The Kernel Mutual Information

Arthur Gretton, Ralf Herbrich, Alex Smola

16 April 2003

Abstract

We introduce two new functions, the kernel covariance (KC) and the kernel mutual information (KMI), to measure the degree of independence of several continuous random variables. The former is guaranteed to be zero if and only if the random variables are pairwise independent; the latter shares this property, and is in addition an approximate upper bound on the mutual information, as measured near independence, and is based on a kernel density estimate. We show that Bach and Jordan's kernel generalised variance (KGV) is also an upper bound on the same kernel density estimate, but is looser. Finally, we suggest that the addition of a regularising term in the KGV causes it to approach the KMI, which motivates the introduction of this regularisation. The performance of the KC and KMI is verified in the context of instantaneous independent component analysis (ICA), by recovering both artificial and real (musical) signals following linear mixing[1].

Acknowledgement. The authors would like to thank to Jean-Yves Audibert, for discovering a gap in our original reasoning for the KMI proof; Francis Bach and Michael Jordan, for providing the kernel ICA code on the web, and for various helpful comments; Aapo Hyvärinen, for his advice on ICA methods and applications; and Malte Kuss, who provided us with an excellent explanation of the geometric properties of the canonical correlation. This work also benefits from helpful discussions with Christophe Andrieu, Olivier Bousquet, Arnaud Doucet, John Fisher, Thore Graepel, James Hopgood, Chih-Jen Lin, Erik Miller, Peter Rayner, and Matthias Seeger.

[1] This version contains changes to the formatting and minor corrections to the background section, compared with that originally posted on April 16 2003.

Contents

Chapter 1

Introduction

The problem of separating mixtures of signals, so as to recover the original signals prior to mixing, is a much studied challenge in signal processing. Methods of solution generally depend on the nature of the signals, and the manner in which they are mixed; in particular, a criterion known as the *contrast function* is required to determine when the demixing is successful. We assume here that the original signals are i.i.d. according to some unknown probability distributions, and are combined in a scalar mixing process: demixing is then achieved by ensuring that the recovered signals are statistically independent. This is the framework for *instantaneous ICA*.

A measure of statistical independence between two random variables is the *mutual information* [25], which for random vectors $\mathbf{x}, \mathbf{y}$ is zero if and only if the random vectors are independent. This may also be interpreted as the KL divergence $D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x},\mathbf{y}}||\mathbf{f}_{\mathbf{x}}\mathbf{f}_{\mathbf{y}})$ between the joint density and the product of the marginal densities; the latter quantity generalises readily to distributions of more than two random variables. We therefore propose two quantities, based on the mutual information, that may be used as contrast functions in ICA. The first, which we call the kernel covariance (KC), can be shown to be zero if and only if the random variables are independent. The second function, the kernel mutual information (KMI), is an upper bound on the Parzen window estimate of the mutual information, and is also zero if and only if the random variables are independent. Both functions bear a strong resemblance to the kernel canonical correlation (KCC) and kernel generalised variance (KGV) introduced by Bach and Jordan [7]: indeed, we demonstrate that the KGV can also be thought of as a (looser) upper bound on the same Parzen window estimate. An important advantage of the derivation described herein, however, is that it addresses the behaviour of the contrast functions for finite kernel sizes, rather than relying on a limiting argument in which the kernel size approaches zero, as in [7] (the latter proof may in any case require further refinement: see Appendix A.6.2). In addition, our approach allows us to apply well established methods for selecting kernel size as a function of the number of observations; see for instance [80].

The ICA framework has found many practical applications. Perhaps one of the earliest and best known is the separation of multiple audio signals recorded in a room, although basic ICA has been superseded by algorithms that take better account of the mixing process and signal properties (reverberation in the room, statistical properties of human speech, movement of the sources; see for instance [2]). A more successful application of instantaneous linear ICA is in removing eye blinks and electronic artefacts from EEG recordings, so as to isolate the weaker signals arising from various mental activities; one such study is [59]. ICA has also been used in [17] to determine brain regions used in visualisation, when applied to event-related fMRI experiments. Compared with generalised linear models, which depend on a particular form of the haemodynamic response being assumed, ICA permits the identification of additional regions of activation relating to the stimulus. Further applications of ICA are described in [49, 23], notably basis function determination for natural images and financial data analysis. The KCC was applied in [85] to find correlations between documents in English and French with identical meanings, thus revealing the features that best represent the semantic information pairs of documents have in common.

The instantaneous linear ICA problem is introduced in Chapter 2, which also contains a review of methods that have previously been used to address it. The KC and KCC contrast functions for the 2-variable case are derived in Chapter 3, and their behaviour at Independence is investigated. Chapter 4 contains the principal results of this study. A multivariate Gaussian approximation to the mutual information (again, for 2 variables), which holds near independence, is introduced in Section 4.1. Upper bounds on the Parzen window estimate of this quantity are derived in Section 4.2; these constitute the KMI and KGV contrast functions. A generalisation to more than 2 variables is presented in Section 4.3. The procedure used to apply the kernel contrast functions to ICA, which includes a method for reducing computational cost and a gradient descent technique, is described in Chapter 5. Finally, we show in Chapter 6 that the performance of the KMI and KC contrasts, when used in ICA, is competitive with that of the KGV and KCC contrasts respectively, and that the kernel based methods outperform many traditional ICA algorithms.

Chapter 2

ICA with linear mixing

In this chapter, we describe the goal of instantaneous independent component analysis (ICA), and review some approaches to this problem. The discussion draws on the numerous existing surveys of ICA and related methods, including [49, 57, 23, 42]; see also [24] for a discussion of older literature on the topic. We begin in Section 2.1 by describing the general framework of linear instantaneous ICA, independent of any particular method used to solve it. The main methodologies that have previously been used for ICA are then described in Section 2.2.

2.1 Problem statement

We begin our discussion with a general description of the problem we wish to solve. We are given m samples $\boldsymbol{t} := (\mathbf{t}_1, \ldots, \mathbf{t}_m)$ of the n dimensional random vector $\mathbf{t}$, which are drawn independently and identically from the distribution $\mathbf{P}_\mathbf{t}$. The vector $\mathbf{t}$ is related to the random vector $\mathbf{s}$ (also of dimension n) by the *scalar* mixing process

$$\mathbf{t} = \mathbf{B}\mathbf{s}, \tag{2.1.1}$$

where $\mathbf{B}$ is a matrix with full rank[1]. We refer to our ICA problem as being *instantaneous* as a way of describing the dual assumptions that any observation $\mathbf{t}$ depends only on the sample $\mathbf{s}$ at that instant, and that the samples $\mathbf{s}$ are drawn independently and identically.

The components $\mathbf{s}_i$ of $\mathbf{s}$ are assumed to be mutually independent: this model codifies the assumption that the sources are generated by unrelated phenomena (for instance, one component might be an EEG signal from the brain, while another could be due to electrical noise from nearby equipment). Mutual independence has the following definition [65]:

Definition 2.1.1 (Mutual independence). Suppose we have a random vector $\mathbf{s}$ of dimension n. We say that the components $\mathbf{s}_i$ are *mutually* independent if and only if

$$\mathbf{f}_\mathbf{s}(\mathbf{s}) = \prod_{i=1}^{n} \mathbf{f}_{\mathbf{s}_i}(s_i). \tag{2.1.2}$$

It follows easily that the random variables are *pairwise* independent if they are *mutually* independent; i.e. $\mathbf{f}_{\mathbf{s}_i}(s_i)\mathbf{f}_{\mathbf{s}_j}(s_j) = \mathbf{f}_{\mathbf{s}_i,\mathbf{s}_j}(s_i, s_j)$ for all $i \neq j$. The reverse does *not* hold, however: pairwise independence does not guarantee mutual independence.

[1] That $\mathbf{B}$ is square means the number of sources is equal to the number of sensors. Full rank is required to ensure that no two sources in $\mathbf{s}$ are mixed with exactly identical coefficients (which would imply the sources coincide perfectly). As discussed in [73], we may consider (2.1.1) when the number of sources is less than the number of sensors by a change of basis, although this may be made more difficult in the presence of noise.

Our goal is to recover $\mathbf{s}$, given only the mixing model (2.1.1), and the fact that the components of $\mathbf{s}$ are mutually independent. Thus, we wish for an estimate $\mathbf{V}$ of the inverse of the matrix $\mathbf{B}$, such that the recovered vector $\mathbf{x} = \mathbf{VBs}$ has mutually independent components (as measured according to Definition 2.1.1).

It turns out that the problem described above is indeterminate in certain respects. For instance, our measure of independence does not change when the ordering of elements in $\mathbf{x}$ is swapped; in addition, components of $\mathbf{x}$ may be scaled by different constant amounts, while still retaining their independence with respect to the remaining components. We therefore modify our problem definition slightly, and choose $\mathbf{V}$ such that $\mathbf{VB} = \mathbf{PS}$, where $\mathbf{P}$ is a permutation matrix and $\mathbf{S}$ is a scaling matrix (in other words, $\mathbf{S}$ is a matrix with non-zero elements on the diagonals *only*). The associated random vector

$$\mathbf{x} = \mathbf{PSs}$$

clearly has independent components.

Mutual independence is generally difficult to determine. In the case of scalar mixing, however, we are able to find a unique optimal unmixing matrix $\mathbf{V}$ using only the *pairwise* independence between elements of $\mathbf{x}$, which is equivalent to recovering the *mutually* independent terms of $\mathbf{s}$ (up to permutation and scaling). This is due to the following theorem [24].

Theorem 2.1.2 (Mutual independence in linear ICA). *Let* $\mathbf{s}$ *and* $\mathbf{x}$ *be two random vectors with dimension* n*, related according to* $\mathbf{x} = \mathbf{As}$*, for which the underlying densities do not contain delta functions. Let* $\mathbf{s}$ *contain at most one Gaussian component. Then the properties*

- *The components of* $\mathbf{x}$ *are pairwise independent*
- *The components of* $\mathbf{s}$ *are mutually independent*
- $\mathbf{A} = \mathbf{PS}$*, where* $\mathbf{P}$ *is a permutation matrix, and* $\mathbf{S}$ *a full rank scaling matrix*

are equivalent.

There is a third identifiability limitation, in addition to those due to permutation and scaling. This is illustrated by the case $\mathbf{s} = [\mathsf{s}_1\ \mathsf{s}_2]^\top$, where s_1 and s_2 are Gaussian random variables with equal variance. These are combined with a pure rotation matrix,

$$\mathbf{B} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Clearly, the density of $\mathbf{t}$ can be factorised in the manner of Definition 2.1.1, regardless of θ; thus, we cannot use the mutual independence of components of $\mathbf{x} = \mathbf{VBs}$ to invert $\mathbf{B}$. Likewise, if $\mathbf{t}$ is *deterministic* (constant for each element of the sample), then $\mathbf{B}$ cannot be inverted using a measure of independence. The following theorem is a more general statement of these concepts [28].

Theorem 2.1.3 (Independence and Gaussianity). *Let* $\mathbf{s}$ *and* $\mathbf{x}$ *be two random vectors, where the components of* $\mathbf{s}$ *are mutually independent, and the components of* $\mathbf{x}$ *are pairwise independent. Furthermore, let* $\mathbf{x} = \mathbf{As}$*, where* $\mathbf{A}$ *has two or more entries in the* j*th column. Then* s_j *is Gaussian or deterministic.*

In any practical instantiation of the above framework, we are provided with a sample $\boldsymbol{t}$ rather than the true distribution $\mathbf{P}_\mathbf{t}$. Thus, any determination of independence must be made empirically on the basis of this sample. A selection of criteria used to accomplish this goal is described in Sections 2.2.2,2.2.3, and 2.2.4; this is also the application we use to test the performance of the kernel covariance and kernel mutual information.

We note at this point that if elements $\mathbf{t}_i$, $\mathbf{t}_j$ in the sample $\boldsymbol{t}$ are *not* drawn independently for $i \neq j$ (for instance, if they are generated by a random process with non-zero correlation between the

outputs at different times), then an entirely different set of approaches can be brought to bear[2]. For instance, the independent sources are modeled in [67] as generalised autoregressive processes (with an inverse *cosh* noise distribution), which allows their separation using maximum likelihood principles. Alternatively, the sources may be modeled as stationary random processes, in which case they can be recovered when the *spectral* covariance matrices are diagonal [11, 73]; if non-stationary, the sources may be unmixed by jointly diagonalising the covariance matrices at each time point [72] (see also [2, 44] on the subject of demixing random processes given convolutive mixing). An elegant overview of these ICA methods and the links between them is given in [22]. Although the present study concentrates entirely on the i.i.d. case, we will briefly address random processes with time dependencies in Chapter 7, when describing possible extensions to our work.

2.2 Review of ICA methods

In this section, we describe the various approaches that have previously been used in ICA. The problem of finding the inverse $\mathbf{V}$ of $\mathbf{B}$ is broken down into smaller steps, for greater ease of solution. First, we remove the mean from the observations $\mathbf{t}$ (our estimate of $\mathbf{s}$ then also has zero mean). Next, we decompose the demixing matrix as $\mathbf{V} = \mathbf{WQ}$, where $\mathbf{Q}$ is a whitening matrix and $\mathbf{W}$ is an orthogonal matrix, as described in Section 2.2.1. It is generally simpler to determine $\mathbf{Q}$ and $\mathbf{W}$ separately than to compute $\mathbf{V}$ in its entirety, although a small loss of accuracy results from this procedure. In the remaining parts of this section, we introduce the three main methods used to determine when the independent sources have been recovered, given the possible choices of $\mathbf{W}$. The first approach maximises the likelihood (Section 2.2.2), the second applies the mutual information with a specific density model (Section 2.2.3), and the third draws on various density estimation techniques to compute the mutual information (Section 2.2.4).

We do not describe gradient descent on $\mathbf{W}$, however; this may for instance be accomplished using the natural gradient [3] and relative gradient [20] techniques, which result in similar algorithms despite their different motivating principles. Our algorithm for computing $\mathbf{W}$ follows [7] in performing gradient descent on the Stiefel manifold, as described in Section 5.2.

2.2.1 Preprocessing

The steps described in this section may be found in any discussion on ICA, for instance [23, 49, 42]. First, we *whiten* the observations $\boldsymbol{t}$, using the operation

$$\tilde{\mathbf{t}}_i = \mathbf{Q}\mathbf{t}_i, \tag{2.2.1}$$

for each $\mathbf{t}_i \in \boldsymbol{t}$, such that the new observations $\tilde{\boldsymbol{t}} := (\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_m)$ have a *unit (empirical) covariance matrix*. Our estimate of the demixing matrix then becomes

$$\mathbf{V} := \mathbf{WQ}, \tag{2.2.2}$$

where $\mathbf{W}$ is an *orthogonal* matrix (which is to say that $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$; $\mathbf{W}$ may thus be written as a rotation matrix with arbitrary row/column permutations). Although the determination of $\mathbf{W}$ remains difficult, there are only $n(n-1)$ degrees of freedom involved in this problem [49], as opposed to the n^2 degrees of freedom present in the estimation of $\mathbf{V}$. In the remainder of this section, we describe a way to estimate[3] $\mathbf{Q}$. If we wish the *population* random variable $\tilde{\mathbf{t}}$ to have a unit covariance matrix, we may write

$$\begin{aligned}
\mathbf{I} &= \mathsf{E}_{\tilde{\mathbf{t}}}\left(\tilde{\mathbf{t}}\tilde{\mathbf{t}}^\top\right) - \mathsf{E}_{\tilde{\mathbf{t}}}\left(\tilde{\mathbf{t}}\right)\mathsf{E}_{\tilde{\mathbf{t}}}\left(\tilde{\mathbf{t}}\right)^\top \\
&= \mathsf{E}_{\mathbf{t}}\left(\mathbf{Q}\mathbf{t}\mathbf{t}^\top\mathbf{Q}^\top\right) - \mathsf{E}_{\mathbf{t}}\left(\mathbf{Q}\mathbf{t}\right)\mathsf{E}_{\mathbf{t}}\left(\mathbf{Q}\mathbf{t}\right)^\top \\
&= \mathbf{Q}\left[\mathsf{E}_{\mathbf{t}}\left(\mathbf{t}\mathbf{t}^\top\right) - \mathsf{E}_{\mathbf{t}}\left(\mathbf{t}\right)\mathsf{E}_{\mathbf{t}}\left(\mathbf{t}^\top\right)\right]\mathbf{Q}^\top,
\end{aligned}$$

[2] In particular, it becomes possible to separate Gaussian processes when they are correlated over time.

[3] Our reasoning makes use of the population random variables, although in practice we use empirical estimates.

or

$$\mathbf{Q}^{-1}\mathbf{I}\left(\mathbf{Q}^{-1}\right)^{\top} = \mathbf{C}_{tt},$$

where $\mathbf{C}_{tt}$ is the covariance matrix of $\mathbf{t}$. Since $\mathbf{C}_{tt}$ is positive definite, it is real and symmetric. Thus it can be written in the form

$$\mathbf{C}_{tt} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{\top},$$

where $\mathbf{\Lambda}$ is a diagonal matrix of real values (the eigenvalues), and $\mathbf{S}$ contains the *orthonormal* eigenvectors (which means $\mathbf{S}^{-1} = \mathbf{S}^{\top}$). It can then be trivially seen that

$$\mathbf{Q}^{-1} = \mathbf{S}\mathbf{\Lambda}^{1/2},$$

and thus

$$\mathbf{Q} = \mathbf{\Lambda}^{-1/2}\mathbf{S}^{\top}.$$

The pre-whitening process described here is not a statistically efficient means of estimating $\mathbf{B}$ (see [20]), although in practice the performance penalty incurred by using pre-whitening is small.

2.2.2 Maximum likelihood

The computation of the remaining portion $\mathbf{W}$ of $\mathbf{V}$ must rely on the high order statistics of the output random vector x, since these statistics indicate when the components of x are pairwise independent. We use the term *contrast function* to denote both the expression that specifies the statistical dependencies between elements of x (which is therefore a function of the high order statistics), and the empirical estimate of this expression. We solve for $\mathbf{W}$ by optimising over the empirical contrast function. The idea of using the expectation of a nonlinear function to measure independence was first proposed by Jutten and Hérault (a summary of this early research is in [53]), although the choice of function is in this case less mathematically rigorous than the methods set out below.

The first approach we discuss for measuring independence is the *maximum likelihood* method [74]. This is equivalent to the Infomax algorithm [10], as described in [19]: the objective is to find the orthogonal matrix $\mathbf{W}$ that causes the distribution of x to most closely approach a certain model of f_{s}, written

$$\widehat{\mathsf{f}}_{\mathsf{s}}(\mathbf{s}) := \prod_{i=1}^{n} \widehat{\mathsf{f}}_{\mathsf{s}_i}(s_i). \tag{2.2.3}$$

The independent sources are therefore recovered by minimising the KL divergence $D_{\mathrm{KL}}\left(\mathsf{f}_{\mathsf{x}} || \widehat{\mathsf{f}}_{\mathsf{s}}\right) = D_{\mathrm{KL}}\left(\mathsf{f}_{\mathbf{W}\bar{\mathsf{t}}} || \widehat{\mathsf{f}}_{\mathsf{s}}\right)$ with respect to $\mathbf{W}$ (see Definition A.8.11 for the definition of the KL divergence, and Theorem A.8.12 for its properties). Equivalently, we can solve for $\mathbf{W}$ by maximising

$$-D_{\mathrm{KL}}\left(\mathsf{f}_{\mathbf{W}\bar{\mathsf{t}}} || \widehat{\mathsf{f}}_{\mathsf{s}}\right) = h\left(\left[\begin{array}{ccc} \widehat{\mathsf{F}}_{\mathsf{s}_1}\left(\mathbf{W}\tilde{\mathbf{t}}\right)_1 & \cdots & \widehat{\mathsf{F}}_{\mathsf{s}_n}\left(\mathbf{W}\tilde{\mathbf{t}}\right)_n \end{array}\right]\right), \tag{2.2.4}$$

where $\widehat{\mathsf{F}}_{\mathsf{s}_i}$ is the *model* C.D.F. of s_i, and $h(\cdot)$ is the differential entropy.

Of course, this leaves open the nature of the model $\widehat{\mathsf{f}}_{\mathsf{s}}$ to be chosen. One strategy is to use the property that the signals ought to become less Gaussian as the estimate of $\mathbf{W}$ yields increasingly independent x. In effect, we know from Theorem 2.1.3 that at most one component of s can be Gaussian, which makes non-Gaussianity a property of the remaining x by our problem definition. With this in mind, Bell and Sejnowski set the densities $\widehat{\mathsf{f}}_{\mathsf{s}_i}(s_i)$ to be derivatives of generalised logistic sigmoids, which corresponds to the assumption the elements of s are *super-Gaussian*. In practice, this particular choice of contrast is known to perform poorly for sub-Gaussian signals [10], since in this case stationary points of the contrast function can be found greater than those obtained for independent x. This is addressed in [36], in which a second contrast function is used for sub-Gaussian cases. The associated density models are therefore

$$\widehat{\mathsf{f}}_{\mathsf{s}_i}(s_i) \propto \begin{cases} \exp(-s_i^2/2)\mathrm{sech}^2(s_i) & \mathrm{super-Gaussian}, \\ \exp\left(-(s_i-1)^2/2\right) + \exp\left(-(s_i+1)^2/2\right) & \mathrm{sub-Gaussian}. \end{cases}$$

This model is applied in [59] to the separation of source signals and the removal of noise in EEG recordings.

We now show that this method corresponds to the maximum likelihood approach, as described for instance in [60, 73, 67] (a concise explanation of this link is in [19], although the connection was made independently in [60, 67]). We assume that m (whitened) observations $\tilde{\boldsymbol{t}}$ are generated from the *model* distribution[4] $\widehat{\mathbf{f}}_{\mathbf{W}^{-1}\mathbf{s}}$ in accordance with the model in (2.1.1) and (2.2.1)[5]. The expected log likelihood of the whitened random vector $\tilde{\mathbf{t}}$, which must be maximised with respect to $\mathbf{W}$, is

$$\begin{aligned} \mathbf{E}_{\tilde{\mathbf{t}}} \log\left(\widehat{\mathbf{f}}_{\tilde{\mathbf{t}}}\left(\tilde{\mathbf{t}}\right)\right) &= \mathbf{E}_{\tilde{\mathbf{t}}} \log\left(\widehat{\mathbf{f}}_{\mathbf{W}^{-1}\mathbf{s}}\left(\tilde{\mathbf{t}}\right)\right) && (2.2.5) \\ &= -D_{\mathrm{KL}}\left(\mathbf{f}_{\tilde{\mathbf{t}}} || \widehat{\mathbf{f}}_{\mathbf{W}^{-1}\mathbf{s}}\right) - h\left(\tilde{\mathbf{t}}\right). && (2.2.6) \end{aligned}$$

Only the first term need be considered when $\mathbf{W}$ changes, however. Comparing with (2.2.4), we note that $D_{\mathrm{KL}}\left(\mathbf{f}_{\tilde{\mathbf{t}}} || \widehat{\mathbf{f}}_{\mathbf{W}^{-1}\mathbf{s}}\right) = D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{W}\tilde{\mathbf{t}}} || \widehat{\mathbf{f}}_{\mathbf{s}}\right)$, which completes the proof.

In applying the method described above, we would like to know whether the proposed contrast function still works when the model (2.2.3) is incorrect. In fact, it has been shown independently in [73, 88] that the KL divergence exhibits a zero gradient with respect to $\mathbf{W}$ when the elements of $\mathbf{x} = \mathbf{W}\tilde{\mathbf{t}}$ are independent. A sufficient condition for the stability of this point (which ensures that it is not a saddle point, for instance), taken from [20], is

$$\mathbf{E}_{\mathsf{s}_i}\left(\varphi_i'\left(\mathsf{s}_i\right)\right)\mathbf{E}_{\mathsf{s}_i}\left(\mathsf{s}_i^2\right) - \mathbf{E}\left(\varphi_i\left(\mathsf{s}_i\right)\mathsf{s}_i\right) \geq 0, \tag{2.2.7}$$

where the *score functions* φ_i are defined as

$$\varphi_i := \frac{\widehat{\mathbf{f}}_{\mathsf{s}_i}'}{\widehat{\mathbf{f}}_{\mathsf{s}_i}}.$$

It is not guaranteed that the solution at independence is a *global* optimum; indeed, both [10] and [20] give specific instances in which errors in model formulation cause the global minimum of the contrast to be far from independence. Finally, we should expect to be penalised when our source density model is incorrect, given that the contrast is in practice computed using an empirical expectation with m samples. Thus, given a maximum likelihood estimate of $\mathbf{W}$, it is shown in [20] that the asymptotic ratio in output x_i of the *unwanted* signal power $P(\mathsf{s}_j)$ to the *desired* signal power $P(\mathsf{s}_i)$ is minimised when the source density model is correct. In [73], the effects of incorrect source density models are reduced using a parametric estimate for $\widehat{\mathbf{f}}_{\mathsf{s}_i}$, which is adapted according to the observations $\mathbf{t}$. Alternative methods for adapting the source density models include [54], which is designed to work well near zero kurtosis, and [41], which uses regularised splines to approximate the departure of the sources from Gaussianity.

2.2.3 KL divergence

We next introduce an alternative approach to determining independence, as summarised from [24, 42]. In this case, we use a criterion suggested directly by Theorem 2.1.2: namely, that when the pairwise independence of components of $\mathbf{x}$ is maximised (through adjustment of $\mathbf{W}$), we recover $\mathbf{s}$ up to the natural indeterminacies of permutation and scaling. Recall from (2.1.2) that the random variables $\mathbf{x}$ are mutually independent when the joint density can be written as the product of the marginals. Another way of looking at this is to say that the *KL divergence* between the joint density and the product of the marginals must be zero; in other words,

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}} \middle\| \prod_{i=1}^{n} \mathbf{f}_{\mathsf{x}_i}\right) = 0. \tag{2.2.8}$$

[4] Note that the model requires *both* an approximate density $\widehat{\mathbf{f}}_{\mathbf{s}}$ *and* an estimate $\mathbf{W}^{-1}$ of $\mathbf{QB}$.

