

$$\begin{aligned} \left.\frac{\partial q_s(z|m,\lambda)}{\partial s}\right|_{s=0} &= \frac{1}{4}q_0(z|m,\lambda) - \frac{1}{2}q_0(z|m+2,\lambda) + \frac{1}{4}q_0(z|m+4,\lambda) \\ &= g(z|m,\lambda) = \frac{\partial^2}{\partial\lambda^2}q_0(z|m,\lambda) \end{aligned} \tag{G.18}$$

Figure G.3 shows that in our region of interest ($\lambda > \lambda_0$), $g(\lambda_0|m,\lambda)$ is always positive. Now we can say for sure that, for a constant offset λ, the type II error will increase when $s = \sum \delta_i^2$ becomes non-zero (see Figure G.4). The total change in T_2 at λ_0 has two parts to it: the increase in $q_0(z|m,\lambda)$ due to a non-zero s, and the increase in $q_0(z|m,\lambda)$ due to $\Delta\lambda$. The total is

$$\Delta T_2 = T_{2,s} - T_{2,0} = \Delta q_1 + \Delta q_2 \approx \left[g(z|m,\lambda) - p_0(z|m,\lambda)\frac{g(z|m,0)}{p_0(z|m)}\right]_{z=\lambda_0}(s) \tag{G.19}$$

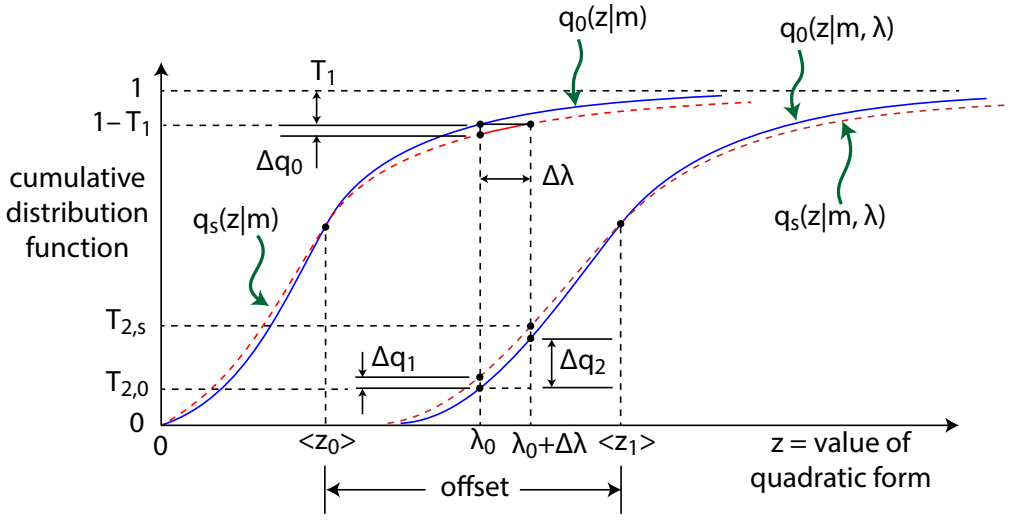


Figure G.4: Finding the increase in type II error when $s > 0$, for fixed offset λ

and we have now shown that both terms (Δq_1 and Δq_2) are positive in the region of interest.

That is not the whole story, though; what we are really after is the mean value of T_2, taken over all directions. The effective value of λ changes with direction. In particular, we have

$$\langle z_1 \rangle = \langle \sum_{i=1}^{m} (1+\delta_i)(x_i + v\cos\theta_i)^2 = m + v^2(1 + \sum_{i=1}^{m} \delta_i \cos^2\theta_i) \tag{G.20}$$

where the expectation is with respect to the distribution of $\vec{x}$. The quantity $\sum \delta_i \cos^2\theta_i$, when averaged over all directions, is zero:

$$\langle \sum_{i=1}^{m} \delta_i \cos^2\theta_i \rangle_d = \sum_{i=1}^{m} \delta_i \langle \cos^2\theta_i \rangle_d = \frac{1}{m} \sum_{i=1}^{m} \delta_i = 0 \tag{G.21}$$

where the averaging $\langle \bullet \rangle_d$ is over the probability simplex that the $\{\cos^2\theta_i\}$ live on. Since $\sum \delta_i \cos^2\theta_i$ is a linear function of the $\{\cos^2\theta_i\}$, its extreme values occur at vertices of the simplex [10], so we have

$$\max(\sum_{i=1}^{m} \delta_i \cos^2\theta_i) = (\delta_i)\text{max}, \quad \min(\sum_{i=1}^{m} \delta_i \cos^2\theta_i) = (\delta_i)\text{min} \tag{G.22}$$

So, as we vary the direction cosines, the effective offset λ goes from $v^2[1 + (\delta_i)_{\min}]$ to $v^2[1 + (\delta_i)_{\max}]$, with mean value v^2. Our assumption is that $|\delta_i| \ll 1$, so λ undergoes a small jitter about its mean value v^2. We know from Figure G.3 that $g(\lambda_0|m, \lambda) = \partial^2 q_0(\lambda_0|m, \lambda)/\partial\lambda^2$ is positive in the region of interest $\lambda > \lambda_0$ and well beyond. Assuming that $\partial^2 q_s(\lambda_0|m, \lambda)/\partial\lambda^2 \approx \partial^2 q_0(\lambda_0|m, \lambda)/\partial\lambda^2$, we can conclude by Jensen's inequality [28] that

$$\langle q_s(\lambda_0 + \Delta\lambda|m, \lambda)\rangle_d \geq q_s(\lambda_0 + \Delta\lambda|m, \langle\lambda\rangle_d) \tag{G.23}$$

where the averaging is over the squares of the direction cosines $\{\cos^2\theta_i\}$. But we already know from equation G.19 that

$$q_s(\lambda_0 + \Delta\lambda|m, \langle\lambda\rangle_d) > q_0(\lambda_0|m, \langle\lambda\rangle_d) \tag{G.24}$$

and equations G.23 and G.24 taken together imply that

$$\langle q_s(\lambda_0 + \Delta\lambda|m, \lambda)\rangle_d > q_0(\lambda_0|m, v^2) \tag{G.25}$$

which is what we intended to show all along, namely that for non-zero $\{\delta_i\}$, the type II error increases when averaged over all directions. Put another way: for positive definite matrices Q whose trace equals their rank, a type I error of 0.05, and offsets λ greater than the type I error threshold λ_0, the directional average of the type II error is minimized at $Q = I_m$.

As an example of the foregoing theory, consider the $m = 2$ case, where

$$\begin{aligned} \vec{x} \sim N(0, I_2), \quad z_0 = (1+\delta)x_1^2 + (1-\delta)x_2^2, \quad \text{and} \\ z_1 = (1+\delta)(x_1 - v\cos\theta)^2 + (1-\delta)(x_2 - v\sin\theta)^2 \end{aligned} \tag{G.26}$$

Figure G.5 shows the corresponding clusters and defines the offset vector $\vec{v}$ and offset angle θ. A positive value of δ means that the distance measures z_0 and z_1 favor the x_1 direction over the x_2 direction, i.e. the x_1 direction gets boosted.

Figure G.6 shows how the type II error varies with the angle between the offset vector $\vec{v}$ and the favored axis x_1, for $v^2 = 6$. That value of v^2 is approximately the type I threshold for a chi-square(2) variable at $1 - T_1 = 0.95$. I calculated these type II error rates using the Imhof method [61], as implemented in the R package CompQuadForm. Note in particular that, as we would expect, the type II error decreases when the offset vector $\vec{v}$ lies along the boosted direction.

Figure G.7 shows that, when averaged over all directions, the type II error has a net increase with respect to its value when $\delta = 0$. Although equation G.25 was based on the assumption that $|\delta_i| \ll 1$, Figure G.7 suggests that it is valid for any set $\{\delta_i\}$.