[5] In other words, $\widehat{\mathbf{f}}_{\tilde{\mathbf{t}}}(\tilde{\mathbf{t}}) := \widehat{\mathbf{f}}_{\mathbf{W}^{-1}\mathbf{s}}\left(\mathbf{W}^{-1}\mathbf{s}\right) = \widehat{\mathbf{f}}_{\mathbf{s}}(\mathbf{s}) / \left|\det\left(\mathbf{W}^{-1}\right)\right|$.

This contrast was proposed in [24]. We now discuss the link between this method and the maximum likelihood algorithms introduced in the previous section, following [20]. We decompose the contrast in (2.2.4) as

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}}||\widehat{\mathbf{f}}_{\mathbf{s}}\right) = D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}}\left\|\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}\right.\right) + D_{\mathrm{KL}}\left(\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}\left\|\widehat{\mathbf{f}}_{\mathbf{s}}\right.\right).$$

Since both terms are *non-negative*, the maximum likelihood based score is minimised when $\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i} = \widehat{\mathbf{f}}_{\mathbf{s}}$ (i.e., when the model (2.2.3) is correct[6]) *and* $\mathbf{f}_{\mathbf{x}} = \prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}$ (as required by (2.2.8)). Thus, for a correct density model, the maximum likelihood and KL divergence based solutions coincide. An advantage of the KL divergence, however, is that it makes no assumptions regarding the source model $\mathbf{f}_{\mathbf{s}}$; on the other hand, there remains the problem of empirically estimating $\mathbf{f}_{\mathbf{x}}$ and its marginals.

We now describe methods used to compute the empirical estimate of (2.2.8), or at least that part of the criterion which varies with our choice of $\mathbf{W}$. We begin with the following simplification.

Theorem 2.2.1 (Change in KL divergence following instantaneous linear mixing). *Given random vectors $\mathbf{x}, \tilde{\mathbf{t}}$ related according to $\mathbf{x} = \mathbf{W}\tilde{\mathbf{t}}$, the KL divergence between $\mathbf{f}_{\mathbf{x}}$ and the product of its marginals can be written*

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}}\left\|\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}\right.\right) = \sum_{i=1}^{n} h\left(\mathsf{x}_i\right) - h\left(\tilde{\mathbf{t}}\right) - \log\left|\det\mathbf{W}\right|. \tag{2.2.9}$$

This result is proved using Theorem A.8.17. We recall from the previous section that $\mathbf{W}$ is orthogonal, and hence $\log\left|\det\mathbf{W}\right| = 0$. In addition, $h\left(\tilde{\mathbf{t}}\right)$ is invariant with respect to $\mathbf{W}$. It remains, therefore, to estimate the entropies $h\left(\mathsf{x}_i\right)$.

One method for computing the entropies is by using cumulant based expansions about the Gaussian density, as described in [21, 20]. Two methods for empirically estimating the entropies include the Edgeworth expansion [24] and the Gram-Charlier expansion [4]. Using the former expansion (Definition A.9.3), and subject to the constraint that $\mathbf{x}$ be whitened, the contrast can be approximated as

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}}\left\|\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}\right.\right) \approx -\frac{1}{48}\sum_{i=1}^{n}\kappa_4^2\left(\mathsf{x}_i\right) + C,$$

where C is a constant term with respect to $\mathbf{W}$, and the 4th order cumulant $\kappa_4\left(\mathsf{x}_i\right)$ is defined in Appendix A.9. We note that the stability analysis described at the end of the previous section can also applied to cumulant based contrasts [20]. A related contrast function is used in the Jade algorithm [21], which differs slightly from the above function to simplify the calculation of $\mathbf{W}$. The Gram-Charlier expansion yields

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}}\left\|\prod_{i=1}^{n}\mathbf{f}_{\mathsf{x}_i}\right.\right) \approx -\sum_{i=1}^{n}\left(\frac{\left(\kappa_3^i\right)^2}{3!2} + \frac{\left(\kappa_4^i\right)^2}{4!2} - \frac{5\left(\kappa_3^i\right)^2\kappa_4^i}{8} - \frac{\left(\kappa_4^i\right)^3}{16}\right) + C.$$

A non-cumulant based method for approximating the KL divergence is proposed in [47], which is of particular interest since it permits the design of a contrast that takes specific features of the source distributions into account (aside from cumulants), as well as providing an alternative link between the KL divergence based contrasts and those derived from maximum likelihood. In this case, the observation densities $\mathbf{f}_{\mathsf{x}_i}$ are modeled by densities $\widehat{\mathbf{f}}_{\mathsf{x}_i}$ of maximum entropy, subject to the constraints

$$\widehat{\mathbf{E}}_{\mathsf{x}_i}\left(g_l\left(\mathsf{x}_i\right)\right) = c_{l,i} = \mathbf{E}_{\mathsf{x}_i}\left(g_l\left(\mathsf{x}_i\right)\right) \tag{2.2.10}$$

for $l = 1 \ldots P$ and functions $g_l(\cdot)$ that describe certain known properties of our sources (here $\widehat{\mathbf{E}}_{\mathsf{x}_i}$ denotes the expectation with respect to $\widehat{\mathbf{f}}_{\mathsf{x}_i}$, and $\mathbf{E}_{\mathsf{x}_i}$ the expectation with respect to $\mathbf{f}_{\mathsf{x}_i}$). The right hand equality applies since the quantities $c_{l,i}$ are in practice estimated from samples drawn

[6] We gloss over the question of ambiguities in permutation and scaling, which can be dealt with for instance by using a convention in source ordering [73].

according to $\mathbf{f}_{\mathsf{x}_i}$. It is assumed in addition that the $\widehat{\mathbf{f}}_{\mathsf{x}_i}$ are close to Gaussian (this being the distribution of highest entropy for a given mean and variance), and that the $g_l(\cdot)$ are orthonormal with respect to the metric induced by the Gaussian, orthogonal to all quadratic polynomials, and do not rise faster than quadratically as a function of their argument. Under these circumstances, we may write

$$\widehat{\mathbf{f}}_{\mathsf{x}_i}(x) = (2\pi)^{-1/2} \exp\left(-x^2/2\right) \left(1 + \sum_{l=1}^{P} c_{l,i} g_l\left(x_i\right)\right),$$

in accordance with [25]. Then

$$h(\mathsf{x}_i) \quad \approx \quad h(\mathsf{x}_G) - \frac{1}{2}\sum_{l=1}^{P} c_{l,i}^2$$

where we make use of the orthogonality relations required of the $g_l(\cdot)$, and x_G is a Gaussian random variable with zero mean and unit covariance. A single such function is often used in ICA (i.e., $P = 1$), since this results in simple ICA algorithms. Given that we have some desirable property measured by some *even* function $f(\cdot)$ (more on these properties below), then minimising $h(\mathsf{x})$ can be shown to be equivalent to minimising

$$h(\mathsf{x}_i) \approx h(\mathsf{x}_G) - \gamma \left(\mathbf{E}_{\mathsf{x}_i}\left(f\left(\mathsf{x}_i\right)\right) - \mathbf{E}_{\mathsf{x}_G}\left(f\left(\mathsf{x}_G\right)\right)\right)^2,$$

where γ is a constant[7]. The minimisation of $h(\mathsf{x}_i)$ is thus achieved by maximising $\mathbf{E}_{\mathsf{x}_i}\left(f\left(\mathsf{x}_i\right)\right)$. If we write as $\mathbf{w}$ a particular row of $\mathbf{W}$, then replacing $x = \mathbf{w}\tilde{\mathbf{t}}$ and $f(x) := \log\left(\widehat{\mathbf{f}}_{\mathsf{s}}(\mathbf{w}\tilde{\mathbf{t}})\right)$ for each such row allows us to recover the maximum likelihood contrast (compare with (2.2.5)). It is shown in [50] that $\mathbf{E}_{\tilde{\mathbf{t}}}\left(f\left(\mathbf{w}\tilde{\mathbf{t}}\right)\right)$ has a local maximum or minimum when $\mathbf{w}$ is chosen so as to recover an independent component. Moreover, denoting by $\hat{\mathbf{w}}$ our estimate of this extremum for m observations, a proof is given in [46] that the trace of the asymptotic[8] covariance matrix of $\hat{\mathbf{w}}\sqrt{m}$ is minimised when $f = \log \mathbf{f}_{\mathsf{s}}$, which corresponds approximately to a minimum least squares estimate of $\mathbf{w}$. Given that we do not know the source densities, contrast functions are proposed in [46] that achieve good results in terms of robustness to outliers and small asymptotic covariance; these are

$$\begin{aligned} f_1(x_i) &= \tfrac{1}{a_1}\log\cosh(a_1 x_i), \\ f_2(x_i) &= -\tfrac{1}{a_2}\exp(-a_2 x_i^2/2), \\ f_3(x_i) &= \tfrac{1}{4}x_i^4, \end{aligned}$$

where $a_1 \geq 1$ and $a_2 \approx 1$. The first function is recommended for general use with super-Gaussian sources, and is robust to outliers; the second is still more robust, and is used in highly super-Gaussian situations; the third is equivalent to a kurtosis based contrast. A method for performing ICA quickly and robustly with these contrasts is given in [48].

2.2.4 Semi-parametric entropy estimates for KL divergence

The results in this section also describe ways to compute the entropies $h\left(\mathsf{x}_i\right)$, in the context of the KL divergence based contrast in (2.2.9). These methods are sufficiently different from the cumulant and nonlinearity based methods to merit a separate discussion, however, and represent promising directions in recent research. We begin with a kernel density estimate, which was proposed in the context of ICA in [69][9], and refined to decrease computational cost in [70]. Let us divide

[7] The operations required to modify an arbitrary even function $f(\cdot)$, so as to satisfy the orthogonality constraints required of $g(\cdot)$ in (2.2.10), are accomplished in the course of the derivation of this expression.

[8] In the sense that $m \to \infty$.

[9] A related method was rediscovered independently in [86], which like [69] uses the binning and FFT density estimation method described in [80].

the support of $\mathbf{f}_{\mathsf{x}}$ into a grid of size l_x with even spacing Δ_x, where l_x is assumed to be odd for notational convenience. The proposed approximation of the differential entropy[10] is

$$\begin{aligned}\widehat{h}(\mathsf{x}) &= \sum_{j=-\lfloor l_x/2 \rfloor}^{\lfloor l_x/2 \rfloor} \Delta_x \widehat{\mathbf{f}}_{\mathsf{x}}(j\Delta_x) \log\left(\Delta_x \widehat{\mathbf{f}}_{\mathsf{x}}(j\Delta_x)\right) \\ &= H(\hat{\mathsf{x}}) + \log \Delta_x,\end{aligned}$$

where $H(\hat{\mathsf{x}})$ is the entropy associated with the discretised approximation $\hat{\mathsf{x}}$ (with distribution $\mathbf{P}_{\hat{\mathsf{x}}}(j) = \widehat{\mathbf{f}}_{\mathsf{x}}(j\Delta_x)\,\Delta_x$) of the continuous random variable x, and $\widehat{\mathbf{f}}_{\mathsf{x}}(x)$ is the kernel density estimate of $\mathbf{f}_{\mathsf{x}}$. Given an i.i.d. sample $\boldsymbol{x} := (x_1, \ldots x_m)$ of size m from $\mathbf{f}_{\mathsf{x}}$, a simple kernel density estimate [80] is

$$\widehat{\mathbf{f}}_{\mathsf{x}}(x) = \frac{1}{m\sigma} \sum_{j=1}^{m} k\left(\frac{x - x_j}{\sigma}\right), \tag{2.2.11}$$

where σ determines the degree of smoothing, and $k(x)$ is a valid probability density (see also the more detailed discussion at the start of Section 4.2). Unfortunately, (2.2.11) does not lend itself to computationally efficient numerical integration: instead, we apply *binning*, the simplest version of which involves setting the effective number of samples at each grid point proportional to the number of observations that fall closer to it than to its neighbours. A more accurate estimate of the density (in the mean square sense) can be found by assigning weights to the grid points in the immediate vicinity of each observation, where the weights are a function of the distance to the observation in question. This turns out to be equivalent to replacing $k(x)$ in (2.2.11) with

$$\tilde{k}(x) := \sum_{l=0}^{L} \frac{\sigma c_l}{\Delta_x} S^r\left(\frac{x\sigma}{\Delta_x} + \frac{L+r}{2} - l\right),$$

where S^r are the cardinal splines of order r (simple binning corresponds to $r = 1$). The c_l are positive coefficients that sum to 1 and are symmetric about $L/2$, and are functions of the original kernel $k(x)$. In [70], however, a single coefficient $c_0 = 1$ is used in the above sum, and the kernel is in effect a spline kernel alone. We shall see in Section 4.2 that the method of Pham bears certain similarities to our approach.

An alternative estimate of entropy is proposed in [64], which does not require a density estimate as an intermediate step, but is based instead on order statistics. This estimate represents a modification to that proposed in [84], and the two are in fact asymptotically equivalent; both are, in addition, asymptotically efficient. Writing as $\breve{\boldsymbol{x}}$ the sample $\boldsymbol{x}$ with terms ordered from smallest to largest, the entropy can be approximated as

$$\widehat{h}(\mathsf{x}) = \frac{1}{m-l} \sum_{i=1}^{m-l} \log \frac{m+1}{l} (\breve{x}_{i+l} - \breve{x}_i),$$

where l must satisfy

$$\lim_{m\to\infty} \frac{l}{m} = 0$$

for the estimate to be consistent (in [64], $l = \sqrt{m}$ is used). The estimate is smoothed by augmenting the data with Gaussian clusters of points about each observation, and a grid search over the Jacobi angles parameterising $\mathbf{W}$ is used to find the global optimum. Results in [64] indicate that this estimate is highly resistant to outliers, and performs better than the KGV on many of the data sets described in [7]. On the other hand, performance of this method when the sources are near-Gaussian remains problematic.

[10] The relation below derives from Theorem A.8.8 in Appendix A.8.

Chapter 3

The kernel covariance and kernel canonical correlation

In this chapter, we focus on the formulation of measures of independence (that is, *contrast functions*) for two random variables. This reasoning uses well established principles, going back to [75], in which study a list of desirable properties was given for a measure of statistical dependence $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}})$ between random variables x, y. These include

1. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}})$ is defined for random variables x, y that are not constant with probability 1,
2. $0 \leq \mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) \leq 1$,
3. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) = 0$ if and only if x, y independent,
4. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) = 1$ if and only if $\mathsf{y} = f(\mathsf{x})$ or $\mathsf{x} = g(\mathsf{y})$, where f and g are Borel measurable functions.

It is shown in [75] that one measure satisfying these constraints is $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) = \sup_{f,g} \operatorname{corr}(f(\mathsf{x}), g(\mathsf{y}))$, where $f(\mathsf{x}), g(\mathsf{y})$ must have finite positive variance, and f, g are Borel measurable. This is similar to the kernel canonical correlation (KCC) introduced in [7], although we shall see that the latter is more restrictive in its choice of f, g. We propose a different measure, the *kernel covariance* (KC), which omits the second and fourth properties above; in the context of ICA, however, the first and third properties are adequate. Comparing with the contrast functions described in Section 2, the kernel methods in this section use a supremum over a function class, rather than a fixed nonlinearity: we shall see that this guarantees a *global* minimum of the contrast at independence, which in traditional methods is contingent on the accuracy of the source density model.

We begin in Section 3.1 with some useful definitions. In section 3.2, we introduce the normalised covariance, and demonstrate that this quantity is a measure of independence when computed in a RKHS. Several related approaches in the area of spectral methods and clustering are presented in Section 3.3. Finally, we introduce the canonical correlation, and the associated interpretation when this is accomplished in a RKHS, in Section 3.4.

3.1 Definitions

We begin by defining the terms and concepts needed to describe our contrast functions. Let $\vec{x}$ and $\vec{y}$ be vectors in $\mathcal{X}, \mathcal{Y}$ respectively, where $\mathcal{X}$ is a bounded subset in $\mathbb{R}^{n_x}$ and $\mathcal{Y}$ is a bounded subset in $\mathbb{R}^{n_y}$. Let $\vec{\mathsf{x}}$ and $\vec{\mathsf{y}}$ be *random* vectors in $\mathcal{X}$ and $\mathcal{Y}$. We define the vectors $\mathbf{x}$ and $\mathbf{y}$ and the

random vectors $\mathbf{x}$ and $\mathbf{y}$ in the feature spaces $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, and the mappings $\phi_x : \mathcal{X} \to \mathcal{F}_\mathcal{X}$ and $\phi_y : \mathcal{Y} \to \mathcal{F}_\mathcal{Y}$ such that

$$\mathbf{x} := \phi_x(\vec{x}) \quad \text{and} \quad \mathbf{y} := \phi_y(\vec{y}).$$

The feature spaces may be the reproducing kernel Hilbert spaces (and subspaces of ℓ_2^∞) associated with the Gaussian or Laplace kernels; we also consider feature spaces $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$ on occasion, in which case the feature and input spaces coincide. We define the variance and covariance matrices for $\mathbf{x}$ and $\mathbf{y}$ as

$$\mathbf{C}_{xy} := \mathbf{E}_{\mathbf{x},\mathbf{y}}\left((\mathbf{x} - \mathbf{E}_\mathbf{x}(\mathbf{x}))(\mathbf{y} - \mathbf{E}_\mathbf{y}(\mathbf{y}))^\top\right), \tag{3.1.1}$$

$$\mathbf{C}_{xx} := \mathbf{E}_\mathbf{x}\left((\mathbf{x} - \mathbf{E}_\mathbf{x}(\mathbf{x}))(\mathbf{x} - \mathbf{E}_\mathbf{x}(\mathbf{x}))^\top\right), \tag{3.1.2}$$

$$\mathbf{C}_{yy} := \mathbf{E}_\mathbf{y}\left((\mathbf{y} - \mathbf{E}_\mathbf{y}(\mathbf{y}))(\mathbf{y} - \mathbf{E}_\mathbf{y}(\mathbf{y}))^\top\right), \tag{3.1.3}$$

$$\mathbf{C} := \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{C}_{yy} \end{bmatrix}. \tag{3.1.4}$$

Assume we are given m i.i.d. samples of data; $\boldsymbol{z} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$, where $\mathbf{x}_i \in \mathcal{F}_\mathcal{X}$ and $\mathbf{y}_i \in \mathcal{F}_\mathcal{Y}$. Defining the matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_m \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_m \end{bmatrix}, \tag{3.1.5}$$

we obtain the *empirical estimates*

$$\mathbf{C}_{xy} = \frac{1}{m-1}\mathbf{X}\mathbf{H}\mathbf{Y}^\top \quad \mathbf{C}_{xx} = \frac{1}{m-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top \quad \mathbf{C}_{yy} = \frac{1}{m-1}\mathbf{Y}\mathbf{H}\mathbf{Y}^\top, \tag{3.1.6}$$

where

$$\mathbf{H} = \mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top, \tag{3.1.7}$$

and $\mathbf{1}_m$ is an $m \times 1$ vector of ones. These results are obtained by noting that multiplying $\mathbf{X}$ by $\mathbf{H}$ is equivalent to subtracting the column mean from each column of $\mathbf{X}$;

$$\begin{aligned} \mathbf{X}\mathbf{H} &= \mathbf{X}\left(\mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right) \\ &= \mathbf{X} - \frac{1}{m}\left(\sum_{i=1}^m \mathbf{x}_i\right)\mathbf{1}_m^\top \\ &= \begin{bmatrix} \mathbf{x}_1 - \frac{1}{m}\left(\sum_{i=1}^m \mathbf{x}_i\right) & \ldots & \mathbf{x}_m - \frac{1}{m}\left(\sum_{i=1}^m \mathbf{x}_i\right) \end{bmatrix}. \end{aligned}$$

We define the centered sample matrices as

$$\widetilde{\mathbf{X}} := \mathbf{X}\mathbf{H}, \qquad \widetilde{\mathbf{Y}} := \mathbf{Y}\mathbf{H}, \tag{3.1.8}$$

for greater conciseness. Writing $\mathbf{X}\mathbf{H}\mathbf{X}^\top = \mathbf{X}\mathbf{H}\mathbf{H}\mathbf{X}^\top = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top$ yields the empirical estimate for $\mathbf{C}_{xx}$ (see Appendix A.7.3 for the proof that $\mathbf{H}$ is idempotent). Another interpretation is that the *rows* of $\mathbf{X}\mathbf{H}$ are the projection of the rows of $\mathbf{X}$ onto the space perpendicular to $\mathbf{1}_m$. Thus, following the discussion of Appendix A.7.2 (in which it is proved that $\mathbf{H}$ has rank $m-1$), the matrix $\mathbf{X}\mathbf{H}$ can have rank *at most* $m-1$.

Finally, we define the Gram matrices of inner products between the centered observations above, in the case where $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces with associated kernels[1]

[1] The argument of the kernel specifies whether the kernel pertains to $\mathcal{F}_\mathcal{X}$ or $\mathcal{F}_\mathcal{Y}$.

$k(\vec{x}_i - \vec{x}_j) := \mathbf{x}_i^\top \mathbf{x}_j$ and $k(\vec{y}_i - \vec{y}_j) := \mathbf{y}_i^\top \mathbf{y}_j$. Beginning with the *uncentered* Gram matrices[2],

$$\mathbf{K}_{mm}^{(x)} := \begin{bmatrix} k(\vec{x}_1 - \vec{x}_1) & \dots & k(\vec{x}_1 - \vec{x}_m) \\ \vdots & \ddots & \vdots \\ k(\vec{x}_m - \vec{x}_1) & \dots & k(\vec{x}_m - \vec{x}_m) \end{bmatrix} = \mathbf{X}^\top \mathbf{X}$$

$$\mathbf{K}_{mm}^{(y)} := \begin{bmatrix} k(\vec{y}_1 - \vec{y}_1) & \dots & k(\vec{y}_1 - \vec{y}_m) \\ \vdots & \ddots & \vdots \\ k(\vec{y}_m - \vec{y}_1) & \dots & k(\vec{y}_m - \vec{y}_m) \end{bmatrix} = \mathbf{Y}^\top \mathbf{Y},$$

then Gram matrices for the centered variables are

$$\widetilde{\mathbf{K}}_{mm}^{(x)} := \mathbf{H}\mathbf{K}_{mm}^{(x)}\mathbf{H} = \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}, \quad \widetilde{\mathbf{K}}_{mm}^{(y)} := \mathbf{H}\mathbf{K}_{mm}^{(y)}\mathbf{H} = \widetilde{\mathbf{Y}}^\top \widetilde{\mathbf{Y}}.$$

This result is taken from [78].

3.2 The normalised covariance and kernel covariance

In this section, we define the normalised covariance, and describe the properties of the kernelised version. This quantity, in non-kernelised form, has been widely used in the partial least squares method for regression [87], which is applied to problems in which the dimension of the regressors $\vec{x}_i$ is greater than their number m, and the regressors are highly multicollinear (this situation is often encountered in chemometrics). A kernelised version of the (highly robust) nonlinear iterative partial least squares algorithm was derived in [76], although in this case only the variables in the space $\mathcal{X}$ of regressors are mapped to an RKHS $\mathcal{F}_\mathcal{X}$, whereas in this study we also map the outputs in $\mathcal{Y}$ to $\mathcal{F}_\mathcal{Y}$[3]. It is noted in [76] that the feature spaces commonly used in kernel methods are of high dimension, and the large ratio of the maximum to minimum eigenvalues in the associated Gram matrices imply the feature space representations of the observations $\boldsymbol{x}$ to be highly collinear. Thus, the partial least squares algorithm is well suited to being kernelised.

We begin by giving a general description of the normalised covariance, which applies whether or not the feature spaces $\mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces. Let $\mathbf{x}$ and $\mathbf{y}$ be random vectors in $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, as defined in the previous section. We wish to find vectors $\boldsymbol{\alpha}_i \in \mathcal{F}_\mathcal{X} : \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_i \leq 1$ and $\boldsymbol{\beta}_i \in \mathcal{F}_\mathcal{Y} : \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_i \leq 1$ onto which $\mathbf{x}$ and $\mathbf{y}$ respectively project, such that the covariance γ_i between these projections is a stationary point with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$. The following theorem yields an equivalent eigenvalue problem.

Theorem 3.2.1 (Stationary points of the normalised covariance). *The stationary points γ_i with respect to $\boldsymbol{\alpha}_i \boldsymbol{\beta}_i$ of*

$$\operatorname{cov}\left(\boldsymbol{\alpha}^\top \mathbf{x}, \boldsymbol{\beta}^\top \mathbf{y}\right) : \|\boldsymbol{\alpha}\|_{\mathcal{F}_\mathcal{X}} \leq 1, \|\boldsymbol{\beta}\|_{\mathcal{F}_\mathcal{Y}} \leq 1 \tag{3.2.1}$$

are given by the solutions to the eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{3.2.2}$$

The proof may be found in Appendix A.2.1. The eigenvalues are in pairs $\pm\gamma_i$, since for each solution $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ with eigenvalue γ_i there is a solution $\boldsymbol{\alpha}_i, -\boldsymbol{\beta}_i$ with eigenvalue $-\gamma_i$. We replace $\mathbf{C}_{xy}$ with its empirical estimate $\mathbf{X}\mathbf{H}\mathbf{Y}^\top$ to obtain an empirical estimate of the stationary points;

$$\begin{bmatrix} \widetilde{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{Y}} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_i \\ \widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{X}\mathbf{H}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{H}\mathbf{X}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} \tag{3.2.3}$$

[2] The subscripts associated with the Gram matrices may at this point seem redundant, however they are required to distinguish between different Gram matrices in later sections.

[3] Of course, the use of the terms "input" and "output" are in our case meaningless, since we require only the dependency between the two quantities.

It is clear from the above expression that the eigenvectors for non-zero eigenvalues γ_i can be written as linear combinations of the centered observations. We define the coefficient vectors $\mathbf{c}_i, \mathbf{d}_i$ of these linear combinations as

$$\boldsymbol{\alpha}_i = \widetilde{\mathbf{X}}\mathbf{c}_i, \quad \boldsymbol{\beta}_i = \widetilde{\mathbf{Y}}\mathbf{d}_i. \tag{3.2.4}$$

(this argument can be thought of as a specific instance of the representer theorem; see Schölkopf *et al.* [77]).

In the remainder of this section, we consider *only* the case where $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces, with the intent of demonstrating that the *maximum* eigenvalue in (3.2.2) may be used as a measure of independence. We make this replacement in (3.2.3), and premultiply both sides with the matrix $\begin{bmatrix} \widetilde{\mathbf{X}}^\top & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{Y}}^\top \end{bmatrix}$ (this does not alter the non-zero eigenvalues, in the light of (3.2.4)), to obtain the equivalent *generalised* eigenvalue problem in terms of the centered Gram matrices;

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}^{(x)}_{mm}\widetilde{\mathbf{K}}^{(y)}_{mm} \\ \widetilde{\mathbf{K}}^{(y)}_{mm}\widetilde{\mathbf{K}}^{(x)}_{mm} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \widetilde{\mathbf{K}}^{(x)}_{mm} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}^{(y)}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} \tag{3.2.5}$$

The contrast function associated with this maximum eigenvalue is explicitly defined below.

Definition 3.2.2 (Kernel covariance). Let the feature spaces $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ be reproducing kernel Hilbert spaces with associated kernels $k(\vec{x}_i - \vec{x}_j)$ and $k(\vec{y}_i - \vec{y}_j)$. The kernel covariance is defined as

$$\mathcal{I}(\mathsf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}) = \sup_{f \in \widetilde{\mathcal{F}}_\mathcal{X},\, g \in \widetilde{\mathcal{F}}_\mathcal{Y}} \left| \mathsf{E}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}\left[f(\vec{\mathsf{x}})g(\vec{\mathsf{y}})\right] - \mathsf{E}_{\vec{\mathsf{x}}}\left[f(\vec{\mathsf{x}})\right]\mathsf{E}_{\vec{\mathsf{y}}}\left[g(\vec{\mathsf{y}})\right] \right|,$$

where $\widetilde{\mathcal{F}}_\mathcal{X} := \left\{ f \in \mathcal{F}_\mathcal{X} \,:\, \|f\|_{\mathcal{F}_\mathcal{X}} \leq 1 \right\}$, and $\widetilde{\mathcal{F}}_\mathcal{Y} := \left\{ g \in \mathcal{F}_\mathcal{Y} \,:\, \|g\|_{\mathcal{F}_\mathcal{Y}} \leq 1 \right\}$.

The empirical estimate of this quantity follows simply from the definition.

Definition 3.2.3 (Empirical kernel covariance). Given a training sample $\boldsymbol{x}, \boldsymbol{y}$ drawn independently and identically according to $\mathsf{P}_{\mathsf{x},\mathsf{y}}$, the empirical estimate of the kernel covariance is given by

$$\mathcal{I}_{\text{emp}}(\boldsymbol{x}, \boldsymbol{y}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}) = \sup_{f \in \widetilde{\mathcal{F}}_\mathcal{X},\, g \in \widetilde{\mathcal{F}}_\mathcal{Y}} \frac{1}{m-1} \left| \sum_{l=1}^{m} f(\vec{x}_l) g(\vec{y}_l) - \frac{1}{m} \left(\sum_{l=1}^{m} f(\vec{x}_l) \right) \left(\sum_{l=1}^{m} g(\vec{y}_l) \right) \right|.$$

In particular, if we replace

$$f(\vec{x}) = \sum_{l=1}^{m} c_l k(\vec{x} - \vec{x}_l) = \sum_{l=1}^{m} c_l \mathbf{x}^\top \mathbf{x}_l, \tag{3.2.6}$$

$$g(\vec{y}) = \sum_{l=1}^{m} d_l k(\vec{y} - \vec{y}_l) = \sum_{l=1}^{m} d_l \mathbf{y}^\top \mathbf{y}_l, \tag{3.2.7}$$

we observe that the kernel covariance is simply the maximum stationary point with respect to $\mathbf{c}, \mathbf{d}$ of the normalised covariance in Theorem 3.2.1.

We now demonstrate the link between the kernel covariance and independence.

Theorem 3.2.4 (Kernel covariance and independence). *Let $f \in \widetilde{\mathcal{F}}$ and $g \in \widetilde{\mathcal{G}}$, where $\widetilde{\mathcal{F}}, \widetilde{\mathcal{G}}$ are the absolutely bounded, continuous functions on the respective bounded sets $\mathcal{X} \subset \mathbb{R}^{n_x}, \mathcal{Y} \subset \mathbb{R}^{n_y}$. Then the kernel correlation $\mathcal{I}(\mathsf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}) = 0$ if and only if $\vec{\mathsf{x}}, \vec{\mathsf{y}}$ are independent.*

The proof below is a classical result [32, 75].

Proof. We first show that $\mathcal{I}\left(\mathbf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}\right) = 0$ if $\vec{\mathsf{x}}, \vec{\mathsf{y}}$ are independent. This is quite simple;

$$\begin{aligned}\mathcal{I}\left(\mathbf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}\right) &:= \sup_{f\in\widetilde{\mathcal{F}}_{\mathcal{X}},\, g\in\widetilde{\mathcal{F}}_{\mathcal{Y}}} \left|\mathbf{E}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}\left[f(\vec{\mathsf{x}})g(\vec{\mathsf{y}})\right] - \mathbf{E}_{\vec{\mathsf{x}}}\left[f(\vec{\mathsf{x}})\right]\mathbf{E}_{\vec{\mathsf{y}}}\left[g(\vec{\mathsf{y}})\right]\right| \\ &= \sup_{f\in\widetilde{\mathcal{F}}_{\mathcal{X}},\, g\in\widetilde{\mathcal{F}}_{\mathcal{Y}}} \left|\mathbf{E}_{\vec{\mathsf{x}}}\left[f(\vec{\mathsf{x}})\right]\mathbf{E}_{\vec{\mathsf{y}}}\left[g(\vec{\mathsf{y}})\right] - \mathbf{E}_{\vec{\mathsf{x}}}\left[f(\vec{\mathsf{x}})\right]\mathbf{E}_{\vec{\mathsf{y}}}\left[g(\vec{\mathsf{y}})\right]\right| \\ &= 0,\end{aligned}$$

where the second step makes use of the independence of $\vec{\mathsf{x}}, \vec{\mathsf{y}}$. We now prove the converse. To simplify the discussion, we describe only the case $n_x, n_y = 1$, and write x, y without the vector notation. Let $[q_1, q_2] \subseteq \mathcal{X}$ and $[r_1, r_2] \subseteq \mathcal{Y}$, on which the strictly positive functions $u(x) \in \widetilde{\mathcal{F}}_{\mathcal{X}}$ and $v(y) \in \widetilde{\mathcal{F}}_{\mathcal{Y}}$ are compactly supported. In this case, $u^{1/l}(x) \in \widetilde{\mathcal{F}}_{\mathcal{X}}$ and $v^{1/l}(y) \in \widetilde{\mathcal{F}}_{\mathcal{Y}}$ for $l \geq 1$. Using the limits

$$\lim_{l\to\infty} u^{1/l}(x) = \mathbb{I}_{x\in[q_1,q_2]} \qquad \text{and} \qquad \lim_{l\to\infty} v^{1/l}(y) = \mathbb{I}_{y\in[r_1,r_2]},$$

then if the *supremum* of the covariance over all functions in $\widetilde{\mathcal{F}}_{\mathcal{X}}, \widetilde{\mathcal{F}}_{\mathcal{Y}}$ is zero, it follows that

$$\lim_{l\to\infty} \left|\mathbf{E}_{\mathsf{x},\mathsf{y}}\left[u^{1/l}(\mathsf{x})v^{1/l}(\mathsf{y})\right] - \mathbf{E}_{\mathsf{x}}\left[u^{1/l}(\mathsf{x})\right]\mathbf{E}_{\mathsf{y}}\left[v^{1/l}(\mathsf{y})\right]\right| = 0,$$

and thus

$$\mathbf{P}_{\mathsf{x},\mathsf{y}}\left([q_1,q_2],[r_1,r_2]\right) = \mathbf{P}_{\mathsf{x}}\left([q_1,q_2]\right)\mathbf{P}_{\mathsf{y}}\left([r_1,r_2]\right).$$

Since the σ-algebra over $\{[q_1, q_2] \times [r_1, r_2] : q_1, q_2 \in \mathcal{X},\, r_1, r_2 \in \mathcal{Y}\}$ constitute the Borel sets over $\mathcal{X} \times \mathcal{Y}$, $\mathcal{I}\left(\mathbf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}\right) = 0$ implies x, y are independent. □

We end by noting that the kernel covariance is generalised to more than two random variables in Section 4.3. This more general expression is zero if and only if the random variables are *pairwise* independent.

3.3 Concepts similar to the kernel covariance

The kernel covariance turns out to be similar in certain respects to a number of kernel algorithms, for an appropriate choice of the spaces $\mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}$. By contrast with ICA, however, these methods seek to *maximise* the kernel covariance through the correct choice of feature space elements. To describe these similarities, we rewrite (3.2.5) in a more suitable form.

Lemma 3.3.1 (Kernel covariance: alternative form). *The* square *of the empirical kernel covariance (Definition 3.2.3) can be written*

$$\mathcal{I}^2_{\text{emp}}\left(\boldsymbol{x}, \boldsymbol{y}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}\right) = \left\|\widetilde{\mathbf{K}}^{(x)}_{mm}\widetilde{\mathbf{K}}^{(y)}_{mm}\right\|,$$

where the matrix norm is the maximum eigenvalue.

The proof may be found in Appendix A.2.2. We now describe two examples which are demonstrated to perform optimisation using this contrast.

The first algorithm related to the kernel covariance is kernel principal component analysis (kPCA) [78, 79]. Given a data sample $\boldsymbol{x}$ of size m, kPCA entails diagonalising the covariance matrix $\mathbf{C}_{xx}$ *in the feature space* $\mathcal{F}_{\mathcal{X}}$. In other words, we solve

$$\mathbf{C}_{xx}\mathcal{A} = \mathcal{A}\boldsymbol{\Lambda},$$

where $\mathcal{A}$ contains the eigenvectors $\boldsymbol{\alpha}_i$ in its columns, and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Given the empirical estimate of the covariance matrix in (3.1.6) and (3.1.5), this is written[4]

$$\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathcal{A} = \mathcal{A}\boldsymbol{\Lambda}.$$

[4] The $(m-1)^{-1}$ factor in the empirical covariance is absorbed into the eigenvalues.

Each eigenvector $\boldsymbol{\alpha}_i$ with non-zero eigenvalue λ_i can be written as a linear combination of the centered observations;

$$\boldsymbol{\alpha}_i = \mathbf{X}\mathbf{H}\mathbf{c}_i. \tag{3.3.1}$$

The eigenvalue problem is therefore rewritten

$$\mathbf{H}\mathbf{K}_{mm}^{(x)}\mathbf{H}\mathbf{c}_i = \lambda_i \mathbf{H}\mathbf{c}_i$$

for each entry λ_i in the diagonal of $\boldsymbol{\Lambda}$, or equivalently $\mathbf{c}_i^\top \mathbf{H}\mathbf{K}_{mm}^{(x)}\mathbf{H}\mathbf{c}_i = \lambda_i$, where we know from (3.3.1) that $\mathbf{H}\mathbf{c}_i = \mathbf{c}_i$ for non-zero λ_i. We now define our second feature space $\mathcal{F}_\mathcal{Y} := \mathbb{R}$, with kernel $k(y_k, y_l) = y_k y_l$. If we consider the maximum eigenvalue $\lambda_{\max}$ with associated eigenvector $\mathbf{c}$, we obtain

$$\begin{aligned}\lambda_{\max} = \mathbf{c}^\top \mathbf{H}\mathbf{K}_{mm}^{(x)}\mathbf{H}\mathbf{c} &= \left\| \widetilde{\mathbf{K}}_{mm}^{(x)}\mathbf{H}\mathbf{c}\mathbf{c}^\top\mathbf{H} \right\| \\ &= \left\| \widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \right\|,\end{aligned}$$

where the uncentered Gram matrix in $\mathcal{F}_\mathcal{Y}$ is $\mathbf{K}_{mm}^{(y)} := \mathbf{c}\mathbf{c}^\top$. In other words, the largest eigenvalue obtained in kernel PCA is found by *maximising the kernel covariance* through the optimal choice of m elements $\mathbf{c}_i$ in $\mathcal{F}_\mathcal{Y}$, with the constraint $\|\mathbf{c}_i\| \leq 1$ (an inequality is used to keep the constraint set convex).

The second algorithm that relates to the kernel covariance is part of the spectral clustering/kernel target alignment framework [26, 27]. Consider the case where we observe a data sample $\boldsymbol{x}$ of size m, with a matrix of distances between these observations defined via a kernel as $\widetilde{\mathbf{K}}_{mm}^{(x)}$. In the case of spectral clustering, we wish to assign each point in $\boldsymbol{x}$ to one of two classes, such that points in a given class are similar to each other (as defined by the kernel), and different to points in the opposite class. Writing as $\mathbf{c}$ the vector containing the m proposed labels $c_j \in \{-1, 1\} : j \in \{1, \ldots, m\}$ for each point in $\boldsymbol{x}$, we would like to choose $\mathbf{c}$ to ensure $\widetilde{\mathbf{K}}_{mm}^{(x)}$ and $\mathbf{c}\mathbf{c}^\top$ are similar. One measure of similarity, the *alignment*, is written

$$\begin{aligned}\mathcal{Q}(\widetilde{\mathbf{K}}_{mm}^{(x)}, \mathbf{c}\mathbf{c}^\top) &:= \frac{\left\langle \widetilde{\mathbf{K}}_{mm}^{(x)}, \mathbf{c}\mathbf{c}^\top \right\rangle}{\sqrt{\left\langle \widetilde{\mathbf{K}}_{mm}^{(x)}, \widetilde{\mathbf{K}}_{mm}^{(x)} \right\rangle \left\langle \mathbf{c}\mathbf{c}^\top, \mathbf{c}\mathbf{c}^\top \right\rangle}} \\ &= \frac{\left\langle \widetilde{\mathbf{K}}_{mm}^{(x)}, \mathbf{c}\mathbf{c}^\top \right\rangle}{m\sqrt{\left\langle \widetilde{\mathbf{K}}_{mm}^{(x)}, \widetilde{\mathbf{K}}_{mm}^{(x)} \right\rangle}},\end{aligned}$$

where the inner product used is given in Definition A.1.5 (the alignment is simply the angle between two matrices: see Definition A.1.7). The numerator can be rewritten $\left\langle \widetilde{\mathbf{K}}_{mm}^{(x)}, \mathbf{c}\mathbf{c}^\top \right\rangle = \mathbf{c}^\top \widetilde{\mathbf{K}}_{mm}^{(x)}, \mathbf{c}$, while the denominator is constant with respect to $\mathbf{c}$. Since the optimisation over the discrete labels is intractable, a relaxed problem is solved instead: we replace $\mathbf{c}$ with $\tilde{\mathbf{c}} \in \mathbb{R}^m$ such that $\|\tilde{\mathbf{c}}/m\| \leq 1$, and find the $\tilde{\mathbf{c}}$ that yields the maximum eigenvalue of $\tilde{\mathbf{c}}^\top \widetilde{\mathbf{K}}_{mm}^{(x)}, \tilde{\mathbf{c}}$. Thus, we return to the kernel PCA framework described earlier. The signs of the elements of $\tilde{\mathbf{c}}$ provide an estimate of $\mathbf{c}$, bearing in mind that an offset should be added to $\tilde{\mathbf{c}}$ to account for imbalances in the proportion of points in each class. Alternatively, we might be given the labels $\mathbf{c}$, and wish to build up a kernel $\widetilde{\mathbf{K}}_{mm}^{(x)}$ to maximise the alignment; further detail regarding this approach may be found in [26].

3.4 The canonical correlation

The results in this section are taken from [37, 55, 7]. In particular, a much more extensive discussion of the properties of canonical correlation analysis and its kernelisation may be found in [55] and

[7], and this section simply summarises the properties and derivations relevant to our requirements for ICA. We again start by defining the canonical correlation in the general case, without reference to its interpretation when $\mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces. The random vectors x and y in $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are defined in Section 3.1, as are the related variance and covariance matrices. For a set of indices $i \in (1, \ldots, s)$ (the determination of s will be addressed shortly), we would like to find vectors $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ onto which x and y respectively project, such that the *correlation* ρ_i between these projections is a stationary point with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$. A related interpretation is that we are attempting to find the linear combination of the elements of x that is the best "predictor", and the linear combination of the y that is the most "predictable", subject to being uncorrelated with other such linear combinations.

We define the *linear variates* $\mathsf{a}_i, \mathsf{b}_i$ as respective linear combinations of the random variables x and y;

$$\mathsf{a}_i = \boldsymbol{\alpha}_i^\top \left(\mathsf{x} - \mathbf{E}_\mathsf{x}\left(\mathsf{x}\right)\right), \quad \mathsf{b}_i = \boldsymbol{\beta}_i^\top \left(\mathsf{y} - \mathbf{E}_\mathsf{y}\left(\mathsf{y}\right)\right). \tag{3.4.1}$$

The correlation between a_i and b_i is

$$\rho_i = \operatorname{corr}\left(\mathsf{a}_i, \mathsf{b}_i\right) := \frac{\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathsf{a}_i \mathsf{b}_i\right)}{\sqrt{\mathbf{E}_\mathsf{x}\left(\mathsf{a}_i^2\right)\mathbf{E}_\mathsf{y}\left(\mathsf{b}_i^2\right)}} \tag{3.4.2}$$

$$= \frac{\boldsymbol{\alpha}_i^\top \mathbf{C}_{xy} \boldsymbol{\beta}_i}{\sqrt{\left(\boldsymbol{\alpha}_i^\top \mathbf{C}_{xx} \boldsymbol{\alpha}_i\right)\left(\boldsymbol{\beta}_i^\top \mathbf{C}_{yy} \boldsymbol{\beta}\right)}}. \tag{3.4.3}$$

These definitions lead to the following theorem.

Theorem 3.4.1 (Stationary points of the correlation between projections). *The stationary points of (3.4.3) with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are known as the* canonical correlations, *and are given by the solutions to the generalised eigenvector problem*

$$\begin{bmatrix} 0 & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{3.4.4}$$

An alternative form is

$$\mathbf{C} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = (1 + \rho_i) \begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{3.4.5}$$

The proof may be found in Appendix A.3.1. We next describe the correlation relations between the canonical variates. This result is proved in Appendix A.3.2.

Theorem 3.4.2 (Correlation relations between canonical variates). *Given $\mathbf{C}_{xx}, \mathbf{C}_{yy}$ have full rank, and writing*

$$\mathcal{A} := \begin{bmatrix} \boldsymbol{\alpha}_1 & \ldots & \boldsymbol{\alpha}_s \end{bmatrix} \quad \mathcal{B} := \begin{bmatrix} \boldsymbol{\beta}_1 & \ldots & \boldsymbol{\beta}_s \end{bmatrix},$$

then

$$\mathcal{A}^\top \mathbf{C}_{xx} \mathcal{A} = \mathbf{I}, \quad \mathcal{B}^\top \mathbf{C}_{yy} \mathcal{B} = \mathbf{I}, \quad \mathcal{A}^\top \mathbf{C}_{xy} \mathcal{B} = \operatorname{diag}\left(\begin{bmatrix} \rho_1 & \ldots & \rho_s \end{bmatrix}\right),$$

where $s = 2\min\{n_x, n_y\}$ when $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ coincide with $\mathcal{X}$ and $\mathcal{Y}$, and the solutions are in pairs $\pm\rho_i$. Note that the linear variates $\mathsf{a}_i, \mathsf{b}_i$ and $\mathsf{a}_j, \mathsf{b}_j$ corresponding to different roots $\rho_i, \rho_j : j \neq i$ of (3.4.3) are uncorrelated. The first canonical correlation *is that for which ρ_i is largest.*

Replacing $\mathbf{C}_{xy} = \mathbf{X}\mathbf{H}\mathbf{Y}^\top$, $\mathbf{C}_{xx} = \mathbf{X}\mathbf{H}\mathbf{X}^\top$, and $\mathbf{C}_{yy} = \mathbf{Y}\mathbf{H}\mathbf{Y}^\top$, it is again clear that the projection directions for non-zero ρ_i can be written as linear combinations of the centered observations;

$$\boldsymbol{\alpha}_i = \mathbf{X}\mathbf{H}\mathbf{c}_i, \quad \boldsymbol{\beta}_i = \mathbf{Y}\mathbf{H}\mathbf{d}_i.$$

We next introduce a theorem taken from [37], which will be useful when describing the behaviour of the canonical variates in high dimensional spaces.

Theorem 3.4.3 (Geometric interpretation of canonical variates). *Let the sample matrices* $\mathbf{X}, \mathbf{Y}$ *be defined as in (3.1.5), and let the vectors* $\mathbf{a}_i$, $\mathbf{b}_i$ *be the* m *empirical estimates of the linear variates in (3.4.1) for this sample. Then* $\rho_i \mathbf{a}_i$ *is the projection of* $\mathbf{b}_i$ *onto the column space of* $\widetilde{\mathbf{X}}^\top$*, and* $\rho_i \mathbf{b}_i$ *is the projection of* $\mathbf{a}_i$ *onto the column space of* $\widetilde{\mathbf{Y}}^\top$*, where* $\widetilde{\mathbf{X}}$ *and* $\widetilde{\mathbf{Y}}$ *represent the centered sample matrices, as defined in (3.1.8).*

The proof of this theorem is in Appendix A.3.3.

We now discuss the properties of the canonical correlation when $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces. This kernelisation has been investigated by [7, 83, 55, 56, 62]. Following the procedure used to derive (3.2.5) in Section 3.2, we obtain an expression for the *kernel* canonical correlations (KCC) in terms of the Gram matrices;

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}^{(x)}_{mm}\widetilde{\mathbf{K}}^{(y)}_{mm} \\ \widetilde{\mathbf{K}}^{(y)}_{mm}\widetilde{\mathbf{K}}^{(x)}_{mm} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \left(\widetilde{\mathbf{K}}^{(x)}_{mm}\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(\widetilde{\mathbf{K}}^{(y)}_{mm}\right)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} \tag{3.4.6}$$

By analogy with Definition 3.2.2, we can define a candidate contrast function on the basis of (3.4.4).

Definition 3.4.4 (First kernel canonical correlation). Let the feature spaces $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ be reproducing kernel Hilbert spaces with associated kernels $k(\vec{x}_i - \vec{x}_j)$ and $k(\vec{y}_i - \vec{y}_j)$. The *first* kernel canonical correlation is defined as

$$\begin{aligned} \mathcal{J}\left(\mathbf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}\right) &= \sup_{f\in\mathcal{F}_\mathcal{X},\, g\in\mathcal{F}_\mathcal{Y}} \operatorname{corr}\left(f(\vec{\mathsf{x}}), g(\vec{\mathsf{y}})\right)) \\ &= \sup_{f\in\mathcal{F}_\mathcal{X},\, g\in\mathcal{F}_\mathcal{Y}} \frac{\mathsf{E}\left(f(\vec{\mathsf{x}})g(\vec{\mathsf{y}})\right) - \mathsf{E}_{\vec{\mathsf{x}}}\left(f(\vec{\mathsf{x}})\right)\mathsf{E}_{\vec{\mathsf{y}}}\left(g(\vec{\mathsf{y}})\right)}{\sqrt{\mathsf{E}_{\vec{\mathsf{x}}}\left(f^2(\vec{\mathsf{x}})\right) - \mathsf{E}^2_{\vec{\mathsf{x}}}\left(f(\vec{\mathsf{x}})\right)}\sqrt{\mathsf{E}_{\vec{\mathsf{y}}}\left(g^2(\vec{\mathsf{y}})\right) - \mathsf{E}^2_{\vec{\mathsf{y}}}\left(g(\vec{\mathsf{y}})\right)}}. \end{aligned}$$

As in the case of the kernel covariance, we may specify an empirical estimate $\mathcal{J}_{\text{emp}}\left(\boldsymbol{x}, \boldsymbol{y}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}\right) := \max_i(\rho_i)$ similar to that in Definition 3.2.3, and use (3.2.6) and (3.2.7) to recover (3.4.6). It is pointed out in [7] that the first canonical correlation is very similar to the function maximised by the *alternating conditional expectation* algorithm in [15], although in the latter case $f(\vec{x})$ may be replaced with a linear combination of several functions of $\vec{x}$.

We note that the numerator of the contrast in Definition 3.4.4 is the same as that in Definition 3.2.2, which suggests that the first kernel canonical correlation might also be a useful contrast function; this was proposed by Bach and Jordan [7]. A problem with using the kernel canonical correlation in the form described above is discussed in both [55] and [7]; we now describe this problem, and the two main ways in which it has been solved.

Lemma 3.4.5 (Without regularisation, the kernel canonical correlation is independent of the data). *Suppose that the Gram matrices* $\mathbf{K}^{(x)}_{mm}, \mathbf{K}^{(y)}_{mm}$ *have full rank. The* $2(m-1)$ *non-zero solutions to (3.4.6) are then* $\rho_i = \pm 1$*, regardless of* $\boldsymbol{z}$*.*

The proof may be found in Appendix A.6.1. This result does not come as a surprise, given Theorem 3.4.3: when the respective dimensions of $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are much greater than m (which is the case when they are RKHSs associated with the Gaussian or Laplace kernel) then the rank of both $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ is $m-1$. In other words, the columns of $\widetilde{\mathbf{X}}^\top$ and $\widetilde{\mathbf{Y}}^\top$ both span $\mathbb{R}^m \setminus \{\mathbf{1}_m\}$, and the projection of $\mathbf{b}_i$ onto the column space of $\widetilde{\mathbf{X}}^\top$ (and of $\mathbf{a}_i$ onto the column space of $\widetilde{\mathbf{Y}}^\top$) gives $\mathbf{b}_i$ (resp. $\mathbf{a}_i$) exactly for all i.