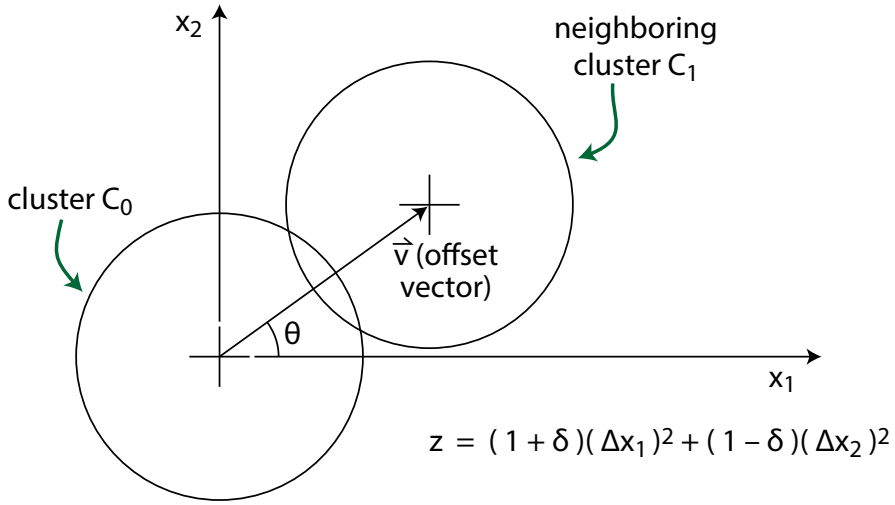


Figure G.5: Definition of offset vector $\vec{v}$ and angle θ for quadratic forms with $m = 2$

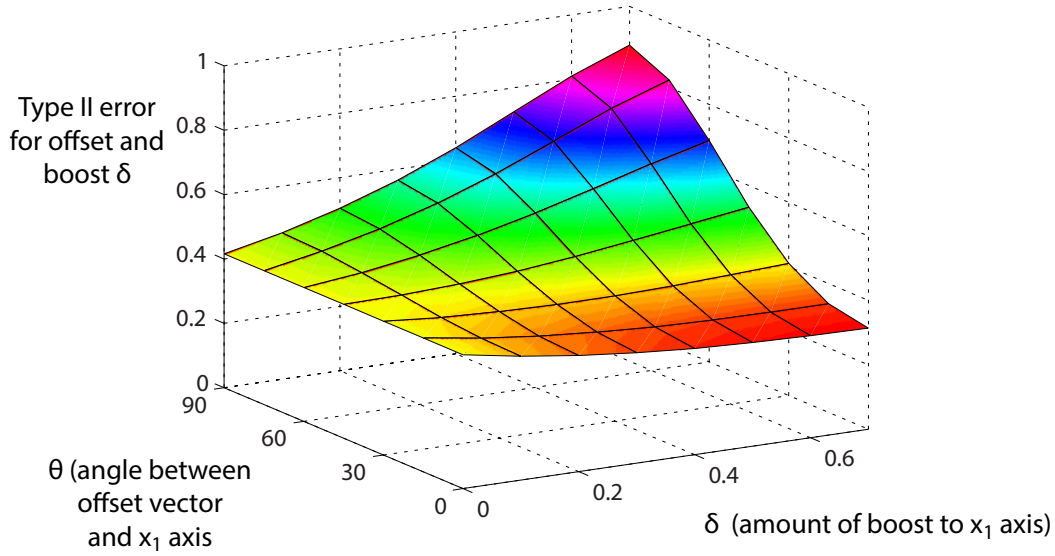


Figure G.6: Variation of type II error with angle between offset vector and boosted axis, for $m = 2$ and $v^2 = 6$