This argument is used in [7] to justify a regularised contrast function,

$$\mathcal{J}_\kappa\left(\mathbf{P}_{\vec{\mathsf{x}},\vec{\mathsf{y}}}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}\right) := \sup_{f\in\mathcal{F}_\mathcal{X},\, g\in\mathcal{F}_\mathcal{Y}} \frac{\operatorname{cov}\left(f(\vec{\mathsf{x}}), g(\vec{\mathsf{y}})\right)}{\left(\operatorname{var}\left(f\left(\vec{\mathsf{x}}\right)\right) + \kappa\left\|f\right\|^2_{\mathcal{F}_\mathcal{X}}\right)^{1/2}\left(\operatorname{var}\left(g\left(\vec{\mathsf{y}}\right)\right) + \kappa\left\|g\right\|^2_{\mathcal{F}_\mathcal{Y}}\right)^{1/2}}, \tag{3.4.7}$$

although this requires an additional parameter κ, which complicates the model selection problem. It is further stated in [7] that this regularisation is responsible for ensuring the consistency of the estimate obtained from the observations, with respect to the regularised population quantity. Comparing with the definition of the kernel covariance (Definition 3.2.2), we note that the KC differs from the KCC in that the former does not contain the variance terms $\mathrm{var}\left(f\left(\vec{\mathrm{x}}\right)\right), \mathrm{var}\left(g\left(\vec{\mathrm{y}}\right)\right)$. In the light of the derivation of Theorem 3.2.4, however, which relies only on the covariance in the numerator of both the KC and KCC, it seems that both contrasts ought to perform similarly when used to measure independence; this observation is borne out in our experimental results (Section 6).

An alternative solution to the problem described in Lemma 3.4.5 is given in [55], in which the projection directions used to compute the canonical correlations are expressed in terms of a more restricted set of basis functions, rather than the respective subspaces of $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Y}}$ spanned by the entire series of mapped observations. These basis functions can be chosen using kernel PCA, for instance.

Chapter 4

Approximations to the mutual information

In this chapter, we investigate the mutual information as a contrast function for measuring independence. We begin in Section 4.1 by introducing the mutual information between two multivariate Gaussian random variables, for which a closed form solution exists. We then describe the discrete approximation to the mutual information between *two continuous, univariate* random variables, and discuss how it relates to the continuous mutual information. Next, we show that the discrete mutual information may be approximated by the Gaussian mutual information near independence. A more general discussion of the basic principles of information theory may be found in Appendix A.8.

In Section 4.2, we begin by deriving a Parzen window estimate of the Gaussian mutual information between two random variables. Next, we give an upper bound on this quantity, which defines the *kernel mutual information* (KMI). We show this is zero if and only if the two random variables are independent. Next, we give more detail on the computation of certain data dependent normalising factors used in the KMI. Finally, we demonstrate that the regularised kernel generalised variance (KGV) proposed in [7] is also an upper bound on the Gaussian mutual information, but is looser than the KMI. A comparison with the original KGV proof is given in Appendix A.6.2.

In Section 4.3, we derive generalisations of the KC and KMI to more than two univariate random variables. We demonstrate that both the KC and KMI are zero if and only if the associated random variables are pairwise independent, which makes them suited for application as contrast functions in ICA.

4.1 Mutual information approximated by multivariate Gaussian random variables

4.1.1 The mutual information between two multivariate Gaussian random variables

We begin by introducing the Gaussian mutual information, and its relation with the canonical correlation. A more detailed and general discussion of these principles may be found in [9]. If $\mathbf{x}_G, \mathbf{y}_G$ are Gaussian random variables[1] in $\mathbb{R}^{l_x}, \mathbb{R}^{l_y}$ respectively, with mean vector $[\boldsymbol{\mu}_x \, \boldsymbol{\mu}_x]$ and

[1] The subscripts G are used to emphasise that $\mathbf{x}_G, \mathbf{y}_G$ are Gaussian; this notation is introduced here to make the reasoning clearer in subsequent sections.

covariance matrix $\mathbf{C}$, then the mutual information between them can be written

$$I\left(\mathbf{x}_G;\mathbf{y}_G\right) = -\frac{1}{2}\log\left(\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|}\right), \tag{4.1.1}$$

using Definition A.8.13 and Theorem A.8.14 in Appendix A.8. A more useful representation, however, is given in the following theorem from [6, 7], which is proved in Appendix A.3.4.

Theorem 4.1.1 (Gaussian mutual information as a function of canonical correlations). *The mutual information of the Gaussian random variables $\mathbf{x}_G, \mathbf{y}_G$ may be written*

$$I\left(\mathbf{x}_G;\mathbf{y}_G\right) = -\frac{1}{2}\log\left(\prod_{i=1}^{\min(l_x,l_y)}\left(1-\rho_i^2\right)\right), \tag{4.1.2}$$

where the ρ_i are the canonical correlations, given by the stationary points in (3.4.4).

We next simplify the ratio of determinants in (4.1.1);

$$\begin{aligned}
\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} &= \frac{\left|\begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}\right|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} && (4.1.3)\\
&= \frac{|\mathbf{C}_{xx}|\left|\mathbf{C}_{yy}-\mathbf{C}_{yx}\mathbf{C}_{x,x}^{-1}\mathbf{C}_{xy}\right|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} && (4.1.4)\\
&= \left|\mathbf{I}-\mathbf{C}_{y,y}^{-1}\mathbf{C}_{yx}\mathbf{C}_{x,x}^{-1}\mathbf{C}_{xy}\right|. && (4.1.5)
\end{aligned}$$

We may rearrange this slightly, to obtain an expression that will be shown to possess useful mathematical properties in the subsequent discussion. Thus,

$$\begin{aligned}
\tfrac{|\mathbf{C}|}{|\mathbf{C}_{xx}||\mathbf{C}_{yy}|} &= \left|\mathbf{I}-\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\right| &&= \left|\mathbf{I}-\mathbf{C}_{yy}^{-1/2}\mathbf{C}_{yy}^{-1/2}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\right|\\
&= \left|\mathbf{C}_{yy}^{-1/2}\right|\left|\mathbf{C}_{yy}^{1/2}-\mathbf{C}_{yy}^{-1/2}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\right| &&= \left|\mathbf{I}-\mathbf{D}^\top\mathbf{D}\right|,
\end{aligned}$$

where $\mathbf{D} = \mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}$.

4.1.2 Mutual information between discretised univariate parameters

In this section, we describe a discretised approximation to the mutual information between two continuous, univariate random variables $\mathsf{x} \in \mathcal{X}$ and $\mathsf{y} \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are both bounded intervals on $\mathbb{R}$ (in other words, $n_x = n_y = 1$). All results and proofs in this section are taken from [7]. Consider a grid of size $l_x \times l_y$ over the range of x and y. Let the indices i, j denote the point $(q_i, r_j) \in \mathcal{X} \times \mathcal{Y}$ on this grid, and let $\boldsymbol{q} = (q_1, \ldots, q_{l_x}), \boldsymbol{r} = (r_1, \ldots, r_{l_y})$ be the complete sequences of grid coordinates. Assume, further, that the spacing between points along the x and y axes is respectively Δ_x and Δ_y. We define two random multinomial random variables $\hat{\mathsf{x}}, \hat{\mathsf{y}}$ with a distribution $\mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$ over the grid (we write the complete $l_x \times l_y$ matrix of such probabilities as $\mathbf{P}_{xy}$); this corresponds to the probability that x, y is within a small interval surrounding the grid position q_i, r_j, so

$$\begin{aligned}
\mathsf{P}_{\hat{\mathsf{x}}}(i) &= \int_{q_i}^{q_i+\Delta_x} \mathsf{f}_{\mathsf{x}}(x)dx, \qquad \mathsf{P}_{\hat{\mathsf{y}}}(j) = \int_{r_j}^{r_j+\Delta_y} \mathsf{f}_{\mathsf{y}}(y)dy,\\
\mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) &= \int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y} \mathsf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy.
\end{aligned}$$

Thus $\mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$ is a discretisation of $\mathsf{f}_{\mathsf{x},\mathsf{y}}$. Finally, we denote as $\mathbf{p}_x$ the vector for which $(\mathbf{p}_x)_i = \mathsf{P}_{\hat{\mathsf{x}}}(i)$, with a similar $\mathbf{p}_y$ definition. The mutual information between $\hat{\mathsf{x}}$ and $\hat{\mathsf{y}}$ is defined as

$$I\left(\hat{\mathsf{x}};\hat{\mathsf{y}}\right) = \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)\log\left(\frac{\mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)}{\mathsf{P}_{\hat{\mathsf{x}}}(i)\,\mathsf{P}_{\hat{\mathsf{y}}}(j)}\right);$$

it is well known that $I(\mathsf{x},\mathsf{y})$ represents the upper bound on $I(\hat{\mathsf{x}};\hat{\mathsf{y}})$ as the discretisation becomes infinitely fine [25]. We may always write $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) = \mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)\,(1+\epsilon_{i,j})$ for an appropriate choice of $\epsilon_{i,j}$, where $\epsilon_{i,j}$ is small near independence. Making this substitution in the above yields

$$I(\hat{\mathsf{x}};\hat{\mathsf{y}}) \approx \frac{1}{2}\sum_{i=1}^{l_x}\sum_{j=1}^{l_y}\mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)\,\epsilon_{i,j}^2, \tag{4.1.6}$$

where the proof may be found in Appendix *A.4*.

4.1.3 Multivariate Gaussian approximation to the discretised mutual information

The results in this section are again from [7], although the proof of (4.1.13) below is novel. We begin by defining an equivalent multidimensional representation $\check{\mathsf{x}},\check{\mathsf{y}}$ of $\hat{\mathsf{x}},\hat{\mathsf{y}}$ in the previous section, where $\check{\mathbf{x}} \in \mathbb{R}^{l_x}$ and $\check{\mathbf{y}} \in \mathbb{R}^{l_y}$, such that $\hat{x} = i$ is equivalent to $(\check{\mathbf{x}})_i = 1$ and $(\check{\mathbf{x}})_{j\,:\,j\neq i} = 0$. To be precise, we define the functions

$$\kappa_i(x) = \begin{cases} 1 & x \in [q_i, q_i+\Delta_x) \\ 0 & \text{otherwise} \end{cases}, \qquad \kappa_j(y) = \begin{cases} 1 & x \in [r_j, r_j+\Delta_y) \\ 0 & \text{otherwise} \end{cases},$$

such that

$$\mathbf{E}_{\mathsf{x}}(\kappa_i(\mathsf{x})) = \mathbf{E}_{\mathsf{x}}((\check{\mathbf{x}})_i) = \int_{-\infty}^{\infty}\kappa_i(\mathsf{x})\mathbf{f}_{\mathsf{x}}(x)dx = \mathbf{P}_{\hat{\mathsf{x}}}(i)$$

and

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}(\kappa_i(\mathsf{x})\kappa_j(\mathsf{y})) = \mathbf{E}_{\mathsf{x},\mathsf{y}}\left((\check{\mathbf{x}})_i\,(\check{\mathbf{y}})_j\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa_i(x)\kappa_j(y)\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy = \mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j).$$

A specific instance of the second formula is when $\mathsf{y} = \mathsf{x}$, and $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y) = \delta_x(y)\mathbf{f}_{\mathsf{x}}(x)$. Then

$$\begin{aligned}\mathbf{E}_{\mathsf{x}}(\kappa_i(\mathsf{x})\kappa_j(\mathsf{x})) = \mathbf{E}_{\mathsf{x}}\left((\check{\mathbf{x}}\check{\mathbf{x}}^\top)_{i,j}\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa_i(x)\kappa_j(y)\mathbf{f}_{\mathsf{x}}(x)\delta_x(y)dxdy \\ &= \begin{cases}\mathbf{P}_{\hat{\mathsf{x}}}(i) & i=j \\ 0 & \text{otherwise}\end{cases}.\end{aligned}$$

In summary,

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right) = \mathbf{P}_{xy} \tag{4.1.7}$$

$$\mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\right) = \mathbf{p}_x \tag{4.1.8}$$

$$\mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right) = \mathbf{D}_x \tag{4.1.9}$$

where $\mathbf{D}_x = \mathrm{diag}(\mathbf{p}_x)$. Using these results, it is possible to define the covariances

$$\mathbf{C}_{xy} = \mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right) - \mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\right)\mathbf{E}_{\mathsf{y}}\left(\check{\mathbf{y}}\right)^\top = \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top, \tag{4.1.10}$$

$$\mathbf{C}_{xx} = \mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right) - \mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\right)\mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\right)^\top = \mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top, \tag{4.1.11}$$

$$\mathbf{C}_{yy} = \mathbf{E}_{\mathsf{y}}\left(\check{\mathbf{y}}\check{\mathbf{y}}^\top\right) - \mathbf{E}_{\mathsf{y}}\left(\check{\mathbf{y}}\right)\mathbf{E}_{\mathsf{y}}\left(\check{\mathbf{y}}\right)^\top = \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top. \tag{4.1.12}$$

Using the above definitions, we may define Gaussian random variables $\mathsf{x}_G, \mathsf{y}_G$ with the same covariance structure as $\check{\mathsf{x}},\check{\mathsf{y}}$, and with mutual information given by (4.1.1). We prove in Appendix A.5.1 that the mutual information for this Gaussian case is

$$I(\mathsf{x}_G;\mathsf{y}_G) = -\frac{1}{2}\log\left(\left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}\right|\right), \tag{4.1.13}$$

which may be approximated by

$$I(\mathsf{x}_G;\mathsf{y}_G) \approx \frac{1}{2}\sum_{i=1}^{l_x}\sum_{j=1}^{l_y}\mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)\,\epsilon_{i,j}^2 \tag{4.1.14}$$

(Appendix A.5.2). The latter expression is identical to the approximate mutual information in (4.1.6) of the discretised approximation $\hat{\mathsf{x}}, \hat{\mathsf{y}}$ to x, y. On this basis, it is possible to show that the multivariate Gaussian mutual information in (4.1.14) is a good approximation to the mutual information between continuous, univariate variables with arbitrary probability distribution.

4.2 Kernel density estimate of the discretised mutual information

In this section, we describe a kernel density estimate of the approximate mutual information in (4.1.13). Before proceeding, however, we motivate this discussion by reviewing the Parzen window estimate and its properties, as discussed in [80, 31] (this discussion pertains to the general case of multivariate $\vec{\mathsf{x}}$, although our application requires only univariate random variables). Given a sample $\boldsymbol{x}$ of size m, each point $\vec{x}_l$ of which is assumed generated i.i.d. according to some unknown distribution with density $\mathbf{f}_{\vec{\mathsf{x}}}$, the associated Parzen window estimate of this density is written[2]

$$\widehat{\mathbf{f}}_{\vec{\mathsf{x}}}(\vec{x}) = \frac{1}{m} \sum_{l=1}^{m} k\left(\vec{x}_l - \vec{x}\right).$$

The kernel function $k\left(\vec{x}_l - \vec{x}\right)$ must be a legitimate probability density function, in that it should be correctly normalised,

$$\int_{\mathcal{X}} k\left(\vec{x}\right) d\vec{x} = 1, \tag{4.2.1}$$

and $k\left(\vec{x}\right) \geq 0$. We may rescale the kernel according to $\frac{1}{V_x} k\left(\frac{\vec{x}}{\sigma_x}\right)$, where the term V_x is needed to preserve (4.2.1)[3]. The following theorem then describes the convergence of the Parzen window estimate to the true probability density.

Theorem 4.2.1 (Convergence of the Parzen window estimate). *Let $k_m(\vec{x})$ be the Parzen window chosen for a sample $\boldsymbol{x}$ of size m, where the window size is chosen by varying $\sigma_{x,m}$ with increasing m, with associated normalisation constant $V_{x,m}$. The density estimate $\widehat{\mathbf{f}}_{\vec{\mathsf{x}}}(\vec{x})$ converges to $\mathbf{f}_{\vec{\mathsf{x}}}$ in the* mean square sense *as $m \to \infty$, subject to the kernel $k_m\left(\vec{x}\right)$ being a legitimate probability density, and to*

- $\sup_{\vec{x}} k_m(\vec{x}) < \infty$,
- $\lim_{\|\vec{x}\|_{\mathcal{X}} \to \infty} \prod_i x_i \, k_m(\vec{x}) = 0$,
- $\lim_{m \to \infty} V_{x,m} = 0$,
- $\lim_{m \to \infty} m V_{x,m} = \infty$.

The last two properties in the above definition give an indication of how one might choose the kernel size as a function of the number of observed samples. Common choices are $V_{x,m} = V_{x,1}/\sqrt{m}$ and $V_{x,m} = V_{x,1}/\log m$. This method requires an initial parameter choice for a particular sample size (written $V_{x,1}$ in this case, in a slight abuse of index notation), which can be obtained by cross validation.

[2] The Parzen window is deliberately written in the same way as the RKHS kernel in Section 3.2, since it is in fact the same function; this link is described in detail in the present section.

[3] Here we scale each component x_i of $\vec{x}$ by the same factor σ_x.

4.2.1 Exact expression for the kernel density estimate

We return now to the problem described in Sections 4.1.2 and 4.1.3. We are given a sample of length m, $\boldsymbol{z} := ((x_1, y_1), \ldots, (x_m, y_m))$, from $\mathbf{P}_{\mathsf{x},\mathsf{y}}$. The kernel density (Parzen window) estimates are then

$$\begin{aligned} \widehat{\mathbf{f}}_{\mathsf{x}}(x) &= \frac{1}{m}\sum_{l=1}^{m} k\left(x_l - x\right), \qquad \widehat{\mathbf{f}}_{\mathsf{y}}(y) = \frac{1}{m}\sum_{l=1}^{m} k\left(y_l - y\right), \\ \widehat{\mathbf{f}}_{\mathsf{x},\mathsf{y}}(x,y) &= \frac{1}{m}\sum_{l=1}^{m} k\left(x_l - x\right) k\left(y_l - y\right), \end{aligned}$$

where the kernel argument is used to specify which kernel is used, to simplify notation. We require approximations to the terms in the Gaussian mutual information, as described in (4.1.13). We therefore define the vectors $\hat{\mathbf{p}}_x, \hat{\mathbf{p}}_y$, and the matrix $\widehat{\mathbf{P}}_{xy}$, using the expectations in (4.1.7)-(4.1.9) computed with these kernel expressions;

$$\widehat{\mathbf{E}}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\,\check{\mathbf{y}}^\top\right) = \widehat{\mathbf{P}}_{xy}, \tag{4.2.2}$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\check{\mathbf{x}}\right) = \hat{\mathbf{p}}_x, \tag{4.2.3}$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\check{\mathbf{x}}\,\check{\mathbf{x}}^\top\right) = \widehat{\mathbf{D}}_x. \tag{4.2.4}$$

In the limit where Δ_x, Δ_y are small (and thus, by implication, $l_x \gg m$, $l_y \gg m$, $\sigma_k \gg \Delta_x$, and $\sigma_k \gg \Delta_y$, where σ_k defines the kernel size), we make the approximations

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left((\check{\mathbf{x}})_i\right) = \widehat{\mathbf{P}}_{\hat{\mathsf{x}}}\left(i\right) = \frac{1}{m}\int_{q_i}^{q_i+\Delta_x}\sum_{l=1}^{m} k\left(x_l - x\right)dx \approx \frac{\Delta_x}{m}\sum_{l=1}^{m} k\left(x_l - q_i\right),$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\left(\check{\mathbf{x}}\,\check{\mathbf{x}}^\top\right)_{i,j}\right) \approx \begin{cases} \frac{\Delta_x}{m}\sum_{l=1}^{m} k\left(x_l - q_i\right) & i = j \\ 0 & \text{otherwise} \end{cases},$$

and

$$\begin{aligned} \widehat{\mathbf{E}}_{\mathsf{x},\mathsf{y}}\left(\left(\check{\mathbf{x}}\,\check{\mathbf{y}}^\top\right)_{i,j}\right) = \widehat{\mathbf{P}}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}\left(i,j\right) &= \frac{1}{m}\int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y}\sum_{l=1}^{m} k\left(x_l - x\right) k\left(y_l - y\right) dxdy \\ &\approx \frac{\Delta_x\Delta_y}{m}\sum_{l=1}^{m} k\left(x_l - q_i\right) k\left(y_l - r_j\right). \end{aligned}$$

Finally, the normalisation condition (4.2.1) becomes

$$1 = \int_{-\infty}^{\infty} k(x-s)ds \approx \Delta_x \sum_{i=1}^{l_x} k\left(q_i - x\right),$$

and therefore

$$\frac{1}{\Delta_x} \approx \sum_{i=1}^{l_x} k\left(q_i - x\right) \quad \text{and} \quad \frac{1}{\Delta_y} \approx \sum_{j=1}^{l_y} k\left(r_j - y\right). \tag{4.2.5}$$

Before proceeding further, we define two matrices of kernel inner products to simplify our notation. Namely,

$$\mathbf{K}^{(x)}_{lm} := \begin{bmatrix} k\left(q_1 - x_1\right) & \ldots & k\left(q_1 - x_m\right) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ k\left(q_{l_x} - x_1\right) & \ldots & k\left(q_{l_x} - x_m\right) \end{bmatrix}, \quad \mathbf{K}^{(y)}_{lm} := \begin{bmatrix} k\left(r_1 - y_1\right) & \ldots & k\left(r_1 - y_m\right) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ k\left(r_{l_y} - y_1\right) & \ldots & k\left(r_{l_y} - y_m\right) \end{bmatrix},$$

where we write the above in such a manner as to indicate $l_x \gg m$ and $l_y \gg m$. The first subscript specifies whether the grid ($\boldsymbol{q}$ and $\boldsymbol{r}$) or the sample ($\boldsymbol{x}$ and $\boldsymbol{y}$) is used in the rows of the kernel matrix, and the second subscript whether the grid or sample is used in the columns. By analogy, we may also define the matrices $\mathbf{K}_{ll}^{(x)}, \mathbf{K}_{mm}^{(x)}, \mathbf{K}_{ll}^{(y)}, \mathbf{K}_{mm}^{(y)}$. Two useful results that follow from (4.2.5) are

$$\left(\mathbf{K}_{lm}^{(x)}\right)^\top \mathbf{1}_{l_x} \approx \frac{1}{\Delta_x}\mathbf{1}_m \quad \text{and} \quad \left(\mathbf{K}_{lm}^{(y)}\right)^\top \mathbf{1}_{l_y} \approx \frac{1}{\Delta_y}\mathbf{1}_m.$$

We now redefine the estimates of the matrices used to compute the Gaussian mutual information. These are

$$\widehat{\mathbf{P}}_{xy} - \hat{\mathbf{p}}_x\hat{\mathbf{p}}_y^\top \approx \frac{\Delta_x\Delta_y}{m}\left(\mathbf{K}_{lm}^{(x)}\left(\mathbf{K}_{lm}^{(y)}\right)^\top - \frac{1}{m}\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\mathbf{1}_m^\top\left(\mathbf{K}_{lm}^{(y)}\right)^\top\right),$$

$$\widehat{\mathbf{D}}_x \approx \frac{\Delta_x}{m}\mathrm{diag}\left(\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\right) = \frac{\Delta_x^2}{m}\mathrm{diag}\left(\mathbf{K}_{lm}^{(x)}\left(\mathbf{K}_{lm}^{(x)}\right)^\top \mathbf{1}_{l_x}\right),$$

and

$$\widehat{\mathbf{D}}_y \approx \frac{\Delta_y^2}{m}\mathrm{diag}\left(\mathbf{K}_{lm}^{(y)}\left(\mathbf{K}_{lm}^{(y)}\right)^\top \mathbf{1}_{l_y}\right).$$

Substituting these terms into (4.1.13), it can be seen that the factors $\Delta_x\Delta_y/m$, Δ_x^2/m and Δ_y^2/m cancel. Furthermore, the matrices being inverted have full rank, as required. Comparing (4.1.13) with (4.1.5), we may proceed by analogy with (4.1.2), and (3.4.3), using the proof strategy in Appendix A.3.4, to obtain the kernel density approximation to the discretised mutual information,

$$\widehat{I}\left(\hat{\mathrm{x}};\hat{\mathrm{y}}\right) = -\frac{1}{2}\log\left(\prod_{i=1}^{\min(l_x,l_y)}\left(1-\hat{\rho}_i^2\right)\right), \tag{4.2.6}$$

where

$$\hat{\rho}_i = \frac{\hat{\mathbf{c}}_i^\top\left(\widehat{\mathbf{P}}_{xy} - \hat{\mathbf{p}}_x\hat{\mathbf{p}}_y^\top\right)\hat{\mathbf{d}}_i}{\sqrt{\hat{\mathbf{c}}_i^\top\widehat{\mathbf{D}}_x\hat{\mathbf{c}}_i\,\hat{\mathbf{d}}_i^\top\widehat{\mathbf{D}}_y\hat{\mathbf{d}}_i}} \tag{4.2.7}$$

$$\approx \frac{\left(\frac{\Delta_x\Delta_y}{m}\right)\hat{\mathbf{c}}_i^\top\left(\mathbf{K}_{lm}^{(x)}\left(\mathbf{K}_{lm}^{(y)}\right)^\top - \frac{1}{m}\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\mathbf{1}_m^\top\left(\mathbf{K}_{lm}^{(y)}\right)^\top\right)\hat{\mathbf{d}}_i}{\sqrt{\left(\frac{\Delta_x}{m}\right)\hat{\mathbf{c}}_i^\top\mathrm{diag}\left(\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\right)\hat{\mathbf{c}}_i\,\hat{\mathbf{d}}_i^\top\left(\frac{\Delta_x}{m}\right)\mathrm{diag}\left(\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\right)\hat{\mathbf{d}}_i}} \tag{4.2.8}$$

$$= \frac{\hat{\mathbf{c}}_i^\top\left(\mathbf{K}_{lm}^{(x)}\left(\mathbf{K}_{lm}^{(y)}\right)^\top - \frac{1}{m}\mathbf{K}_{lm}^{(x)}\mathbf{1}_m\mathbf{1}_m^\top\left(\mathbf{K}_{lm}^{(y)}\right)^\top\right)\hat{\mathbf{d}}_i}{\sqrt{\hat{\mathbf{c}}_i^\top\mathrm{diag}\left(\mathbf{K}_{lm}^{(x)}\left(\mathbf{K}_{lm}^{(x)}\right)^\top\mathbf{1}_{l_x}\right)\hat{\mathbf{c}}_i\,\hat{\mathbf{d}}_i^\top\mathrm{diag}\left(\mathbf{K}_{lm}^{(y)}\left(\mathbf{K}_{lm}^{(y)}\right)^\top\mathbf{1}_{l_y}\right)\hat{\mathbf{d}}_i}}. \tag{4.2.9}$$

The form of (4.2.8) might mislead us into supposing that the expressions in the numerator and denominator are expressible as Riemann integrals in the limit as the grid becomes increasingly fine. This is not correct, however, since the solutions $\hat{\mathbf{c}}_i, \hat{\mathbf{d}}_i$ are *not* constant at a given set of grid points when new grid points are added. In the next section, we derive an upper bound for $\hat{\rho}_i$ that is independent of the discretisation.