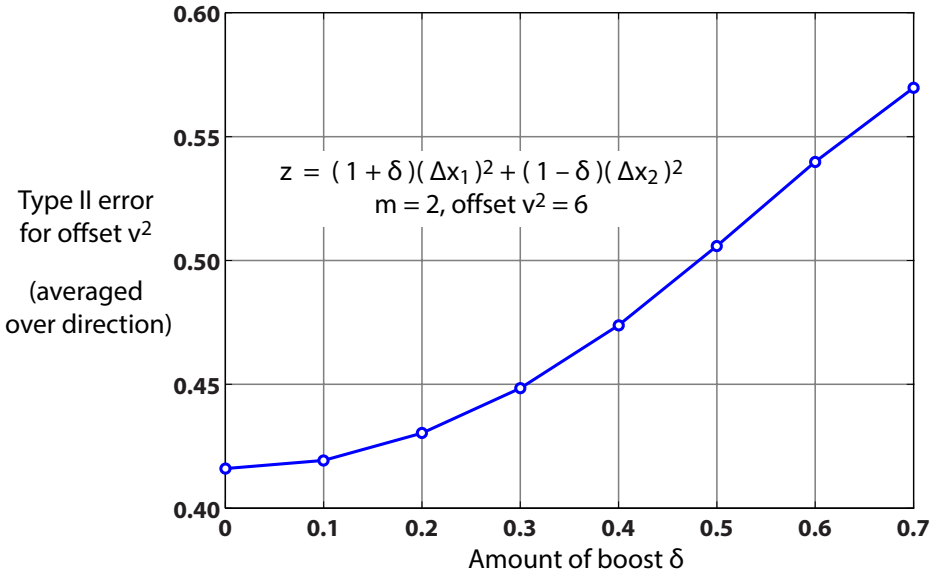


Figure G.7: Variation of type II error with amount of boost δ, for $m = 2$ and $v^2 = 6$. The type II error is averaged over direction and is smallest at $\delta = 0$.

Next, we will apply a similar style of analysis to the case of L_k distance measures. In particular, our goal is to show that when

$$\vec{z} \sim N(0, I_m), \quad d_0 = \sum_{i=1}^{m} |z_i|^k, \quad d_1 = \sum_{i=1}^{m} |z_i - v_i|^k, \tag{G.27}$$

for an offset vector $\vec{v}$, then the directionally averaged type II error for offsets $|\vec{v}\,|^2$ greater than the type I threshold is minimized at $k = 2$ (see Figure G.8). In the context of cluster separation, the offset vector $\vec{v}$ represents the difference between cluster centres.

Our starting point is the pdf of the chi-square(m) density, namely

$$p_0(x|m) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} \exp\left(-\frac{1}{2}x\right) \tag{G.28}$$

with corresponding cumulative density function

$$q_0(x|m) = \int_0^x p_0(u|m)\, \mathrm{d}u \tag{G.29}$$

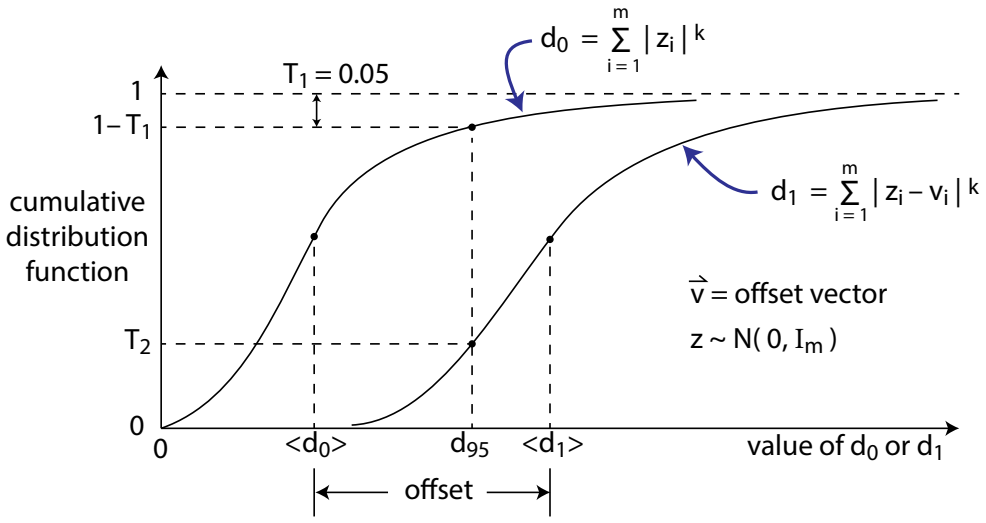


Figure G.8: Terminology for central and non-central L_k distance measures

The expectation of x^p under the chi-square(m) density is

$$\langle x^p \rangle = \int_0^\infty x^p p_0(x|m)\, \mathrm{d}x = 2^p \frac{\Gamma(m/2+p)}{\Gamma(m/2)} \tag{G.30}$$

If $z_i \sim N(0,1)$, then $|z_i|$ is a half-normal variate. We can also view $|z_i|$ as $\sqrt{x_1}$ where x_1 is a chi-square(1) variate. Thus

$$\langle |z_i|^k \rangle = \langle x_1^{k/2} \rangle = 2^{k/2} \frac{\Gamma(1/2+k/2)}{\Gamma(1/2)} \quad \text{and } \langle |z_i|^{2k} \rangle = \langle x_1^k \rangle = 2^k \frac{\Gamma(1/2+k)}{\Gamma(1/2)} \tag{G.31}$$

and the corresponding variance is

$$\mathrm{var}(|z_i|^k) = \langle |z_i|^{2k} \rangle - \langle |z_i|^k \rangle^2 = 2^k \left[\frac{\Gamma(1/2+k)}{\Gamma(1/2)} - \left(\frac{\Gamma(1/2+k/2)}{\Gamma(1/2)} \right)^2 \right] \tag{G.32}$$

If we now consider $d_0 = \sum_{i=1}^m |z_i|^k$, where the $\{z_i\}$ are independent $N(0,1)$ variates, then

its first two moments are

$$\begin{aligned}\langle d_0\rangle &= m\langle|z_i|^k\rangle = 2^{k/2}m\frac{\Gamma(1/2+k/2)}{\Gamma(1/2)}\\ \langle d_0^2\rangle &= m\ \mathrm{var}(|z_i|^k) + [m\langle|z_i|^k\rangle]^2\\ &= 2^k m\left[\frac{\Gamma(1/2+k)}{\Gamma(1/2)} + (m-1)\left(\frac{\Gamma(1/2+k/2)}{\Gamma(1/2)}\right)^2\right]\end{aligned} \tag{G.33}$$

Next, we need an approximation for the pdf of d_0. We will view $d_0^{2/k}$ as a "warped" chi-square(m) variate. Setting $z_i = \sqrt{u}\cos\theta_i$, where u is a true chi-square(m) variate and $\cos\theta_i$ is the direction cosine between $\vec{z}$ and the i-th axis, we get

$$d_0 = \sum_{i=1}^m |z_i|^k = u^{k/2}\left(\sum_{i=1}^m \cos^k\theta_i\right), \quad d_0^{2/k} = u\left(\sum_{i=1}^m \cos^k\theta_i\right)^{2/k} \tag{G.34}$$

Let x represent $d_0^{2/k}$. The dimensional factor $x^{m/2-1}$ in $p_0(x|m)$ of equation G.28 remains unchanged, but the exponential factor $\exp(-x/2)$ is replaced by the directional average

$$\exp\left(-\frac{1}{2}x\right) \Rightarrow \langle \exp -\frac{1}{2}x\left(\sum_{i=1}^m \cos^k\theta_i\right)^{-2/k} \rangle_d \tag{G.35}$$

where the averaging is over all directions. Since the $\{z_i\}$ are interchangeable, the all-ones axis $(1,1,\ldots 1)(m$ ones) is a symmetry axis, and we will approximate the spatial average of equation G.35 as