4.2.2 An upper bound on the kernel density estimate

To compute an upper bound on the kernel density estimate of the discretised mutual information, we begin with the following theorem.

Theorem 4.2.2 (Effect on norm of taking sums of rows). *If* $\mathbf{B}$ *is a symmetric* $n \times n$ *matrix with positive elements* $b_{i,j}$*, and* $\mathbf{c}$ *an arbitrary* $n \times 1$ *vector with elements* c_i*, then*

$$\mathbf{c}^\top \operatorname{diag}\left(\mathbf{B}\mathbf{1}_n\right)\mathbf{c} \geq \mathbf{c}^\top \mathbf{B}\mathbf{c}.$$

The proof is in Appendix A.7.1. We now apply this result to the computation of an upper bound on $\hat{\rho}_i$ in (4.2.8). We define the quantities

$$\nu_{\boldsymbol{x}} := \min_{j \in \{1 \ldots l_x\}} \sum_{l=1}^{m} k\left(x_l, q_j\right), \qquad \nu_{\boldsymbol{y}} := \min_{j \in \{1 \ldots l_y\}} \sum_{l=1}^{m} k\left(y_l, r_j\right). \tag{4.2.10}$$

Then

$$\begin{aligned}
\nu_{\boldsymbol{x}} \hat{\mathbf{c}}_i^\top \mathbf{K}_{ll}^{(x)} \hat{\mathbf{c}}_i &\leq \nu_{\boldsymbol{x}} \hat{\mathbf{c}}_i^\top \operatorname{diag}\left(\mathbf{K}_{ll}^{(x)} \mathbf{1}_{l_x}\right) \hat{\mathbf{c}}_i \\
&\approx \frac{\nu_{\boldsymbol{x}}}{\Delta_x} \hat{\mathbf{c}}_i^\top \mathbf{I}_{l_x} \hat{\mathbf{c}}_i \\
&\leq \hat{\mathbf{c}}_i^\top \begin{bmatrix} \frac{1}{\Delta_x}\sum_{l=1}^{m} k\left(q_1, x_l\right) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\Delta_x}\sum_{l=1}^{m} k\left(q_{l_x}, x_l\right) \end{bmatrix} \hat{\mathbf{c}}_i \\
&\approx \hat{\mathbf{c}}_i^\top \operatorname{diag}\left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)}\right)^\top \mathbf{1}_p\right) \hat{\mathbf{c}}_i,
\end{aligned}$$

where both the second and final lines use (4.2.5). Defining

$$\tilde{\gamma}_i := \frac{\hat{\mathbf{c}}_i^\top \left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(y)}\right)^\top - \frac{1}{m}\mathbf{K}_{lm}^{(x)} \mathbf{1}_m \mathbf{1}_m^\top \left(\mathbf{K}_{lm}^{(y)}\right)^\top\right) \hat{\mathbf{d}}_i}{\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}} \sqrt{\hat{\mathbf{c}}_i^\top \mathbf{K}_{ll}^{(x)} \hat{\mathbf{c}}_i \hat{\mathbf{d}}_i^\top \mathbf{K}_{ll}^{(y)} \hat{\mathbf{d}}_i}}, \tag{4.2.11}$$

we find $\tilde{\gamma}_i \geq \hat{\rho}_i$ (this new quantity is a normalised covariance, and *not* a correlation, hence the notation).

We next find an upper bound γ_i on $\tilde{\gamma}_i$ that is independent of the discretisation, using the geometric properties of the RKHS interpretation of (4.2.11). We define the representations in $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Y}}$ of the grid points $\boldsymbol{q} = (q_1, \ldots, q_{l_x}), \boldsymbol{r} = \left(r_1, \ldots, r_{l_y}\right)$ as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1 & \ldots & \mathbf{q}_{l_x} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{r}_1 & \ldots & \mathbf{r}_{l_y} \end{bmatrix}.$$

Using the feature space representations $\mathbf{X}, \mathbf{Y}$ of $\boldsymbol{x}, \boldsymbol{y}$ defined in (3.1.5), and the definition of the centering matrix $\mathbf{H}$ in (3.1.7), we may rewrite (4.2.11) as

$$\tilde{\gamma}_i := \frac{\hat{\mathbf{c}}_i^\top \mathbf{Q}^\top \mathbf{X}\mathbf{H}\left(\mathbf{Y}\mathbf{H}\right)^\top \mathbf{R}\hat{\mathbf{d}}_i}{\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}} \sqrt{\left(\hat{\mathbf{c}}_i^\top \mathbf{Q}^\top \mathbf{Q}\hat{\mathbf{c}}_i\right)\left(\hat{\mathbf{d}}_i^\top \mathbf{R}^\top \mathbf{R}\hat{\mathbf{d}}_i\right)}}. \tag{4.2.12}$$

The following theorem is proved in Appendix A.2.3.

Theorem 4.2.3 (Eigenvalue problem for normalised covariance with restricted projections). *Let* $\widetilde{\mathbf{X}} := \mathbf{X}\mathbf{H}$ *and* $\widetilde{\mathbf{Y}} := \mathbf{Y}\mathbf{H}$ *be the centered matrices in their respective feature spaces of the points in* $\boldsymbol{z}$*. The stationary points of* $\tilde{\gamma}_i$ *in (4.2.12) with respect to* $\hat{\mathbf{c}}_i, \hat{\mathbf{d}}_i$ *are given by the solutions to the eigenvalue problem*

$$\begin{bmatrix} \mathbf{P}_Q & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_R \end{bmatrix} \begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top \\ \widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_i \\ \widehat{\boldsymbol{\beta}}_i \end{bmatrix} = \left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}} \tilde{\gamma}_i \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_i \\ \widehat{\boldsymbol{\beta}}_i \end{bmatrix}, \tag{4.2.13}$$

where

$$\widehat{\boldsymbol{\alpha}}_i = \mathbf{Q}\hat{\mathbf{c}}_i, \quad \widehat{\boldsymbol{\beta}}_i = \mathbf{R}\hat{\mathbf{d}}_i, \tag{4.2.14}$$

and we define the projection operators

$$\mathbf{P}_Q = \mathbf{Q}\left(\mathbf{Q}^\top \mathbf{Q}\right)^- \mathbf{Q}^\top, \quad \mathbf{P}_R = \mathbf{R}\left(\mathbf{R}^\top \mathbf{R}\right)^- \mathbf{R}^\top.$$

We see immediately that problem described in (4.2.13) is very similar to the normalised covariance problem in (3.2.3), but with the additional requirement that the solutions be projected onto the basis spanned by the columns of $\begin{bmatrix} \mathbf{Q}^\top & \mathbf{R}^\top \end{bmatrix}^\top$, as well as an added scaling factor[4]. In particular, the non-zero solutions γ_i of (3.2.3) enumerate the (potentially) non-zero solutions $\tilde{\gamma}_i$ of (4.2.13), where we bear in mind that the projection operations can increase the nullspace.

Using this insight, we can prove that $|\tilde{\gamma}_i| \leq |\gamma_i|$ for all $\tilde{\gamma}_i$ with a corresponding non-zero γ_i (the eigenvalues obtained as the solutions to both problems come in pairs, with equal value and opposite sign, which is why it is necessary to take the absolute value). We begin by decomposing the solutions of (3.2.3) as

$$\boldsymbol{\alpha}_i = \widehat{\boldsymbol{\alpha}}_i + \boldsymbol{\alpha}_{i,\perp}, \quad \boldsymbol{\beta}_i = \widehat{\boldsymbol{\beta}}_i + \boldsymbol{\beta}_{i,\perp},$$

where $\boldsymbol{\alpha}_{i,\perp}$ is the component of $\boldsymbol{\alpha}_i$ perpendicular to the column space of $\mathbf{Q}$, and $\boldsymbol{\beta}_{i,\perp}$ the component of $\boldsymbol{\beta}_i$ perpendicular to the column space of $\mathbf{R}$. In addition, we write

$$\widetilde{\mathbf{X}} = \mathbf{P}_Q \widetilde{\mathbf{X}} + \widetilde{\mathbf{X}}_\perp, \quad \widetilde{\mathbf{Y}} = \mathbf{P}_R \widetilde{\mathbf{Y}} + \widetilde{\mathbf{Y}}_\perp.$$

The system of equations in (3.2.3) then becomes

$$\left\{ \begin{array}{rcl} \left(\mathbf{P}_Q \widetilde{\mathbf{X}} + \widetilde{\mathbf{X}}_\perp\right) \widetilde{\mathbf{Y}}^\top \left(\widehat{\boldsymbol{\beta}}_i + \boldsymbol{\beta}_{i,\perp}\right) & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \gamma_i \left(\widehat{\boldsymbol{\alpha}}_i + \boldsymbol{\alpha}_{i,\perp}\right) \\ \left(\mathbf{P}_R \widetilde{\mathbf{Y}} + \widetilde{\mathbf{Y}}_\perp\right) \widetilde{\mathbf{X}}^\top \left(\widehat{\boldsymbol{\alpha}}_i + \boldsymbol{\alpha}_{i,\perp}\right) & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \gamma_i \left(\widehat{\boldsymbol{\beta}}_i + \boldsymbol{\beta}_{i,\perp}\right) \end{array} \right. \tag{4.2.15}$$

Considering only those components in the span of the columns of $\mathbf{Q}$ and $\mathbf{R}$, we obtain

$$\left\{ \begin{array}{rcl} \mathbf{P}_Q \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^\top \widehat{\boldsymbol{\beta}}_i + \mathbf{P}_Q \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_{i,\perp} & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \gamma_i \widehat{\boldsymbol{\alpha}}_i, \\ \mathbf{P}_R \widetilde{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \widehat{\boldsymbol{\alpha}}_i + \mathbf{P}_R \widetilde{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_{i,\perp} & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \gamma_i \widehat{\boldsymbol{\beta}}_i, \end{array} \right.$$

or

$$\left\{ \begin{array}{rcl} \mathbf{P}_Q \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^\top \widehat{\boldsymbol{\beta}}_i & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} (\gamma_i - \theta_i) \widehat{\boldsymbol{\alpha}}_i, \\ \mathbf{P}_R \widetilde{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \widehat{\boldsymbol{\alpha}}_i & = & (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} (\gamma_i - \theta_i) \widehat{\boldsymbol{\beta}}_i, \end{array} \right.$$

where $(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \theta_i \widehat{\boldsymbol{\alpha}}_i = \mathbf{P}_Q \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_{i,\perp}$ and $(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \theta_i \widehat{\boldsymbol{\beta}}_i = \mathbf{P}_R \widetilde{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_{i,\perp}$. We conclude that the solutions to (4.2.13) can be written $\tilde{\gamma}_i = \gamma_i - \theta_i$. To complete the proof, we must take into account the geometric properties of the feature spaces $\mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}$. In particular, when the reproducing kernel Hilbert space is induced by a Gaussian or Laplace kernel, the angle between any two vectors in this RKHS must be less than $\pi/2$, and all vectors are of equal magnitude and located in the positive quadrant. Thus $\mathbf{P}_Q \widetilde{\mathbf{X}} \widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_{i,\perp}$ and $\mathbf{P}_R \widetilde{\mathbf{Y}} \widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_{i,\perp}$ have the same sign as γ_i, and $|\tilde{\gamma}_i| \leq |\gamma_i|$ as required.

Drawing the results in this section together, we see that

$$\prod_i \left(1 - \gamma_i^2\right) \leq \prod_i \left(1 - \tilde{\gamma}_i^2\right) \leq \prod_i \left(1 - \hat{\rho}_i^2\right),$$

and

$$-\frac{1}{2} \log \left(\prod_i \left(1 - \gamma_i^2\right) \right) \geq -\frac{1}{2} \log \left(\prod_i \left(1 - \tilde{\gamma}_i^2\right) \right) \geq -\frac{1}{2} \log \left(\prod_i \left(1 - \hat{\rho}_i^2\right) \right).$$

This motivates us to introduce a new contrast function, as described in the following definition, where we use the Gram matrix expression in (3.2.5) for the kernel normalised covariance.

Definition 4.2.4 (The kernel mutual information). Let $\boldsymbol{z}$ be a sample of size m consisting of points $\boldsymbol{x}, \boldsymbol{y}$, and let $\mathbf{X}$, $\mathbf{Y}$ be the corresponding feature space representations. Then an upper

[4]It is therefore implied that the scaling factor $(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}}$ is in this case included in the computation of γ_i, as will be made explicit in (4.2.15) below.

bound on the kernel density approximation to the mutual information near independence is given by the *kernel mutual information*, defined as

$$\mathcal{M}\left(\boldsymbol{x},\boldsymbol{y},\mathcal{F}_{\mathcal{X}},\mathcal{F}_{\mathcal{Y}}\right) := -\frac{1}{2}\log\left(\prod_i\left(1-\gamma_i^2\right)\right),$$

where γ_i are the non-zero solutions to

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \\ \widetilde{\mathbf{K}}_{mm}^{(y)}\widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\gamma_i\begin{bmatrix} \widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}_{mm}^{(y)} \end{bmatrix}\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}$$

Using the fact that $\widetilde{\mathbf{K}}_{mm}^{(x)}$ and $\widetilde{\mathbf{K}}_{mm}^{(y)}$ are positive semidefinite, we may again proceed by analogy with (3.4.4), (4.1.1), and (4.1.2), using the proof strategy in Appendix A.3.4, to obtain a more convenient form for the above;

$$\begin{aligned}\mathcal{M}\left(\boldsymbol{x},\boldsymbol{y},\mathcal{F}_{\mathcal{X}},\mathcal{F}_{\mathcal{Y}}\right) &= -\frac{1}{2}\log\left(\frac{\left|\begin{bmatrix} \left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\widetilde{\mathbf{K}}_{mm}^{(x)} & \widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \\ \widetilde{\mathbf{K}}_{mm}^{(y)}\widetilde{\mathbf{K}}_{mm}^{(x)} & \left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\widetilde{\mathbf{K}}_{mm}^{(y)} \end{bmatrix}\right|}{\left|\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\widetilde{\mathbf{K}}_{mm}^{(x)}\right|\left|\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\widetilde{\mathbf{K}}_{mm}^{(x)}\right|}\right) \\ &= -\frac{1}{2}\log\left(\left|\mathbf{I}-\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)\widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)}\right|\right).\end{aligned}$$

The KMI also inherits the following important property from the kernel covariance.

Theorem 4.2.5 (The population KMI is zero at independence). *The random variables* x,y *are independent if and only if* $\mathcal{M}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}},\mathcal{F}_{\mathcal{X}},\mathcal{F}_{\mathcal{Y}}\right)=0$.

This theorem follows from Theorem 3.2.4, bearing in mind that the kernel correlation is the largest eigenvalue γ_i.

4.2.3 Practical choice of $\nu_{\boldsymbol{x}}$ and $\nu_{\boldsymbol{y}}$

In section 4.2.2, our upper bound on the mutual information incorporated the constants $\nu_{\boldsymbol{x}}$ and $\nu_{\boldsymbol{y}}$, which are defined in (4.2.10). These cannot readily be obtained in practice, however, so we now propose more easily computable replacements. We base our reasoning on the fact that γ_i is an upper bound on $\hat{\gamma}_i$ regardless of the grid spacing. Hence, when the grid is very fine with respect to the spacing of the observations in $\boldsymbol{z}$, then only those coefficients of $\hat{\mathbf{c}}_i$ and $\hat{\mathbf{d}}_i$ corresponding to the grid points nearest to the samples $\boldsymbol{x},\boldsymbol{y}$ are non-zero (assuming that the kernel is continuous).

We now state this idea more formally. Assuming the grid is sufficiently fine, we can define $\breve{\boldsymbol{q}} := (\breve{q}_1,\ldots,\breve{q}_m)$ such that $\breve{q}_l = \arg\min_{q\in\boldsymbol{q}}|x_l-q|$ (we will have $\breve{q}_i\neq\breve{q}_j$ for $i\neq j$ as long as the spacing in the grid is much smaller than the spacing between the samples $\boldsymbol{x}$), with an analogous definition for $\breve{\boldsymbol{r}}$. We also define the Gram matrices $\breve{\mathbf{K}}_{mm}^{(x)}$, $\breve{\mathbf{K}}_{mm}^{(y)}$ on these grid points, and vectors $\breve{\mathbf{c}}_i,\breve{\mathbf{d}}_i$ as the entries in $\hat{\mathbf{c}}_i,\hat{\mathbf{d}}_i$ corresponding to $\breve{\boldsymbol{q}}$ and $\breve{\boldsymbol{r}}$. Recall that the upper bound γ_i on $\tilde{\gamma}_i$ is written

$$\gamma_i := \frac{\mathbf{c}_i^\top\widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)}\mathbf{d}_i}{\left(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}}\right)^{\frac{1}{2}}\sqrt{\mathbf{c}_i^\top\widetilde{\mathbf{K}}_{mm}^{(x)}\mathbf{c}_i\mathbf{d}_i^\top\widetilde{\mathbf{K}}_{mm}^{(y)}\mathbf{d}_i}}. \tag{4.2.16}$$

Thus, when $\boldsymbol{q},\boldsymbol{r}$ are closely spaced, it follows[5] that

[5] The reader is reminded that $\tilde{\gamma}_i$ is a *maximum* over the choice of $\hat{\mathbf{c}}_i$ and $\hat{\mathbf{d}}_i$, and therefore approaches the upper bound γ_i as $\breve{\boldsymbol{q}}$ and $\breve{\boldsymbol{r}}$ approach $\boldsymbol{x}$ and $\boldsymbol{y}$.

$$\begin{aligned}\tilde{\gamma}_i &:= \frac{\hat{\mathbf{c}}_i^\top \mathbf{K}_{lm}^{(x)} \mathbf{H} \left(\mathbf{K}_{lm}^{(y)}\right)^\top \hat{\mathbf{d}}_i}{(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \sqrt{\hat{\mathbf{c}}_i^\top \mathbf{K}_{ll}^{(x)} \hat{\mathbf{c}}_i \hat{\mathbf{d}}_i^\top \mathbf{K}_{ll}^{(y)} \hat{\mathbf{d}}_i}} \\ &\approx \frac{\breve{\mathbf{c}}_i^\top \breve{\mathbf{K}}_{mm}^{(x)} \mathbf{H} \left(\breve{\mathbf{K}}_{mm}^{(y)}\right)^\top \breve{\mathbf{d}}_i}{(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}} \left(\breve{\mathbf{c}}_i^\top \breve{\mathbf{K}}_{mm}^{(x)} \breve{\mathbf{c}}_i\right)^{1/2} \left(\breve{\mathbf{d}}_i^\top \breve{\mathbf{K}}_{mm}^{(x)} \breve{\mathbf{d}}_i\right)^{1/2}},\end{aligned}$$

where we replace $\breve{\mathbf{c}}_i = \mathbf{H}\mathbf{c}_i$ and $\breve{\mathbf{d}}_i = \mathbf{H}\mathbf{d}_i$ to obtain the centered Gram matrices in (4.2.16). In the light of the above, we do not need to compute the minima in (4.2.10) over the entire grid, but merely over those grid points $\breve{\boldsymbol{q}}, \breve{\boldsymbol{r}}$ closest to the observations. Specifically, if we define the new constants

$$\breve{\nu}_{\boldsymbol{x}} := \min_{j \in \{1 \ldots m\}} \sum_{l=1}^{m} k(x_l, x_j) \approx \min_{j \in \{1 \ldots m\}} \sum_{l=1}^{m} k(x_l, \breve{q}_j), \tag{4.2.17}$$

$$\breve{\nu}_{\boldsymbol{y}} := \min_{j \in \{1 \ldots m\}} \sum_{l=1}^{m} k(y_l, y_j) \approx \min_{j \in \{1 \ldots m\}} \sum_{l=1}^{m} k(x_l, \breve{r}_j), \tag{4.2.18}$$

then the leftmost term in the denominator of (4.2.8) may be lower bounded by

$$\begin{aligned}\breve{\nu}_{\boldsymbol{x}} \breve{\mathbf{c}}_i^\top \breve{\mathbf{K}}_{mm}^{(x)} \breve{\mathbf{c}}_i &\leq \breve{\nu}_{\boldsymbol{x}} \breve{\mathbf{c}}_i^\top \operatorname{diag}\left(\breve{\mathbf{K}}_{mm}^{(x)} \mathbf{1}_m\right) \breve{\mathbf{c}}_i \\ &\leq \breve{\mathbf{c}}_i^\top \left(\operatorname{diag}\left(\breve{\mathbf{K}}_{mm}^{(x)} \mathbf{1}_m\right)\right)^2 \breve{\mathbf{c}}_i \\ &\leq \breve{\mathbf{c}}_i^\top \operatorname{diag}\left(\begin{bmatrix} \left(\sum_{i=1}^{l_x} k(q_i - x_1)\right)\left(\sum_{j=1}^{m} k(x_j - x_1)\right) \\ \vdots \\ \left(\sum_{i=1}^{l_x} k(q_i - x_m)\right)\left(\sum_{j=1}^{m} k(x_j - x_m)\right) \end{bmatrix}\right) \breve{\mathbf{c}}_i \\ &\leq \hat{\mathbf{c}}_i^\top \begin{bmatrix} \frac{1}{\Delta_x}\sum_{l=1}^{m} k(q_1, x_l) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\Delta_x}\sum_{l=1}^{m} k(q_{l_x}, x_l) \end{bmatrix} \hat{\mathbf{c}}_i \\ &= \hat{\mathbf{c}}_i^\top \operatorname{diag}\left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)}\right)^\top \mathbf{1}_p\right) \hat{\mathbf{c}}_i.\end{aligned}$$

We may therefore replace the constants $\nu_{\boldsymbol{x}}, \nu_{\boldsymbol{y}}$ in Definition 4.2.4 with the new constants in (4.2.17) and (4.2.18).

We now briefly return to the alternative kernel density estimate of [70], as described in Section 2.2.4. Aside from the fact that we approximate the mutual information, rather than the entropy, an important difference in our respective approaches is that our contrast is computed in the limit of infinitely small grid size, which removes the need for binning. Thus, we retain our original kernel, rather than using a spline kernel in all cases. This allows us greater freedom to choose a kernel density appropriate to the characteristics of the sources.

4.2.4 An alternative upper bound on the kernel density estimate

Bach and Jordan [7] propose two related quantities as contrast functions for ICA: the kernel canonical correlation (KCC), as discussed in Section 3.4, and the kernel generalised variance (KGV). In this section, we demonstrate that the latter quantity may be derived by finding an upper bound

on (4.2.8). This derivation, which is based on the kernel density estimate, takes a completely different approach to the proof in [7], which uses a limit as the kernel becomes infinitely small. In any event, there may be some problems with the limiting argument in [7]; see Appendix A.6.2 for further discussion.

We begin with the simpler case of the *unregularised* KGV[6]. The kernel generalised variance is given by

$$\mathcal{N}(\boldsymbol{x}, \boldsymbol{y}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}) = -\frac{1}{2} \log \left(\prod_i \left(1 - \rho_i^2\right) \right),$$

where in this case ρ_i are the stationary points with respect to $\mathbf{c}_i$ and $\mathbf{d}_i$ of

$$\rho_i = \frac{\mathbf{c}_i^\top \widetilde{\mathbf{K}}_{mm}^{(x)} \widetilde{\mathbf{K}}_{mm}^{(y)} \mathbf{d}_i}{\left(\mathbf{c}_i^\top \left(\widetilde{\mathbf{K}}_{mm}^{(x)} \right)^2 \mathbf{c}_i \right)^{1/2} \left(\mathbf{d}_i^\top \left(\widetilde{\mathbf{K}}_{mm}^{(y)} \right)^2 \mathbf{d}_i \right)^{1/2}}, \tag{4.2.19}$$

given by the generalised eigenvalue solutions of (3.4.6). Starting with the expression for $\hat{\rho}_i$ in (4.2.9), we use Theorem 4.2.2 to show

$$\hat{\mathbf{c}}_i^\top \mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)} \right)^\top \hat{\mathbf{c}}_i \quad \leq \quad \hat{\mathbf{c}}_i^\top \operatorname{diag} \left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)} \right)^\top \mathbf{1}_{l_x} \right) \hat{\mathbf{c}}_i. \tag{4.2.20}$$

Thus, replacing the right hand term with the left hand term in the denominator of (4.2.9) (and likewise for the term in $\hat{\mathbf{d}}_i$), we get

$$\tilde{\rho}_i = \frac{\hat{\mathbf{c}}_i^\top \left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(y)} \right)^\top - \frac{1}{m} \mathbf{K}_{lm}^{(x)} \mathbf{1}_m \mathbf{1}_m^\top \left(\mathbf{K}_{lm}^{(y)} \right)^\top \right) \hat{\mathbf{d}}_i}{\left(\hat{\mathbf{c}}_i^\top \mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)} \right)^\top \hat{\mathbf{c}}_i \right)^{1/2} \left(\hat{\mathbf{d}}_i^\top \mathbf{K}_{lm}^{(y)} \left(\mathbf{K}_{lm}^{(y)} \right)^\top \hat{\mathbf{d}}_i \right)^{1/2}}, \tag{4.2.21}$$

where $\tilde{\rho}_i \geq \hat{\rho}_i$. We then express this solution $\tilde{\rho}_i$ as a projection onto the respective grid matrices $\mathbf{Q}$ and $\mathbf{R}$ as in Theorem 4.2.3, and use similar reasoning to the previous section to prove $|\rho_i| \geq |\tilde{\rho}_i|$ for ρ_i in (4.2.19). Although the contrast function induced by the stationary points of (4.2.21) is never used in practice, it is nonetheless of interest to compute it: recalling the result of Lemma 3.4.5, we find

$$\mathcal{N}(\boldsymbol{x}, \boldsymbol{y}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}) = -\frac{1}{2} \log \left(1 \times \prod_{i=1}^{m-1} (1 - 1) \right) = \infty.$$

In other words, the approximation made in (4.2.20) is too loose: this is a helpful result, however, in that it provides a guideline to how loosely we may bound the terms in the denominator of (4.2.8) while still obtaining an applicable empirical contrast function. It follows that the regularised kernel canonical correlations (3.4.7), which are used to compute the regularised empirical estimate of the KGV [7], constitute a tighter approximation than that in (4.2.20). In particular, if we make the replacement

$$\hat{\mathbf{c}}_i^\top \operatorname{diag} \left(\mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)} \right)^\top \mathbf{1}_{l_x} \right) \hat{\mathbf{c}}_i \;\Rightarrow\; \hat{\mathbf{c}}_i^\top \left(\theta_1 \mathbf{K}_{lm}^{(x)} \left(\mathbf{K}_{lm}^{(x)} \right)^\top + \theta_2 \nu_{\boldsymbol{x}} \mathbf{K}_{ll}^{(x)} \right) \hat{\mathbf{c}}_i,$$

where $\theta_1 \geq 0$, $\theta_2 \geq 0$, and $\theta_1 + \theta_2 \leq 1$, we recover an expression which, for correct choice of θ_1, θ_2, yields the *regularised* KGV. We therefore expect the performance of both the KGV and KMI to be very similar when used for ICA: this is indeed the case, as seen in Section 6.