$$\left(\sum_{i=1}^m \cos^k\theta_i\right)^{-2/k} \approx a_1(1+\epsilon\cos 4\theta) \quad \theta\in[0,\pi/4] \tag{G.36}$$

where θ is the angle between $\vec{z}$ and the all-ones axis. That covers the positive orthant exactly for $m=2$, and covers it approximately for $m\geq 3$. With that approximation, we get

$$\begin{aligned}\langle \exp -\frac{1}{2}x\left(\sum_{i=1}^m \cos^k\theta_i\right)^{-2/k} \rangle_d &\approx \langle \exp -\frac{1}{2}a_1 x(1+\epsilon\cos 4\theta)\rangle_d\\ &\approx \exp\left(-\frac{1}{2}a_1 x\right)\frac{4}{\pi}\int_0^{\pi/4}\exp\left(-\frac{1}{2}a_1 x\epsilon\cos 4\theta\right)\,\mathrm{d}\theta\\ &\approx \exp\left(-\frac{1}{2}a_1 x\right)\left[1+\frac{\epsilon^2}{4}\left(\frac{a_1 x}{2}\right)^2 + \frac{\epsilon^4}{64}\left(\frac{a_1 x}{2}\right)^4 + \ldots\right]\end{aligned} \tag{G.37}$$

and so the pdf for $x = d_0^{2/k}$ is

$$p(x) \approx (\text{constant}) x^{m/2-1} \left[1 + \frac{\epsilon^2}{4}\left(\frac{a_1 x}{2}\right)^2 + \frac{\epsilon^4}{64}\left(\frac{a_1 x}{2}\right)^4\right] \exp\left(-\frac{1}{2}a_1 x\right) \tag{G.38}$$

That is a linear combination of chi-square variables; normalizing it yields

$$p(x) = a_1[(1 - a_4 - a_6)p_0(a_1 x|m) + a_4 p_0(a_1 x|m+4) + a_6 p_0(a_1 x|m+8)] \tag{G.39}$$

with corresponding cumulative distribution function

$$\begin{aligned} q(x) &= (1 - a_4 - a_6)q_0(a_1 x|m) + a_4 q_0(a_1 x|m+4) + a_6 q_0(a_1 x|m+8)] \quad \text{where} \\ a_4 &= \frac{\epsilon^2}{4}\frac{\Gamma(m/2+2)}{\Gamma(m/2)}, \quad a_6 = \frac{\epsilon^4}{64}\frac{\Gamma(m/2+4)}{\Gamma(m/2)}, \quad \text{and } x^{k/2} = d_0 = \sum_{i=1}^{m} |z_i|^k \end{aligned} \tag{G.40}$$

The approximation variables $\{a_1, \epsilon\}$ are then defined by matching the first and second moments of d_0:

$$\begin{aligned} \langle x^{k/2}\rangle &= \left(\frac{2}{a_1}\right)^{k/2} [(1 - a_4 - a_6)F(m, k/2) + a_4 F(m+4, k/2) + a_6 F(m+8, k/2)] = \langle d_0\rangle \\ \langle x^{k}\rangle &= \left(\frac{2}{a_1}\right)^{k} [(1 - a_4 - a_6)F(m, k) + a_4 F(m+4, k) + a_6 F(m+8, k)] = \langle d_0^2\rangle \end{aligned} \tag{G.41}$$

where $F(m,p) = \Gamma(m/2+p)/\Gamma(m/2)$, and $\langle d_0\rangle$ and $\langle d_0^2\rangle$ are given by equation G.33.

Given m and k, we can solve equation G.41 for a_1 and ϵ, and then calculate the value of d_{95} that satisfies $q(d_{95}^{2/k}) = 1 - T_1$ where T_1 is our desired type I error of 0.05. Table G.3 shows a comparison between experimental values d_{95} and those calculated from the approximation of equation G.40.