4.3 Multivariate KC and KMI

We now describe how our contrast function may be generalised to more than two random variables. Let us define the continuous univariate random variables $\mathsf{x}_1, \ldots, \mathsf{x}_n$ on $\mathcal{X}_1, \ldots, \mathcal{X}_n$ (which are assumed here to be bounded intervals in $\mathbb{R}$), and let $\boldsymbol{z} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be an i.i.d. sample of size

[6] We emphasise that only the regularised empirical estimate of the KGV is used in practice [7].

m from the joint distribution $\mathbf{P}_{\mathsf{x}_1,\ldots,\mathsf{x}_n}$. We also define the associated feature spaces $\mathcal{F}_{\mathcal{X}_1},\ldots,\mathcal{F}_{\mathcal{X}_n}$, each with its corresponding kernel (as in the 2 class case, the kernels may be different, and are identified by their argument).

We begin with a generalisation of the concept of normalised covariance. The derivation takes a similar form to that in [7, Appendix A.3], although Bach and Jordan deal with canonical correlations rather than normalised covariances, which changes the proof strategy in some respects. Let $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n$ be the vectors *of unit magnitude* in their respective feature spaces (as in the 2-variable normalised covariance), onto which the feature space vectors $\mathbf{x}_1,\ldots,\mathbf{x}_n$ respectively project. Defining the i,jth covariance between these projections as

$$\begin{aligned} y_{ij} &:= \mathbf{E}_{\mathbf{x}_i,\mathbf{x}_j}\left(\boldsymbol{\alpha}_i^\top \mathbf{x}_i \mathbf{x}_j^\top \boldsymbol{\alpha}_j\right) - \mathbf{E}_{\mathbf{x}_i}\left(\boldsymbol{\alpha}_i^\top \mathbf{x}_i\right)\mathbf{E}_{\mathbf{x}_j}\left(\boldsymbol{\alpha}_j^\top \mathbf{x}_j\right) \\ &= \boldsymbol{\alpha}_i^\top \mathbf{C}_{ij}\boldsymbol{\alpha}_j, \end{aligned}$$

then the eigenvalues λ of

$$\begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix} \begin{bmatrix} 0 & y_{12} & \ldots & y_{1n} \\ y_{21} & 0 & \ldots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \ldots & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_n \end{bmatrix} = \lambda \tag{4.3.1}$$

are *all* zero if and only if $y_{ij} = 0$ for all i,j; in other words, each pair of projections is uncorrelated. If the feature spaces are sufficiently rich RKHSs, this implies that the variables are pairwise independent using Theorem 3.2.4[7] . In fact, we need only determine that the eigenvalue with the largest *magnitude* is zero, which is less costly to compute (we are not guaranteed that this eigenvalue is positive). This gives us the multivariate contrast

$$\begin{aligned} \mathcal{I}\left(\mathbf{P}_{\mathsf{x}_1,\ldots,\mathsf{x}_n}, \mathcal{F}_{\mathcal{X}_1},\ldots,\mathcal{F}_{\mathcal{X}_n}\right) &:= \max_j \left(|\lambda_j|\right), \\ &= |\lambda_{\max}| \end{aligned}$$

It is instructive to compare with the KCC-based contrast for more than two variables, which uses the smallest eigenvalue of a matrix of correlations (with diagonal terms equal to one, rather than zero), where this correlation matrix has only positive eigenvalues.

This generalisation of the covariance can be rewritten in a form that more closely resembles the normalised covariance expression in (3.2.2), by rewriting (4.3.1) as the equivalent eigenvalue problem

$$\begin{bmatrix} a_1\boldsymbol{\alpha}_1^\top & a_2\boldsymbol{\alpha}_2^\top & \ldots & a_n\boldsymbol{\alpha}_n^\top \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{C}_{12} & \ldots & \mathbf{C}_{1n} \\ \mathbf{C}_{21} & \mathbf{0} & \ldots & \mathbf{C}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{n1} & \mathbf{C}_{n2} & \ldots & \mathbf{0} \end{bmatrix} \begin{bmatrix} a_1\boldsymbol{\alpha}_1 \\ a_2\boldsymbol{\alpha}_2 \\ \vdots \\ a_n\boldsymbol{\alpha}_n \end{bmatrix} = \lambda;$$

the eigenvectors retain unit norm due to the components $\boldsymbol{\alpha}_i$ having unit norm. We may find an empirical estimate of this quantity by writing each projection vector as $a_i\boldsymbol{\alpha}_i^\top = \mathbf{c}_{i,j}\mathbf{X}_i$, where $\mathbf{X}_i$ contains the observations $\boldsymbol{x}_i$ in its columns; this follows from the same argument that was used to obtain (3.2.3) from (3.2.2) in Section 3.2. Making this replacement, and using the same reasoning that was applied in deriving (3.2.5), we write the above in terms of Gram matrices of the observations, giving

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_2 & \ldots & \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_n \\ \widetilde{\mathbf{K}}_2\widetilde{\mathbf{K}}_1 & \mathbf{0} & \ldots & \widetilde{\mathbf{K}}_2\widetilde{\mathbf{K}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{K}}_n\widetilde{\mathbf{K}}_1 & \widetilde{\mathbf{K}}_n\widetilde{\mathbf{K}}_2 & \ldots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{1,j} \\ \mathbf{c}_{2,j} \\ \vdots \\ \mathbf{c}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \widetilde{\mathbf{K}}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \widetilde{\mathbf{K}}_n \end{bmatrix} \begin{bmatrix} \mathbf{c}_{1,j} \\ \mathbf{c}_{2,j} \\ \vdots \\ \mathbf{c}_{n,j} \end{bmatrix}, \tag{4.3.2}$$

[7] We require only pairwise independence to recover the independent sources in the case of linear ICA: see Theorem 2.1.2.

where $\widetilde{\mathbf{K}}_i = \widetilde{\mathbf{X}}_i^\top \widetilde{\mathbf{X}}_i = \mathbf{H}\mathbf{K}_i\mathbf{H}$; the matrix $\widetilde{\mathbf{X}}_i$ represents the centered feature space representation of the sample $\boldsymbol{x}_i$, and $\mathbf{K}_i$ the uncentered Gram matrix of $\boldsymbol{x}_i$.

We now describe a generalisation of the kernel mutual information to more than two variables. By analogy with the 2-variable case in Definition 4.2.4, we propose the contrast

$$\mathcal{M}\left(\boldsymbol{z}, \mathcal{F}_{\mathcal{X}_1}, \ldots, \mathcal{F}_{\mathcal{X}_n}\right) := -\frac{1}{2}\log \prod_{j=1}^{mn}\left(1+\breve{\lambda}_j\right), \tag{4.3.3}$$

where $\nu_{\boldsymbol{z}}\breve{\lambda}_j = \lambda_j$, and

$$\begin{aligned} \nu_{\boldsymbol{z}} &:= \min_{i\in\{1,\ldots,n\}} \nu_{\boldsymbol{x}_i}, \text{ where} \\ \nu_{\boldsymbol{x}_i} &:= \min_{j\,\in\,\{1\ldots m\}} \sum_{l=1}^{m} k\left(x_{i,l}, x_{i,j}\right); \end{aligned}$$

this additional scaling factor will be justified below, on both numerical and theoretical grounds. For (4.3.3) to be defined, it is necessary that $1+\breve{\lambda}_j > 0$ for all j, which is true near independence[8] . For this to be a reasonable choice of contrast function, we wish it to be zero if and only if all pairs of feature space projections are uncorrelated (*i.e.*, the kernel correlation must indicate at this point that the variables are independent). This may be shown via a minor adaptation of the corresponding proof in [7, Appendix A.2]. First, we may rewrite each factor $\breve{\lambda}_j + 1$ in (4.3.3) as the solution to

$$\begin{bmatrix} \mathbf{I} & \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_1^{1/2}\widetilde{\mathbf{K}}_2^{1/2} & \ldots & \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_1^{1/2}\widetilde{\mathbf{K}}_n^{1/2} \\ \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_2^{1/2}\widetilde{\mathbf{K}}_1^{1/2} & \mathbf{I} & \ldots & \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_2^{1/2}\widetilde{\mathbf{K}}_n^{1/2} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_n^{1/2}\widetilde{\mathbf{K}}_1^{1/2} & \nu_{\boldsymbol{z}}^{-1}\widetilde{\mathbf{K}}_n^{1/2}\widetilde{\mathbf{K}}_2^{1/2} & \ldots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \left(\breve{\lambda}_j + 1\right) \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix},$$

where $\mathbf{K}_i^{1/2}\mathbf{c}_{i,j} = \mathbf{d}_{i,j}$, bearing in mind that the determinant of the left hand matrix is the product of these eigenvalues. Since the left hand matrix is symmetric, the trace is equal to the sum of the eigenvalues (Theorem A.1.28). This means

$$\sum_{j=1}^{mn}\left(\breve{\lambda}_j + 1\right) = mn. \tag{4.3.4}$$

Assuming without loss of generality that the the mnth eigenvalue corresponds to $\breve{\lambda}_{\max} := \lambda_{\max}/\nu_{\boldsymbol{z}}$, we rewrite (4.3.3) as

$$\begin{aligned} -\frac{1}{2}\log\prod_{j=1}^{n}\left(1+\breve{\lambda}_j\right) &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{1}{2}\log\prod_{j=1}^{mn-1}\left(1+\breve{\lambda}_j\right) \\ &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\sum_{j=1}^{mn-1}\frac{1}{mn-1}\log\left(1+\breve{\lambda}_j\right) \\ &\geq -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\log\left(\frac{1}{mn-1}\sum_{j=1}^{mn-1}\left(1+\breve{\lambda}_j\right)\right) \\ &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\log\left(\frac{mn-\breve{\lambda}_{\max}-1}{mn-1}\right), \end{aligned}$$

[8] This is our first reason for introducing the factor $\nu_{\boldsymbol{z}}$, since it generally causes $\left|\breve{\lambda}_j\right| < |\lambda_j|$, which results in $\mathcal{M}\left(\boldsymbol{z}, \mathcal{F}_{\mathcal{X}_1}, \ldots, \mathcal{F}_{\mathcal{X}_n}\right)$ being defined further from independence. This is not the only such scaling factor: we provide further justification in due course.

where the penultimate line uses Jensen's inequality, and we substitute (4.3.4) in the final line. The function $-1/2\,(mn-1)\log\left((mn-\breve{\lambda}_{\max}-1)/(mn-1)\right)$ is convex with respect to $\breve{\lambda}_{\max}$, and its tangent at $\breve{\lambda}_{\max}+1=1$ is $\breve{\lambda}_{\max}/2$. Since a convex function is always greater than or equal to its tangent, we find

$$-\frac{1}{2}\log(1+\breve{\lambda}_{\max})-\frac{mn-1}{2}\log\left(\frac{mn-\breve{\lambda}_{\max}-1}{mn-1}\right)\geq-\frac{1}{2}\log(1+\breve{\lambda}_{\max})+\frac{\breve{\lambda}_{\max}}{2}.$$

The left hand side is convex, and has a global minimum at $\breve{\lambda}_{\max}=0$. It follows that (4.3.3) is likewise minimised at $\mathcal{M}\left(\boldsymbol{z},\mathcal{F}_{\mathcal{X}_1},\ldots,\mathcal{F}_{\mathcal{X}_n}\right)=0$ (at which point $\breve{\lambda}_j+1=1$ for all j), and that this corresponds to the point at which all pairs of covariances are zero, using (4.3.1).

We now briefly outline how the contrast function in (4.3.3) relates to the KL divergence (Definition A.8.11 in Section A.8), which is zero if and only if the random variables are pairwise independent [24]. In the case of a Gaussian random vector $\mathbf{x}_G$, which can be segmented as $\mathbf{x}_G^\top:=\left[\begin{array}{ccc}\mathbf{x}_{G,1}^\top & \cdots & \mathbf{x}_{G,n}^\top\end{array}\right]$, the KL divergence between the joint distribution of $\mathbf{x}_G$ and the product of the marginal distributions of the $\mathbf{x}_{G,i}$ can be written in terms of the covariance matrices as

$$D_{\mathrm{KL}}\left(\mathbf{f}_{\mathbf{x}_G}\left\|\prod_{i=1}^{n}\mathbf{f}_{\mathbf{x}_{G,i}}\right.\right)=-\frac{1}{2}\log\left(\frac{|\mathbf{C}|}{\prod_{i=1}^{n}|\mathbf{C}_{ii}|}\right),$$

where

$$\begin{aligned}\mathbf{C}&=\mathbf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G\mathbf{x}_G^\top\right)-\mathbf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G\right)\mathbf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G^\top\right),\\ \mathbf{C}_{ii}&=\mathbf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}\mathbf{x}_{G,i}^\top\right)-\mathbf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}\right)\mathbf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}^\top\right).\end{aligned}$$

These results should allow us to generalise the reasoning in Section 4.1, then substitute the kernel density estimates

$$\begin{aligned}\widehat{\mathbf{P}}_{\mathrm{x}_i}(x_i)&=\frac{1}{m}\sum_{l=1}^{m}k\left(x_{i,l}-x_i\right),\\ \widehat{\mathbf{P}}_{\mathrm{x}_1,\ldots,\mathrm{x}_n}(x_1,\ldots,x_n)&=\frac{1}{m}\sum_{l=1}^{m}\prod_{i=1}^{n}k\left(x_{i,l}-x_i\right),\end{aligned}$$

and apply the bounding technique of Section 4.2, to obtain the contrast in (4.3.3); this is the second reason for our choosing $\nu_{\boldsymbol{z}}$ to scale $\breve{\lambda}_j$. The details of this generalisation are beyond the scope of the present work.

Chapter 5

Implementation issues

In this chapter, we describe the manner in which the KC and KMI contrasts may be used to solve the instantaneous linear ICA problem described in Section 2. This implementation comprises two components: the efficient computation of the KC and KMI contrasts, using low rank approximations of the Gram matrices; and gradient descent on the space of orthogonal matrices $\mathbf{W}$, which follows whitening in our determination of the inverse of the mixing matrix $\mathbf{B}$ (see Section 2.2.1). These results are summarised from the more detailed discussion in [7] (although the low rank decomposition is in our case made easier by the absence of the variance term used in the KCC and KGV contrasts).

5.1 Efficient contrast computation

We note that the KC requires us to determine the eigenvalue of maximum magnitude for an $mn \times mn$ matrix (see (4.3.2)), and the KMI is a determinant of an $mn \times mn$ matrix, as specified in (4.3.3). For any reasonable sample size m, the cost of these computations is prohibitive. We now describe how the computational complexity of this problem may be substantially reduced. First, we note that any positive (semi)definite matrix can be written $\mathbf{K}_i = \mathbf{Z}_i \mathbf{Z}_i^\top$, where $\mathbf{Z}_i$ is lower triangular; this is known as the Cholesky decomposition. If the eigenvalues of the Gram matrix $\mathbf{K}_i$ decay sufficiently rapidly, however, we may make the approximation

$$\mathbf{K}_i \approx \mathbf{Z}_i \mathbf{Z}_i^\top$$

to the Gram matrix $\mathbf{K}_i$, where $\mathbf{Z}_i$ is an $m \times d_i$ matrix; the error due to this approach may be measured via the maximum eigenvalue μ_i of $\mathbf{K}_i - \mathbf{Z}_i \mathbf{Z}_i^\top$. The $\mathbf{Z}_i$ are determined via an *incomplete* Cholesky decomposition, in which the smaller pivots are skipped; symmetric permutation of the rows and columns of $\mathbf{K}_i$ is used in the course of this process to increase the accuracy and numerical stability of the approximation. This method is applied in [34] to decrease the storage and computational requirements of interior point methods used in SVMs, as well as in [7] for faster computation of the KMI and KCC contrasts (pseudocode algorithms may be found in both these references). Once the incomplete Cholesky decomposition is accomplished, we can compute the approximate *centered* Gram matrices according to $\widetilde{\mathbf{K}}_i := \mathbf{H}\mathbf{K}_i\mathbf{H} = (\mathbf{H}\mathbf{Z}_i)\left(\mathbf{H}\mathbf{Z}_i^\top\right) = \widetilde{\mathbf{Z}}_i \widetilde{\mathbf{Z}}_i^\top$.

We now show how this low rank decomposition may be used to more efficiently compute the kernel covariance in (4.3.2). Substituting

$$\mathbf{d}_{i,j} = \widetilde{\mathbf{Z}}_i^\top \mathbf{c}_{i,j},$$

we get

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{Z}}_1\widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_2 & \dots & \widetilde{\mathbf{Z}}_1\widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_n \\ \widetilde{\mathbf{Z}}_2\widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_2\widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{Z}}_n\widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_n\widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_n \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix}$$

We may premultiply both sides by[1] $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top & \dots & \widetilde{\mathbf{Z}}_n^\top \end{bmatrix}\right)$ without increasing the nullspace of this generalised eigenvalue problem, and we then eliminate $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_1 & \dots & \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_n \end{bmatrix}\right)$ from both sides. Making these changes, we are left with

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_2 & \dots & \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_n \\ \widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix}, \tag{5.1.1}$$

which is a much more tractable eigenvalue problem, having dimension $\sum_{i=1}^n d_i$. The same procedure may easily be used to recast (4.3.3) as the determinant of an $(\sum_{i=1}^n d_i) \times (\sum_{i=1}^n d_i)$ matrix. We now briefly consider how to choose the rank d_i for a given precision μ_i: this depends on both the density $\mathbf{f}_{\mathsf{x}_i}$ and the kernel $k(x_i - x)$. For Gaussian kernels and densities with exponential decay rates, it is shown in [7] that the required precision relates to the rank according to $d_i = O(\log(m/\mu_i))$, which demonstrates the slow increase in rank with sample size. In the case of the KGV and KCC, however, the form of the contrast causes eigenvalues less than approximately $10^{-3}m\kappa/2$ to be discarded, which thus serves as a target precision to ensure the $\mathbf{Z}_i$ retain constant rank regardless of m. We also adopt this threshold in our simulations with the Gaussian kernel, although our motivation is purely a reduction of computational cost.

5.2 Gradient descent on the Stiefel manifold

We now describe the method used to minimise our kernel contrast functions over possible choices of the orthogonal demixing matrix $\mathbf{W}$ (the whitening process having been accomplished). The manifold described by $n \times p$ matrices $\mathbf{A}$ for which $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$, where $n \geq p$, is known as the *Stiefel manifold*. Gradient descent for functions defined on this manifold is described in [33]. A clear and intuitive explanation of this procedure is also given in [52], which was kindly provided to the authors by Hyvärinen, and which along with [7] constitutes the basis for the present description. Let $f(\mathbf{W}, \tilde{\boldsymbol{t}})$ be the particular contrast function (KC or KMI) on which we wish to do gradient descent, where $\tilde{\boldsymbol{t}} := (\tilde{\mathbf{t}}_1, \dots, \tilde{\mathbf{t}}_m)$ are the whitened, mixed observations (see Section 2.2.1). A naïve gradient descent algorithm would involve computing the derivative

$$\mathbf{G} := \frac{\partial f(\mathbf{W}, \tilde{\boldsymbol{t}})}{\partial \mathbf{W}},$$

updating $\mathbf{W}$ according to $\mathbf{W} \to \mathbf{W} + \mu\mathbf{G}$ (where μ is chosen to minimise $f(\mathbf{W} + \mu\mathbf{G}, \tilde{\boldsymbol{t}})$), and projecting the resulting matrix back onto the Stiefel manifold. This might not be particularly efficient, however, in that the update can largely be canceled by the subsequent projection operation. Instead, we attempt to find the direction of steepest descent on the Stiefel manifold, and to perform our update with the constraint that we remain on this manifold. To achieve this, we first describe the set of perturbations to $\mathbf{W}$ that retain the orthogonality of $\mathbf{W}$, then choose the direction of

[1] The notation $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top & \dots & \widetilde{\mathbf{Z}}_n^\top \end{bmatrix}\right)$ defines a matrix with blocks $\widetilde{\mathbf{Z}}_i^\top$ along the diagonal, and zeros elsewhere. The matrix need not be square, however, and the diagonal is in this case defined in a manner consistent with the asymmetry of the $\widetilde{\mathbf{Z}}_i^\top$.

steepest descent/ascent within this set, and finally give the expression that parameterises the shifts along the geodesic[2] in this direction.

Let $\mathbf{\Delta}$ be a perturbation with small norm to the orthogonal matrix $\mathbf{W}$, such that $\mathbf{W}+\mathbf{\Delta}$ remains on the Stiefel manifold. For this constraint to hold, we require

$$(\mathbf{W}+\mathbf{\Delta})^\top(\mathbf{W}+\mathbf{\Delta}) = \mathbf{I}, \text{ which implies} \tag{5.2.1}$$

$$\mathbf{W}^\top\mathbf{\Delta}+\mathbf{\Delta}^\top\mathbf{W} \approx \mathbf{0}; \tag{5.2.2}$$

in other words, $\mathbf{W}^\top\mathbf{\Delta}$ is skew-symmetric. To find the particular $\mathbf{\Delta}$ that gives the direction of steepest change of $f(\mathbf{W},\tilde{\boldsymbol{t}})$, we solve

$$\mathbf{\Delta}_{\max} := \arg\max_{\mathbf{\Delta}} f(\mathbf{W}+\mathbf{\Delta},\tilde{\boldsymbol{t}}),$$

subject to $\mathrm{tr}(\mathbf{\Delta}^\top\mathbf{\Delta}) = \text{const}$ and (5.2.2). This yields

$$\mathbf{\Delta}_{\max} = \mathbf{G} - \mathbf{W}\mathbf{G}^\top\mathbf{W},$$

where the proof may be found in [33, 52]. Finally, if we use q to parameterise displacement along a geodesic in the direction $\mathbf{\Delta}_{\max}$ from an initial matrix $\mathbf{W}(0)$, then the resulting $\mathbf{W}(q)$ is given by

$$\mathbf{W}(q) = \mathbf{W}(0)\exp\left(q\mathbf{W}(0)^\top\mathbf{\Delta}_{\max}\right).$$

The ICA algorithm was implemented by modifying the Matlab code in [5] to use our KC and KMI contrast functions. Consequently, we determine an approximation of the gradient of our contrast $f(\mathbf{W},\tilde{\boldsymbol{t}})$ by making small perturbations to $\mathbf{W}$ about each possible Jacobi rotation, and recomputing the contrast for each such perturbation. Gradient descent is then accomplished using a Golden search along this direction of steepest descent. It is interesting to note that performance was empirically found to improve, for *all* the kernel algorithms, when a grid search was used over the Jacobi angles that parameterise $\mathbf{W}$ (although careful tuning of the tolerance used in the Golden search can reduce this difference). This becomes impractical for large numbers of sources, however.

Finally, we note that procedures are given in [33] to compute the Hessian on the Stiefel manifold, as are the implementations of Newton's method and conjugate gradient descent. In addition, an adaptive algorithm for gradient descent on the Stiefel manifold is proposed in [90]. The application of these methods to improve the performance of our algorithm is beyond the scope of the present work.

[2] A geodesic represents the shortest path on a manifold between two points; equivalently, the acceleration involved in moving between two points along a geodesic is perpendicular to the manifold when constant velocity is maintained.

Chapter 6

Experimental results and conclusions

In this chapter, we examine the performance of our contrast functions (KC, KMI) as it compares to the KGV and KCC methods, when used to address the problem of linear instantaneous ICA. Since the objective is to find an estimate $\mathbf{V} := \mathbf{WQ}$ of the *inverse* of the mixing matrix $\mathbf{B}$ (the reader is referred to Section 2.2.1 for a detailed description of the ICA problem), we require a measure of distance between our approximation and the true inverse: this is given by the *Amari divergence*, which is introduced in Section 6.1. Next, we separate a range of artificial signals mixed using randomly generated matrices, including cases in which the observations were corrupted by noise. Results are compared with those obtained using fast ICA, Jade, and the extended Infomax algorithm, as well as the KCC and KGV. Finally, we artificially mix a range of audio signals representing a number of musical genres, and attempt to separate these. We end the chapter with some general observations regarding our results, and give suggestions for further study.

6.1 Measurement of performance

We use the Amari divergence, defined in [4], as an index of ICA algorithm performance.