In Table G.3, the experimental values of d_{95} are derived by estimating the cdf's of 100 sets of 64000 randomly generated values of $\sum_{i=1}^{m} |z_i|^k$ where all the $\{z_i\}$ are $N(0,1)$. Table G.3 shows that this two-parameter approximation to the pdf of $d_0 = \sum_{i=1}^{m} |z_i|^k$ predicts the 95% thresholds to better than 0.3% accuracy for the range $\{m \le 3, k \le 4\}$.

The next step is to see what the approximation of equation G.40 says about the directional average of the type II error, given a specific offset vector $\vec{v} = [v_1 \ldots v_m]$. For that, we will need the cumulative distribution function of $x = d_1^{2/k} = [\sum_{i=1}^{m} |z_i - v_i|^k]^{2/k}$. In the spirit of equation G.16, we will approximate that cdf as

$$q(x, v^2) = (1 - a_4 - a_6)q_0(a_1 x|m, \lambda) + a_4 q_0(a_1 x|m+4, \lambda) + a_6 q_0(a_1 x|m+8, \lambda) \tag{G.42}$$

Table G.3: Close agreement between empirical d_{95} values and the predictions of equation G.40

m	k	a_1	ϵ	d_{95} **calculated**	**95% confidence**
2	1	0.671	-0.312	3.163	3.170±0.002
2	3	1.133	0.119	12.483	12.48±0.02
2	4	1.208	0.194	26.98	26.94±0.06
3	1	0.488	-0.330	4.279	4.283±0.002
3	3	1.240	0.144	16.49	16.47±0.02
3	4	1.420	0.290	37.08	36.99±0.08

where $q_0(u|m,\lambda)$ is the cumulative distribution function of the non-central chi-square distribution (see equation G.12), and λ is a suitable non-centrality parameter. We can choose λ as follows: with no offset, we had

$$\begin{aligned} x = d_0^{2/k} = u\left(\sum_{i=1}^{m}\cos^k\theta_i\right)^{2/k} &\approx u[a_1(1+\epsilon\cos 4\theta)]^{-1} \\ \Rightarrow a_1 x &\approx u/(1+\epsilon\cos 4\theta) \end{aligned} \tag{G.43}$$

where u is a true chi-square(m) variate. So, an offset of v^2 in u-space translates into an offset of $v^2/(1+\epsilon\cos 4\theta)$ in a_1x-space. Averaged over $\theta \in [0, \pi/4]$, that gives

$$\langle\lambda\rangle_d \approx v^2(1+\epsilon^2/2) \quad \text{(directional average)} \tag{G.44}$$

So, our approximation for the directional average of $q(x, v^2)$ is:

$$\begin{aligned} \langle q(x,v^2)\rangle_d = (1-a_4-a_6)q_0(a_1x|m,\lambda) + a_4 q_0(a_1x|m+4,\lambda) + a_6 q_0(a_1x|m+8,\lambda) \\ \text{where } \lambda = v^2(1+\epsilon^2/2) \end{aligned} \tag{G.45}$$

The specific value of $\langle q(x,v^2)\rangle_d$ that interests us is that for $x = d_{95}^{2/k}$ where d_{95} is the value of $d_0 = \sum_{i=1}^{m}|z_i|^k$ that solves $q(d_0^{2/k}) = 1 - T_1 = 0.95$. Table G.4 compares calculated and empirical values of $T_2 = \langle q(d_{95}^{2/k}, v^2)\rangle_d$ for representative values of m, k, and offset square magnitude.