Definition 6.1.1 (Amari divergence). Let $\mathbf{A}$ and $\mathbf{C}$ be two $n \times n$ matrices, and let $\mathbf{B} = \mathbf{AC}^{-1}$. Then the Amari divergence between $\mathbf{A}$ and $\mathbf{C}$ is

$$\mathcal{D}(\mathbf{A}, \mathbf{C}) = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{n} |b_{i,j}|}{\max_j |b_{i,j}|} - 1 \right) + \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left(\frac{\sum_{i=1}^{n} |b_{i,j}|}{\max_i |b_{i,j}|} - 1 \right).$$

Although this measure is not, strictly speaking, a distance metric for general matrices $\mathbf{A}, \mathbf{C}$, it nonetheless possesses certain useful properties, as shown below.

Lemma 6.1.2 (Properties of the Amari divergence). *The Amari divergence $\mathcal{D}(\mathbf{A}, \mathbf{C})$ between the $n \times n$ matrices $\mathbf{A}, \mathbf{C}$ has the following properties:*

- $\mathcal{D}(\mathbf{A}, \mathbf{C}) \geq 0$, *with equality when $\mathbf{A}$ is equal to $\mathbf{C}$, or some permitted scaling or permutation thereof (as described below).*
- $0 \leq \mathcal{D}(\mathbf{A}, \mathbf{C}) \leq 1$.
- *Let $\mathbf{P}_1, \mathbf{P}_2$ be arbitrary permutation matrices, and r, s be arbitrary non-zero scaling factors. Then $\mathcal{D}(\mathbf{A}, \mathbf{C}) = \mathcal{D}(\mathbf{A}(r\mathbf{P}_1), \mathbf{C}(s\mathbf{P}_2))$: the Amari divergence is invariant with respect to scaling and* column *swaps. The Amari divergence is generally* not *invariant with respect to row swaps, or with respect to scaling of rows or columns by different amounts.*

The final property in the above Lemma is particularly useful in the context of ICA, since it causes our performance measure to be invariant to the output ordering ambiguity (see Theorem 2.1.2).

6.2 Experiments and performance assessment

We now describe the experiments conducted to verify the performance of our algorithms. Since the main purpose is to compare the performance with that reported in [7], we tried as far as possible to generate our test distributions in the manner described therein, besides our near-Gaussianity experiment and certain additional investigations we performed. A list of the distributions used in our experiments, and their respective kurtoses, is given in Table 6.1. While these distributions represent a broad range of behaviours, we note that negative kurtoses predominate, which should be borne in mind when evaluating performance. We used the KGV and KCC Matlab implementations downloadable from [5], which we also modified to implement the KMI and KC. The precision of the incomplete Cholesky decomposition, used to approximate the Gram matrices for the Kernel contrasts, was set at $\eta := \epsilon n$; our choice of ϵ represents a tradeoff between accuracy and computation speed. Unless otherwise specified, the kernel contrast results were refined in a *polishing step*, in which the kernel size was halved upon convergence, and the gradient descent procedure recommenced with this smaller kernel. This polishing usually caused a measurable improvement in the results.

We also follow [7] in providing results from FastICA [35], Jade [18], and the extended Infomax algorithm [58], as an additional check of our algorithm performance. In the case of fast ICA, the default (kurtosis based) nonlinearity was used except for sources specifically unsuited to it, in which cases we signal our alternative choice of nonlinearity (the predominantly negative kurtoses in Table 6.1 make this a good choice: see Section 2.2.3). The Jade and Infomax algorithms were likewise used in their default configurations.

We begin with a brief investigation into the form taken by the various kernel contrast functions for a selection of the data in Table 6.1. Contours of the KGV, KC, KMI, and Amari divergence are plotted in Figure 6.2.1, which describes the demixing of samples from three distributions, combined using a product of known Jacobi rotations. All kernel based contrasts in this demonstration were computed with a Gaussian RBF kernel,

$$k_G(x, x') = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}\|x - x'\|^2). \tag{6.2.1}$$

We observe that each of the contrasts exhibits local minima at locations distant from independence, but that each possesses a "basin of attraction" in the vicinity of the correct answer. Moreover, we note that each of the contrasts is smooth (given the choice of kernel size), and that the global minima are fairly symmetric. For these reasons, the gradient descent algorithm described in Section 5.2 should converge rapidly to the global optimum, given a reasonable initialisation point. Our solution method differs from [7], however, in that we generally use Jade (unless specified otherwise) to initialise the kernel based contrast functions (KC, KCC, KGV, KMI), whereas Bach and Jordan only do this when separating large numbers of signals. Initialisation is accomplished in [7] using a one-unit kernel contrast with deflation, and with a less costly polynomial kernel. For more than two signals, this process is repeated several times, starting from different initialising matrices. While Jade is less computationally costly as an initialisation method, it might be less reliable in certain cases (where the sources are near-Gaussian, or when a large number of outliers exist due to noise, both of which can cause Jade to misconverge).

6.2.1 General mixtures of artificial data

We now describe the ICA experiments performed with the distributions in Table 6.1, where the Amari divergence is used to measure the closeness of the estimated mixing matrix to the true matrix. Kernels used include the Gaussian RBF kernel in (6.2.1), and the Laplace kernel,

$$k_L(x, x') = \frac{\lambda}{2} \exp(-\lambda\|x - x'\|).$$

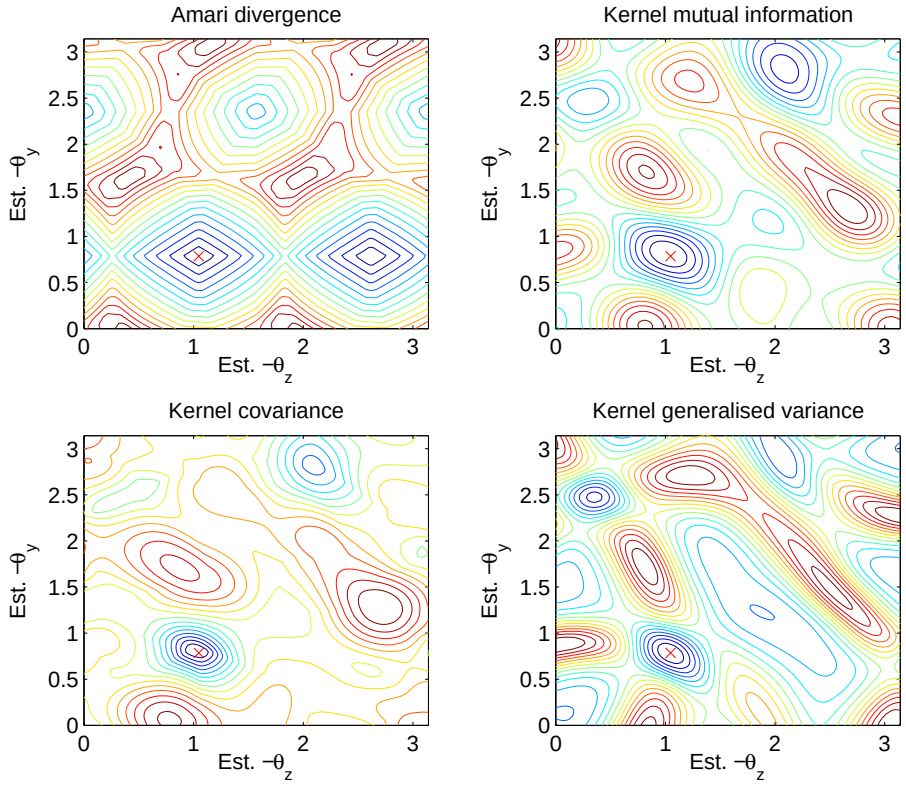

Figure 6.2.1: Contour plots of kernel contrast functions. Top left: Amari divergence. Top right: kernel mutual information. Bottom left: kernel covariance. Bottom right: kernel generalised variance. Three signals of length 1000 and with respective distributions g, k, and q (this choice was random) were combined using a 3×3 orthogonal rotation matrix. This matrix was expressed as a product of Jacobi rotations $\mathbf{A} = \mathbf{R}_{\theta_z}\mathbf{R}_{\theta_y}\mathbf{R}_{\theta_x}$, where $\theta_x = -\pi/6$, $\theta_y = -\pi/4$, and $\theta_z = -\pi/3$; the subscript of the angle denotes the axis about which the rotation occurs. An estimate $\mathbf{W} = \mathbf{R}_{-\theta_x}\mathbf{R}_{\hat{\theta}_y}\mathbf{R}_{\hat{\theta}_z}$ of $\mathbf{A}^{-1}$ was made, in which $\hat{\theta}_y$ and $\hat{\theta}_z$ took values in the range $[0, \pi]$. The red "x" in each plot is located at the coordinates $(-\theta_z, -\theta_y)$ corresponding to the optimal estimate of $\mathbf{A}$. A Gaussian kernel of size $\sigma^2 = 1$ was used in all cases, and $\kappa = 10^{-3}$ for the KGV.

Table 6.1: Labels of distributions used, and their respective kurtoses. All distributions have zero mean and unit variance.

Label	Definition	Kurtosis
a	Student's t distribution, 3 DOF	∞
b	Double exponential	3.00
c	Uniform	-1.20
d	Students's t distribution, 5 DOF	6.00
e	Exponential	6.00
f	Mixture, 2 double exponentials	-1.70
g	Symmetric mixture 2 Gauss., multimodal	-1.85
h	Symmetric mixture 2 Gauss., transitional	-0.75
i	Symmetric mixture 2 Gauss., unimodal	-0.50
j	Asymm. mixture 2 Gauss., multimodal	-0.57
k	Asymm. mixture 2 Gauss., transitional	-0.29
l	Asymm. mixture 2 Gauss., unimodal	-0.20
m	Symmetric mixture 4 Gauss., multimodal	-0.91
n	Symmetric mixture 4 Gauss., transitional	-0.34
o	Symmetric mixture 4 Gauss., unimodal	-0.40
p	Asymm. mixture 4 Gauss., multimodal	-0.67
q	Asymm. mixture 4 Gauss., transitional	-0.59
r	Asymm. mixture 4 Gauss., unimodal	-0.82

We combined the independent sources using random mixing matrices, with condition numbers between 1 and 2, and then whitened the resulting observations before estimating the orthogonal de-mixing matrix[1]. We first present the results obtained by de-mixing two independently generated samples from the same distribution, where each distribution in Table 6.1 was investigated. Results for samples of length 256 are given in Table 6.2, and those for samples of length 1024 in Table 6.3. The average performance in the 256 sample case is best in the case of the KGV, followed by the KMI with the Laplace and Gauss kernels. In addition, the KC and KCC methods outperform the remaining algorithms, despite being based only on a single eigenvalue. It is notable that the extended Infomax method performs badly in every example, which appears due to the small number of observations. In the 1024 sample case, the hierarchy in algorithm performance is retained, although the KGV and KCC outperform the KMI and KC with smaller margin. It is notable that the difference between the KMI with Laplace kernel and that with Gaussian kernel is large for the distributions *a,b,d,e*, all of which are heavy tailed (indeed, the Laplace kernel yields the best performance for double exponential and exponential distributions): this is not surprising, in the sense that a Parzen window estimate of a heavy tailed distribution can be accomplished efficiently with a heavy tailed kernel. We also note that distribution *o* (the symmetric, unimodal mixture of 4 Gaussians) is now accurately demixed, which was not possible with 256 samples.

Our second experiment consisted in de-mixing data drawn independently from several distributions chosen at random from Table 6.1. Results are given in Table 6.4. We note that the KMI with Gaussian kernel outperforms the KGV in the final three experiments, and the KMI with Laplace kernel yields best overall performance in five of the seven experiments. On the other hand, the KGV performs best in the first and third case, where the number m of samples is small. This behaviour is akin to the large performance lead of the KGV in Table 6.2, compared with the smaller difference in Table 6.3, and suggests that the KGV performs better for a smaller number of samples, at the expense of performance at large sample sizes.

A more detailed look at the results behind these averages, as illustrated in Figure 6.2.2, shows that in rare cases (at most 1-2 instances per experiment) the Amari divergences can be substantially greater than the mean, which occasionally (but not always) corresponds to poor convergence in

[1]We did not use a simple orthogonal matrices to mix our sources, since this would lower the variance in our estimate of $\mathbf{W}$, making the problem (slightly) easier than that of estimating a truly random mixing matrix [20].

Jade (which is used to initialise the optimisation process for the kernel contrasts). The effect is most pronounced in the $n = 4, m = 1000$ and $n = 8, m = 2000$ cases, and is absent from the $n = 4, m = 4000$ and $n = 8, m = 4000$ cases. This suggests that these misconvergences are due to local minima in the contrast (caused by insufficient observations relative to the number of sources), and not necessarily an artefact due to poor initialisation. Finally, the extended Infomax algorithm seems unable to separate the signals in 250 sample, 2 signal case: the Amari error was spread almost uniformly over the range $[0, 100]$. While the Lapalce kernel clearly gives superior performance, this comes at an increased computational cost, since the eigenvalues of the associated Gram matrices decay more slowly than for the Gaussian kernel, necessitating the use of a higher rank in the incomplete Cholesky decomposition to maintain good performance.

Table 6.2: Average Amari divergence over 100 runs, for mixtures of two i.i.d. samples of length 256, both drawn from the corresponding distributions in Table 6.1. The Laplace kernel had size $\sigma = 2.5$, with associated $\eta = 0.1$, while the Gaussian kernel used in the KMI and KC cases had size $\sigma = 1.0$ and $\eta = 2 \times 10^{-4}$. The KCC and KGV had Gaussian kernels of size 1.0, $\eta = 2 \times 10^{-4}$, and $\kappa = 2 \times 10^{-2}$. The final row gives the average over all 18 experiments.

	Fica	Jade	Imax	KCC	KC(Gauss)	KC(Lapl)	KGV	KMI(Gauss)	KMI(Lapl)
a	7.3 ± 0.7	$\mathbf{6.0 \pm 0.5}$	44.4 ± 2.9	9.8 ± 0.8	12.2 ± 1.1	9.5 ± 0.9	7.1 ± 0.6	9.2 ± 0.9	6.6 ± 0.6
b	10.4 ± 1.2	7.8 ± 0.7	42.3 ± 2.8	9.1 ± 0.6	9.8 ± 0.7	8.2 ± 0.7	$\mathbf{6.2 \pm 0.4}$	6.9 ± 0.4	6.9 ± 0.5
c	4.3 ± 0.3	$\mathbf{3.2 \pm 0.2}$	45.8 ± 2.9	5.7 ± 0.4	5.8 ± 0.4	4.8 ± 0.3	3.7 ± 0.2	3.7 ± 0.2	3.7 ± 0.3
d	12.1 ± 1.3	$\mathbf{10.7 \pm 1.1}$	44.0 ± 2.9	17.2 ± 1.4	24.1 ± 2.0	19.8 ± 1.8	13.8 ± 1.1	17.8 ± 1.5	16.4 ± 1.5
e	11.7 ± 1.3	10.0 ± 1.1	44.7 ± 2.8	3.9 ± 0.2	4.5 ± 0.2	4.0 ± 0.2	$\mathbf{3.4 \pm 0.2}$	4.3 ± 0.2	4.2 ± 0.2
f	3.7 ± 0.2	3.0 ± 0.2	43.5 ± 2.8	$\mathbf{2.8 \pm 0.2}$	3.0 ± 0.2	2.9 ± 0.2	$\mathbf{2.8 \pm 0.2}$	2.9 ± 0.2	2.9 ± 0.2
g	3.0 ± 0.2	$\mathbf{2.5 \pm 0.2}$	47.8 ± 3.0	2.6 ± 0.2	$\mathbf{2.5 \pm 0.2}$	$\mathbf{2.5 \pm 0.2}$	$\mathbf{2.5 \pm 0.2}$	$\mathbf{2.5 \pm 0.2}$	$\mathbf{2.5 \pm 0.2}$
h	10.6 ± 0.9	$\mathbf{7.3 \pm 0.5}$	42.3 ± 2.8	10.6 ± 0.8	12.6 ± 1.0	11.8 ± 1.0	9.4 ± 0.6	10.0 ± 0.7	9.7 ± 0.8
i	19.0 ± 1.6	$\mathbf{13.5 \pm 1.2}$	43.3 ± 3.0	17.7 ± 1.5	24.6 ± 2.3	25.9 ± 2.4	22.0 ± 1.9	22.8 ± 2.0	22.0 ± 2.0
j	10.0 ± 1.1	9.2 ± 1.4	45.6 ± 2.9	$\mathbf{2.6 \pm 0.2}$	2.9 ± 0.2	2.9 ± 0.2	3.5 ± 1.0	2.9 ± 0.2	2.9 ± 0.2
k	34.4 ± 2.8	32.5 ± 2.5	48.6 ± 2.9	6.8 ± 0.5	6.5 ± 0.5	6.3 ± 0.4	$\mathbf{5.5 \pm 0.3}$	6.1 ± 0.4	6.1 ± 0.4
l	40.7 ± 2.8	40.8 ± 2.8	47.7 ± 3.0	11.1 ± 0.9	11.7 ± 0.9	10.1 ± 0.7	10.4 ± 0.8	9.9 ± 0.7	$\mathbf{9.6 \pm 0.6}$
m	6.8 ± 0.5	4.9 ± 0.3	45.9 ± 2.9	5.6 ± 0.4	15.1 ± 1.8	14.0 ± 2.3	$\mathbf{2.7 \pm 0.2}$	4.5 ± 0.3	3.4 ± 0.2
n	28.5 ± 2.5	26.1 ± 2.4	45.2 ± 2.8	6.4 ± 0.7	6.1 ± 0.4	7.4 ± 0.8	$\mathbf{3.8 \pm 0.2}$	5.2 ± 0.3	5.0 ± 0.3
o	25.2 ± 2.0	$\mathbf{23.1 \pm 1.9}$	44.9 ± 2.8	28.5 ± 2.4	25.0 ± 2.1	36.8 ± 3.0	$\mathbf{23.2 \pm 2.7}$	$\mathbf{23.1 \pm 2.6}$	26.4 ± 2.9
p	13.1 ± 1.1	9.4 ± 0.9	42.6 ± 2.6	5.0 ± 0.5	10.2 ± 1.6	13.8 ± 2.4	$\mathbf{2.9 \pm 0.2}$	4.2 ± 0.2	4.4 ± 0.8
q	13.4 ± 1.2	11.1 ± 1.1	47.5 ± 2.8	10.5 ± 0.8	11.7 ± 0.8	10.5 ± 0.7	$\mathbf{8.0 \pm 0.6}$	8.9 ± 0.6	8.4 ± 0.7
r	9.4 ± 0.8	$\mathbf{6.4 \pm 0.4}$	50.1 ± 2.9	9.2 ± 0.7	11.9 ± 1.0	10.3 ± 1.0	6.7 ± 0.5	8.3 ± 0.6	8.0 ± 0.6
	$\mathbf{14.4 \pm 0.4}$	$\mathbf{12.6 \pm 0.4}$	$\mathbf{45.3 \pm 0.7}$	$\mathbf{9.2 \pm 0.3}$	$\mathbf{11.1 \pm 0.3}$	$\mathbf{11.2 \pm 0.4}$	$\mathbf{7.6 \pm 0.3}$	$\mathbf{8.5 \pm 0.3}$	$\mathbf{8.3 \pm 0.3}$

Table 6.3: Average Amari divergence over 100 runs, for mixtures of two i.i.d. samples of length 1024, both drawn from the corresponding distributions in Table 6.1. The Laplace kernel had size $\sigma = 2$, with associated $\eta = 0.05$, while the Gaussian kernel used in the KMI and KC cases had size $\sigma = 1.0$ and $\eta = 2 \times 10^{-4}$. The KCC and KGV had Gaussian kernels of size 1.0, $\eta = 2 \times 10^{-4}$, and $\kappa = 2 \times 10^{-2}$. The final row gives the average over all 18 experiments.

	Fica	Jade	Imax	KCC	KC(Gauss)	KC(Lapl)	KGV	KMI(Gauss)	KMI(Lapl)
a	4.4 ± 0.4	3.7 ± 0.3	$\mathbf{3.4 \pm 0.5}$	5.1 ± 0.3	5.7 ± 0.5	5.2 ± 0.4	3.9 ± 0.3	4.6 ± 0.3	4.3 ± 0.3
b	6.5 ± 0.6	5.5 ± 0.5	4.6 ± 0.5	3.8 ± 0.2	3.9 ± 0.2	3.3 ± 0.2	3.0 ± 0.2	3.1 ± 0.2	$\mathbf{2.9 \pm 0.2}$
c	2.5 ± 0.2	$\mathbf{1.7 \pm 0.1}$	2.2 ± 0.1	3.0 ± 0.2	2.5 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	2.1 ± 0.1	1.8 ± 0.1
d	7.3 ± 0.6	$\mathbf{6.2 \pm 0.4}$	10.0 ± 1.4	8.6 ± 0.7	12.4 ± 1.1	10.1 ± 0.9	6.9 ± 0.6	8.3 ± 0.7	7.4 ± 0.6
e	6.0 ± 0.8	4.3 ± 0.4	7.1 ± 1.6	1.9 ± 0.1	2.3 ± 0.2	2.3 ± 0.1	1.6 ± 0.1	1.8 ± 0.1	$\mathbf{1.5 \pm 0.1}$
f	1.9 ± 0.1	1.5 ± 0.1	$\mathbf{1.2 \pm 0.1}$	1.4 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	1.4 ± 0.1	1.4 ± 0.1	1.4 ± 0.1
g	1.7 ± 0.1	1.5 ± 0.1	$\mathbf{1.1 \pm 0.1}$	1.5 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	1.5 ± 0.1
h	5.3 ± 0.3	3.7 ± 0.3	3.4 ± 0.2	4.1 ± 0.3	3.8 ± 0.3	3.7 ± 0.2	$\mathbf{3.3 \pm 0.2}$	3.5 ± 0.2	3.5 ± 0.2
i	9.2 ± 0.7	$\mathbf{6.7 \pm 0.6}$	8.8 ± 1.2	9.1 ± 0.7	10.2 ± 0.9	10.2 ± 1.0	7.5 ± 0.7	8.0 ± 0.6	7.3 ± 0.5
j	4.8 ± 0.4	3.7 ± 0.4	35.9 ± 3.4	1.4 ± 0.1	1.8 ± 0.1	2.0 ± 0.1	$\mathbf{1.3 \pm 0.1}$	1.4 ± 0.1	1.4 ± 0.1
k	16.3 ± 1.6	15.3 ± 1.8	35.5 ± 2.9	2.8 ± 0.2	2.9 ± 0.2	2.9 ± 0.2	$\mathbf{2.1 \pm 0.1}$	2.4 ± 0.1	2.4 ± 0.1
l	26.2 ± 2.3	23.6 ± 2.2	36.8 ± 2.9	4.8 ± 0.4	4.6 ± 0.4	4.0 ± 0.3	$\mathbf{3.5 \pm 0.3}$	3.9 ± 0.3	3.8 ± 0.3
m	3.1 ± 0.2	2.3 ± 0.1	3.8 ± 0.2	2.5 ± 0.1	3.5 ± 0.2	1.9 ± 0.1	$\mathbf{1.4 \pm 0.1}$	1.5 ± 0.1	$\mathbf{1.4 \pm 0.1}$
n	9.6 ± 0.9	8.1 ± 0.8	51.0 ± 2.9	3.1 ± 0.4	3.7 ± 0.3	2.8 ± 0.2	$\mathbf{2.0 \pm 0.1}$	2.1 ± 0.1	2.1 ± 0.1
o	9.4 ± 0.9	7.5 ± 0.7	38.3 ± 2.5	9.6 ± 1.2	8.8 ± 0.9	11.6 ± 1.5	$\mathbf{4.8 \pm 0.7}$	6.0 ± 1.0	5.9 ± 0.8
p	4.5 ± 0.4	3.2 ± 0.2	7.7 ± 1.1	1.9 ± 0.1	2.8 ± 0.2	2.1 ± 0.1	$\mathbf{1.5 \pm 0.1}$	$\mathbf{1.5 \pm 0.1}$	$\mathbf{1.5 \pm 0.1}$
q	6.1 ± 0.5	4.1 ± 0.3	8.8 ± 1.3	5.4 ± 0.4	5.7 ± 0.4	4.9 ± 0.3	$\mathbf{3.3 \pm 0.2}$	4.4 ± 0.3	3.7 ± 0.3
r	4.4 ± 0.3	$\mathbf{2.9 \pm 0.2}$	3.4 ± 0.2	4.5 ± 0.3	4.6 ± 0.4	4.2 ± 0.3	3.3 ± 0.2	3.6 ± 0.2	3.6 ± 0.2
	$\mathbf{7.1 \pm 0.2}$	$\mathbf{5.9 \pm 0.2}$	$\mathbf{14.6 \pm 0.5}$	$\mathbf{4.1 \pm 0.1}$	$\mathbf{4.6 \pm 0.1}$	$\mathbf{4.2 \pm 0.1}$	$\mathbf{3.0 \pm 0.1}$	$\mathbf{3.4 \pm 0.1}$	$\mathbf{3.2 \pm 0.1}$

Table 6.4: Illustration of the demixing of n randomly chosen i.i.d. samples of length m, where the n sample distributions are drawn independently with replacement from Table 6.1. For the KC and KMI, a Gaussian kernel of size $\sigma = 1$ was used in all experiments. For the KCC and KGV, we used $\sigma = 1$ and $\kappa = 2 \times 10^{-2}$ for signals of length $m \leq 1000$, and $\sigma = 0.5$ and $\kappa = 2 \times 10^{-3}$ for the remaining signals. In both cases, $\epsilon = 1 \times 10^{-5}$ was used. The initial guess used by the kernel methods was given by Jade in all cases but $n = 16$, for which fast ICA was used (due to its more stable output).

n	m	Rep.	Fica	Jade	Imax	KCC	KC(g)	KC(l)	KGV	KMI(g)	KMI(l)
2	250	1000	10.5 ± 0.4	9.5 ± 0.4	44.4 ± 0.9	7.0 ± 0.3	7.8 ± 0.3	7.0 ± 0.3	$\mathbf{5.3 \pm 0.2}$	6.0 ± 0.2	5.7 ± 0.2
2	1000	1000	6.0 ± 0.3	5.1 ± 0.2	11.3 ± 0.6	3.3 ± 0.1	3.5 ± 0.1	2.9 ± 0.1	$\mathbf{2.3 \pm 0.1}$	2.6 ± 0.1	$\mathbf{2.3 \pm 0.1}$
4	1000	100	5.7 ± 0.4	5.6 ± 0.4	13.3 ± 1.1	4.5 ± 0.4	4.2 ± 0.3	4.6 ± 0.6	$\mathbf{3.1 \pm 0.6}$	4.0 ± 0.7	3.5 ± 0.7
4	4000	100	3.1 ± 0.2	2.3 ± 0.1	5.9 ± 0.7	2.4 ± 0.5	1.9 ± 0.1	1.6 ± 0.1	1.4 ± 0.1	1.4 ± 0.07	$\mathbf{1.2 \pm 0.07}$
8	2000	50	4.1 ± 0.2	3.6 ± 0.2	9.3 ± 0.9	4.8 ± 0.9	3.7 ± 0.9	5.2 ± 1.3	2.6 ± 0.3	2.1 ± 0.1	$\mathbf{1.9 \pm 0.1}$
8	4000	50	3.2 ± 0.2	2.7 ± 0.1	6.4 ± 0.9	2.1 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	1.4 ± 0.1	$\mathbf{1.3 \pm 0.05}$
16	5000	25	2.9 ± 0.1	3.1 ± 0.3	9.4 ± 1.1	3.7 ± 0.6	2.4 ± 0.1	2.6 ± 0.2	1.7 ± 0.1	$\mathbf{1.5 \pm 0.1}$	$\mathbf{1.5 \pm 0.1}$

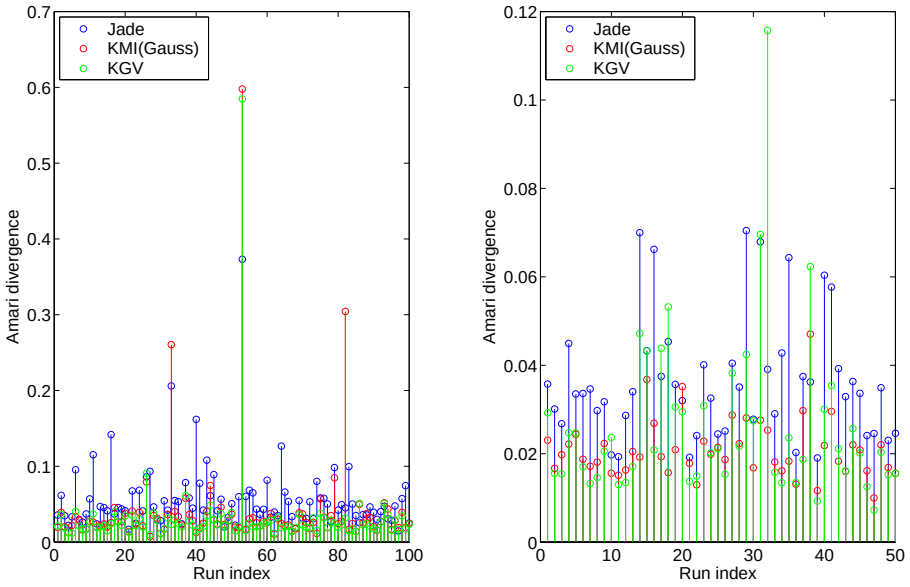

Figure 6.2.2: Amari divergence for demixing n randomly chosen signals, where $n = 4$ (left hand plot) and $n = 8$ (right hand plot), over 100 and 50 experiments respectively. Parameters are in both cases identical to those in Table 6.4.