In Table G.4, the experimental values of d_{95} are calculated by estimating the cdfs of 100 sets of 64000 randomly generated values of $d_0 = \sum_{i=1}^{m}|z_i|^k$ where all the $\{z_i\}$ are $N(0,1)$. The T_2 values are calculated by estimating the cdfs of 400 sets of 64000 randomly

Table G.4: Comparison of empirical T_2 values and the predictions of equation G.45. They agree to within 10%. For fixed m and offset v^2, the L_k distance has the lowest type II error.

m	k	v^2	T_2 **calculated**	**95% confidence**
2	1	6.0	0.439	0.445±0.004
2	2	6.0	0.416	0.416±0.002
2	3	6.0	0.420	0.419±0.002
2	4	6.0	0.424	0.424±0.002
2	1	8.4	0.278	0.291±0.004
2	2	8.4	0.260	0.261±0.002
2	3	8.4	0.263	0.264±0.002
2	4	8.4	0.266	0.266±0.002
3	1	8.0	0.390	0.390±0.004
3	2	8.0	0.346	0.347±0.002
3	3	8.0	0.354	0.355±0.002
3	4	8.0	0.370	0.364±0.002
3	1	11.0	0.229	0.232±0.003
3	2	11.0	0.196	0.198±0.002
3	3	11.0	0.202	0.204±0.001
3	4	11.0	0.214	0.211±0.002
4	1	9.5	0.370	0.355±0.004
4	2	9.5	0.309	0.311±0.002
4	3	9.5	0.323	0.320±0.002
4	4	9.5	0.367	0.334±0.002
4	1	13.3	0.196	0.194±0.004
4	2	13.3	0.154	0.155±0.001
4	3	13.3	0.163	0.164±0.001
4	4	13.3	0.194	0.177±0.002

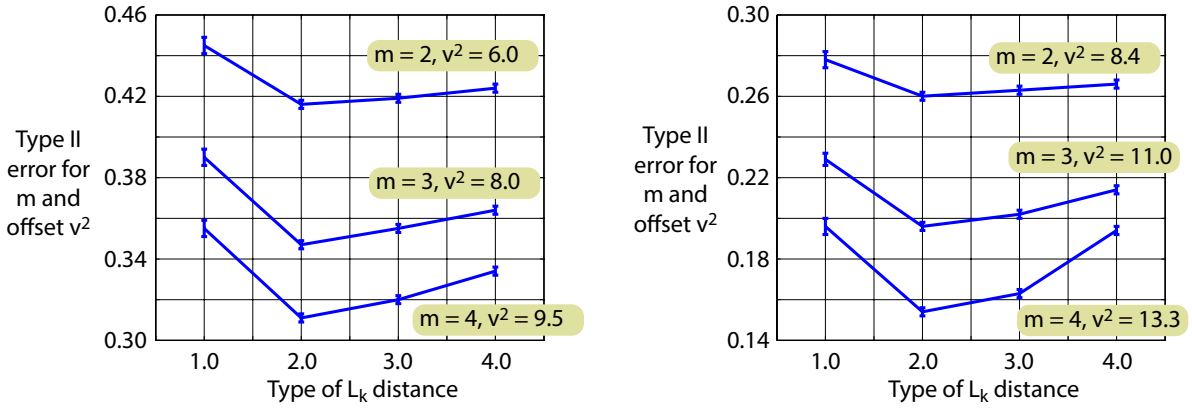


Figure G.9: Empirical type II error rates for L_k distance measures, for representative values of m, k, and offset v^2. The type II errors are averaged over direction. The error bars represent the 95% confidence limits.

generated values of $d_1 = \sum_{i=1}^{m} |z_i - v_i|^k$, where for each set the direction of $\vec{v}$ is chosen from a uniform distribution over all directions. What Table G.4 shows is that our two-parameter approximation of the pdf of $d_1 = \sum_{i=1}^{m} |x_i - v_i|^k$ predicts the directionally-averaged T_2 values to better than 10% accuracy over the range depicted. Note that equation G.45 for the non-central cumulative distribution of d_1 is a rougher approximation than equation G.40 for the central cumulative distribution of d_0.

Figure G.9 presents the empirical results of Table G.4 in chart form. The important observation here is that the L_2 distance always has the lowest directionally-averaged type II error. The foregoing development does not constitute a proof, but is a demonstration that under a simple model of the relevant probability distributions, and also empirically, the directionally-averaged type II error for cluster separation with the L_k distance measure is minimized with $k = 2$.