6.2.2 Performance on difficult artificial problems

In our third experiment, we investigated the effects of noise added to the observations $\boldsymbol{t}$. We selected two generating distributions from Table 6.1, randomly and with replacement. After combining these signals with a randomly generated matrix with condition number between 1 and 2, we added points chosen with equal probability from $\{+5, -5\}$ to *both* signals, at random locations, in various quantities. Results are shown in the left hand plot in Figure 6.2.3. We used the Tanh and Gauss nonlinearities for the fast ICA algorithm, since these are resistant to outliers [46].

We observe that the KMI and KC are more resistant to outliers than the KGV and KCC contrasts, in that the rate of increase of the KC and KMI Amari divergences as a function of the number of corrupted points is less, and the Amari divergences at high noise levels are significantly lower. In addition, the kernel methods all perform very substantially better than the remaining methods, including those that are designed for robustness to outliers. As expected, Jade performed the most poorly, being based on highly noise-sensitive cumulants.

An additional experiment was also carried out on the same data, to test the sensitivity of the KCC and KGV to the choice of κ. Thus, we replaced $\kappa = 2 \times 10^{-2}$ (recommended in [7] for samples with $m \leq 1000$) with $\kappa = 2 \times 10^{-3}$ (used when $m > 1000$). We observe from the right hand plot in Figure 6.2.3 that this greatly reduces the performance of the KCC and KGV with respect to the KC and KMI, although these remain superior to fast ICA, Jade, and the extended Infomax methods.

Our fourth experiment concerns the effect of near-Gaussianity of the sources on the performance of the various algorithms. In this case, the two sources were both generated from members of the exponential power family,

$$\mathbf{f}_x(x) = a_1 \exp(a_2 |x|^\alpha)$$

where parameters a_1 and a_2 were chosen to ensure that the density was appropriately normalised, with zero mean and unit variance. The parameter α determines the nature of the distribution: $0 < \alpha < 2$ is super-Gaussian, while $\alpha > 2$ is sub-Gaussian ($\alpha = 1$ is the Laplace distribution). Results are given in the left hand plot in Figure 6.2.4. At zero Kurtosis, the KMI and KGV perform best, followed by the KCC and the KC. The kernel methods exhibit a decisive performance lead in

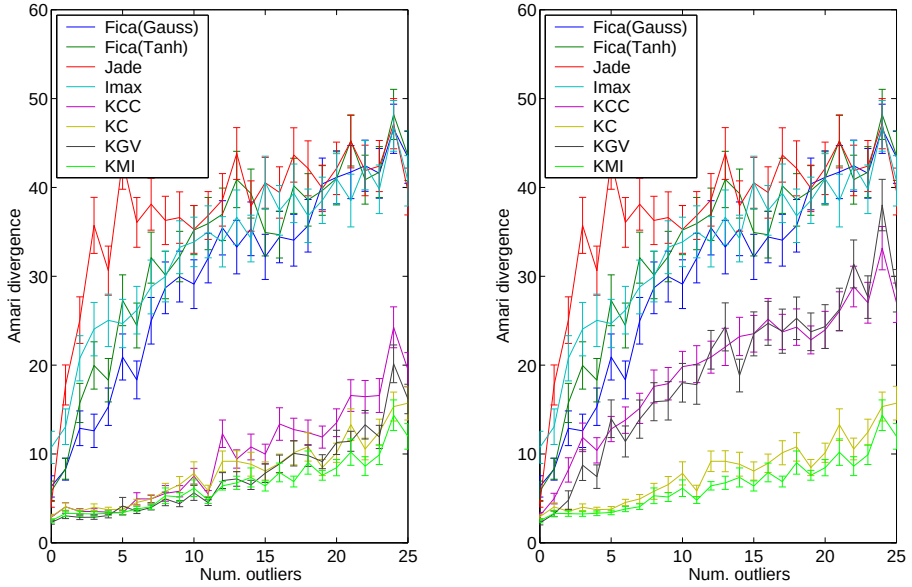

Figure 6.2.3: Effect of outliers on the performance of the ICA algorithms, for signals of length $m = 1000$, drawn independently with replacement from Table 6.1, and corrupted at random observations with outliers at ± 5 (where each sign has probability 0.5). Each point represents an average over 100 independent experiments. The number of corrupted observations in *both* signals is given on the horizontal axis. In both plots, we used $\sigma = 1$ for the kernel contrast functions, and $\epsilon = 2 \times 10^{-5}$. In the left hand plot, $\kappa = 2 \times 10^{-2}$ was used for the KCC and KGV, whereas $\kappa = 2 \times 10^{-3}$ was used in the right hand plot.

regions of positive and near-zero kurtosis, although traditional algorithms (notably Jade) are more competitive when the kurtosis is negative. Finally, the KGV and KCC are somewhat better in the vicinity of zero kurtosis, although the KMI and KC recover more rapidly as kurtosis increases.

Our fifth experiment addresses the effects of low kurtosis on the performance of our contrast functions, since many ICA contrasts rely (sometimes implicitly, through their choice of nonlinearity) on the kurtosis as an index of signal independence. Two signals were drawn from a single distribution, consisting of an asymmetric mixture of two Gaussians; the amplitudes of the Gaussians were adjusted to give both positive and negative kurtosis. Results are given in the right hand plot of Figure 6.2.4. All kernel based contrasts were unaffected by near-zero kurtosis, as opposed to both Jade and Fast ICA (which rely explicitly on the kurtosis) and the extended Infomax method (which does not). Surprisingly, the *Tanh* nonlinearity did not perform as well as the kurtosis based *Pow3* nonlinearity, when used in fast ICA.

6.2.3 Audio signal demixing

Our final experiment involved the demixing of brief extracts from various musical sources, which were combined using a randomly generated matrix (in the same manner as the artificial signals described in the previous section). A total of 17 different extracts were taken from the ICA benchmark set provided in [66]. These consist of 5 second segments sampled at 11 kHz with a precision of 8 bits, and represent a wide variety of musical genres. While adjacent samples of a musical signal are certainly not generated independently and identically, many ICA algorithms have nonetheless been applied successfully to this problem, and in this sense it constitutes a reasonable benchmark for our algorithm. Random permutation of time indices was used to reduce the statistical dependence of adjacent samples in the music, since this was found to improve performance.

A summary of our results is given in Table 6.5: the KMI performs best for two extracts, and the KGV does best with four extracts. In the $n = 2$ case, every possible combination of two different

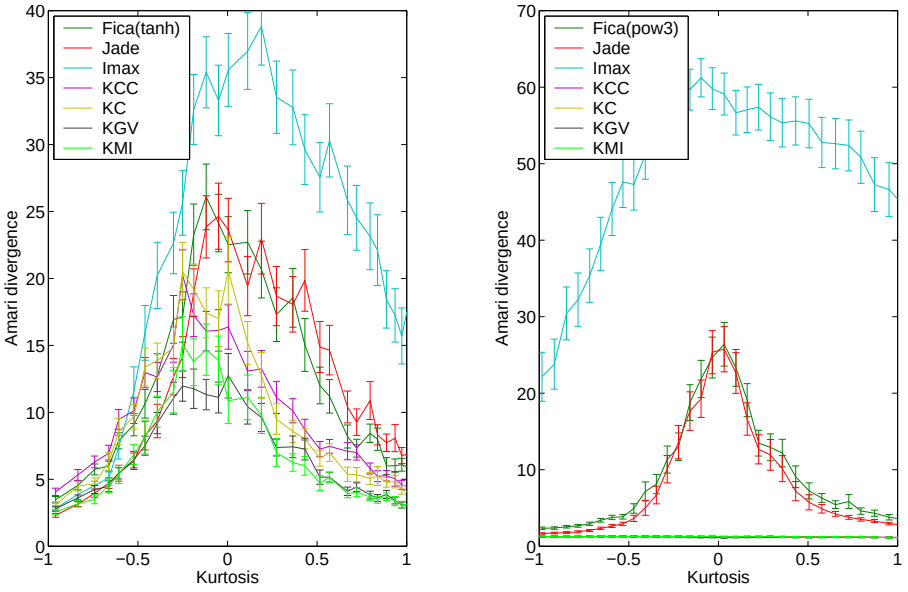

Figure 6.2.4: **Left hand plot:** Effect of near-Gaussianity on the performance of the algorithms, for two signals of length 1000 drawn from a range of generalised exponential distributions (see text). Each point represents an average over 100 independent experiments. We used a Laplace kernel with $\sigma = 3$ and precision $\epsilon = 0.01$ for the KC and KMI, and a Gaussian kernel with size $\sigma = 1$, precision $\epsilon = 2 \times 10^{-5}$, and $\kappa = 2 \times 10^{-2}$ for the KCC and KGV. The *tahn* nonlinearity was used for fast ICA due to its good performance.
Right hand plot: Effect of near-zero kurtosis on the performance of the algorithms, for two signals of length 1000 drawn from a range of mixtures of two Gaussians. Each point represents an average over 100 independent experiments. We used a Gaussian kernel with $\sigma = 1$ and precision $\epsilon = 2 \times 10^{-5}$ for all kernel contrast functions, and $\kappa = 2 \times 10^{-2}$ for the KCC and KGV.

Table 6.5: Illustration of the demixing of n music segments of length $m = 55272$, taken from the collection of 17 music samples in [66], and each representing an average over 120 experiments. Details of the KGV and KMI parameters may be found in Section 6.2.3.

n	Fica	Jade	Imax	KGV	KMI
2	1.10 ± 0.10	1.02 ± 0.06	1.33 ± 0.16	0.70 ± 0.05	$\mathbf{0.66 \pm 0.17}$
4	0.94 ± 0.03	0.89 ± 0.03	1.12 ± 0.06	$\mathbf{0.62 \pm 0.02}$	0.68 ± 0.03

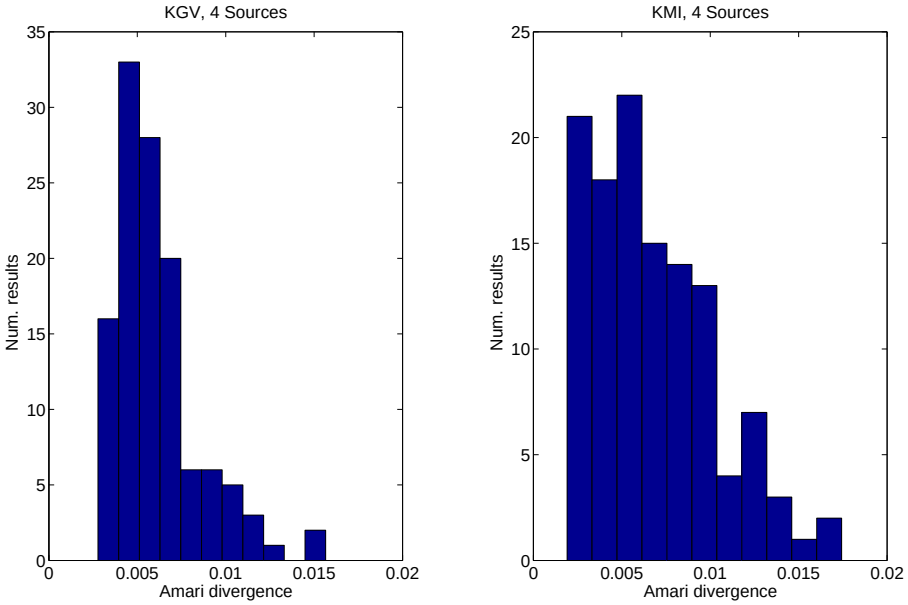

Figure 6.2.5: Histograms of the Amari divergences for the KGV and KMI, plotted for the 120 results obtained when unmixing four randomly selected music signals. Both the associated mean values are given in Table 6.5, as is a comparison with other ICA methods. The settings used for the KMI and KGV are described in Section 6.2.3.

extracts was investigated (for a total of 120 experiments), and the results averaged. We used $\kappa = 2 \times 10^{-3}$, $\sigma = 0.5$, $\epsilon = 1 \times 10^{-5}$, and a Gaussian kernel for the KGV; and $\sigma = 3$, $\epsilon = 1 \times 0.01$, and a Laplace kernel for the KMI. In both cases, a polishing step was applied to refine the result. For each experiment with $n = 4$, music segments were drawn randomly and without replacement from the 17 available extracts, and the results averaged over 120 repetitions. All kernel contrast parameters were the same as in the $n = 2$ case besides the Laplace kernel size, which was increased to $\sigma = 4$. In addition, no polishing step was applied to the KGV or KMI, since it caused a drop in performance for both contrasts [2]. Our use of the Laplace kernel in the KMI was motivated by music generally being super-Gaussian [10].

Although the results in Table 6.5 are quite similar for the KGV and KMI, it is instructive to compare the distribution of the outcomes obtained in each experiment. Histograms of these distributions are given for the $n = 4$ case in Figure 6.2.5. This plot reveals markedly different distributions of the Amari divergences, with the KGV results being more tightly grouped about their mean, while the KMI yields more results at smaller divergences, but a larger number of outliers.

[2] This is perhaps surprising, given that the polishing step caused a minor increase in performance in the $n = 2$ case. On the other hand, the larger dimension of the $n = 4$ problem makes the global minimum harder to find, and diversion to local minima more likely.

Chapter 7

Conclusions

7.1 Conclusions

Our experiments appear to demonstrate the effectiveness of the kernel methods, as compared with Jade, fast ICA, and the extended Infomax algorithms. Indeed, kernel methods proved themselves more resistant to noise in the observations, near-zero kurtosis, and near-Gaussianity. Moreover, the KMI and KGV yield the best performance in most of our experiments, including those conducted with audio signals. This is to be expected, since as we recall from Section 2, these methods all assume particular models for the source densities. Thus, although the contrast functions exhibit stable extrema at independence regardless of the accuracy of the underlying density models, the variance in the values of $\mathbf{W}$ at these extrema becomes larger as the models decrease in accuracy. By contrast, the kernel based methods use nonlinearities that adapt to the sources (as determined by the observations); indeed, the KMI and KGV use upper bounds on the Parzen window estimates of the source densities. This Parzen window interpretation facilitates the kernel choice when properties of the source densities are known: for instance, we have seen that the Laplace kernel-based KMI performs particularly well when separating samples generated by the double exponential and exponential distributions, and also performs better than the Gaussian kernel-based KMI when separating (heavy tailed) music extracts.

The computational cost of this performance is higher, however; this is particularly true when the Laplace kernel is used, since it occasionally requires the retention of a high rank in the Gram matrix approximation, due to the slow decay of the associated eigenvalues. The multimodal nature of the cost function being optimised by the kernel methods also necessitates a good initialisation method; although as pointed out earlier, methods exist to find good starting guesses within the kernel framework, and at a reasonable cost.

The choice between the KGV and KMI (or, alternatively, the KC and KCC) is more difficult. The methods proposed in [7] appear to do well when there is little data available, as in Table 6.2, and also the $n = 2$, $m = 250$ and $n = 4$, $m = 1000$ cases in Table 6.4. The KGV and KCC also yield better performance in the case where both sources have the same distribution (Table 6.3); although the gap is smaller in the $m = 1024$ experiment. The KMI and KC seem to have particular difficulty in separating two Student's t distribitions for a larger number of degrees of freedom (i.e., as the Student's t distribution approaches the Gaussian). Both small sample size and near-Gaussianity cause the variance of the best possible estimate of $\mathbf{W}$ to rise [20], making the ICA problem harder. Thus, it would appear that the use of $\mathrm{var}\left(f\left(\vec{\mathsf{x}}\right)\right) + \kappa \left\|f\right\|^2_{\mathcal{F}_\mathcal{X}}$ in the regularised KCC and KGV (see (3.4.7) in Section 3.4), rather than simply using $\left\|f\right\|^2_{\mathcal{F}_\mathcal{X}}$ (Definition 3.2.2 in Section 3.2), allows us to more closely approach the minimum variance estimate of $\mathbf{W}$ in these difficult cases, although the mechanism by which this is achieved remains unclear. This variance reduction effect is also seen in Figure 6.2.5.

On the other hand, the KCC and KGV appear more susceptible to noise in the observations, which is particularly apparent when κ becomes small (see Figure 6.2.4). The absence of κ in our kernel contrasts therefore greatly simplifies model selection, especially if the observations are known to be corrupted by noise. The KMI and KC also perform better in most cases described in Table 6.4, in which (generally) different sources are being separated. Thus, the good performance obtained by the KGV and KCC at low sample sizes seems to be offset by a drop in accuracy as the number of samples and/or sources increases. This apparent tradeoff cannot easily be explained using the analysis in Section 4.2, and requires further investigation.

A number of extensions to this work are readily apparent. For instance, the behaviour of the KMI has not been studied in detail for more than two univariate random variables, besides the discussion in Section 4.3 which guarantees it to be zero if and only if the sources are pairwise independent. In particular, it would be of interest to prove that (4.3.3) in Section 4.3 is an upper bound on the Gaussian mutual information, in the manner described in Section 4.2.2 for two random variables. This would incidentally require the link between the Gaussian mutual information and the discrete mutual information, described in Section 4.1 for the two variable case, to be extended to a greater number of random variables. These calculations might be made easier with the removal of the discretisation step in Section 4.3, since it is later negated by the subsequent limiting argument in Section 4.2. The optimisation procedure we use for ICA might also be made faster, for instance by implementing Newton's method or conjugate gradient descent on the Stiefel manifold, rather than simple gradient descent.

We have at present no guarantee that the empirical estimates of the KMI and KC converge towards their population expressions, which requires the application of concentration inequalities. It is possible to use Rademacher averages [63] to bound the deviation of the KC from its expected value, by applying the following theorem.

Theorem 7.1.1 (Rademacher bound on the deviation of a function from its expectation). *Let $\mathcal{F}$ be a class of uniformly bounded functions on $\mathcal{X}$, and let $\mathbf{P}_{\mathsf{x}}$ be a probability measure on $\mathcal{X}$. Then there exists a universal constant $C \geq 1$ such that for all $\epsilon > 0$, with probability at least $1 - \exp\left(-\frac{m\epsilon^2}{4C^2}\right)$ over the random draw of $\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}}^m$,*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{m} f(x_i) - \mathbf{E}_{\mathsf{x}}(f(\mathsf{x})) \right| \leq \frac{4}{\sqrt{m}} \mathcal{R}_m(\mathcal{F}, \mathbf{P}_{\mathsf{x}}) + 3\frac{\epsilon}{4},$$

where

$$\mathcal{R}_m(\mathcal{F}, \mathbf{P}_{\mathsf{x}}) := \mathbf{E}_{\mathsf{x}_1, \ldots, \mathsf{x}_m, \boldsymbol{\sigma}} \left(\frac{1}{\sqrt{m}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{m} \sigma_i f(\mathsf{x}_i) \right| \right)$$

and $\mathbf{P}_\sigma(1) = \mathbf{P}_\sigma(-1) = 1/2$.

The application of this result in predicting KMI performance is less clear, however, since the KMI is a *product* of multiple KC-type quantities. More generally, it is necessary to further investigate methods for model selection (i.e., for choosing the kernel size and type) in the KC and KMI. It is not presently known whether performance is most effectively tuned by simple cross-validation, using bounds derived from concentration inequalities, or via the properties of Parzen window estimates discussed in [80].

The good performance of our algorithm in the context of ICA entails a higher computational cost than the other methods investigated. It would thus be of interest to compare with other recent semi- and non-parametric entropy approximations proposed for ICA [41, 64, 70], which also adapt their source density estimates according to the observations. Many real life problems do not fit neatly into the linear ICA framework, however; we now outline ways in which our kernel contrasts might be used to improve performance in these more difficult signal separation problems.

7.2 ICA for stationary random processes

It is rare in practice to encounter signals that do not depend on their previous outputs. Rather, most real signals are drawn from random processes, for which there are statistical dependencies between the observations at different times. These random processes may be stationary, meaning that their statistical properties (for instance the mean and correlation) do not change over time; or they may be nonstationary. In both cases, however, the time dependence greatly assists in separating signals into independent components, the idea being that the independence of different random processes should hold not only between samples drawn at the same time, but also between samples drawn at *different* times.

A simple and computationally efficient criterion is described in [11] to separate linearly mixed, independent, and stationary random processes, using only second order moments. This is achieved by jointly diagonalising the covariance matrices

$$\mathbf{R}(\ell) := \mathbf{E}_{\tilde{\mathbf{t}}(l)\tilde{\mathbf{t}}(l+\ell)}(\tilde{\mathbf{t}}(l)\tilde{\mathbf{t}}^{\top}(l+\ell)),$$

parameterised by various delay values ℓ, where the stationarity of the process causes the expectation to depend only on the time delay [1]. When a large number of delays ℓ are chosen, then the signals become inseparable only when the power spectra of the sources are proportional; it is noted in [11], however, that the expected cross-signal interference in our estimate of the sources becomes high as this indeterminate situation is approached, and the spectral overlap increases. This problem might be avoidable by refining the decorrelation-based solution, using the proposed kernel contrasts (which depend on higher order moments, although this is obviously ineffective if the sources are merely Gaussian random processes). This would require us to extend the KMI to random *vectors*, rather than random variables (the dimension of the vectors corresponding to the number of different lags ℓ considered previously); in other words, we would need to show that the KMI remains an upper bound on $D_{\mathrm{KL}}(\mathbf{f}_{\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_n} || \mathbf{f}_{\vec{\mathbf{x}}_1} \ldots \mathbf{f}_{\vec{\mathbf{x}}_n})$ near independence when the dimension of $\vec{\mathbf{x}}_i$ is greater than one (we remind the reader that the KC and KCC are already known to be valid contrast functions in this circumstance, although they might not perform as well).

Another alternative would be to derive more general features from the signals, and to determine the independence of these features. This was done in [12] using Cohen's class time-frequency kernels, which permits the separation of sources with identical spectra. Cohen's class T-F kernels were also applied in [38, 30] to describe the identifying properties of audio samples: they were used in this case to train a support vector machine to separate polynomial phase signals.

7.3 Nonlinear mixtures

The difficulty of modeling nonlinear mixtures is considerably greater, since the complexity of the problem is far higher, and the number of indeterminacies much larger. One way of dealing with these problems is to assume a specific functional form for the nonlinearity, as in [89]. An alternative is to use a simplified nonlinearity, such that the ith component of the observation vector $\mathbf{t}$ is

$$t_i = f(\mathbf{b}_i \mathbf{s}),$$

where f_i is the ith (unknown) nonlinearity, and $\mathbf{b}_i$ is the ith row of the mixing matrix $\mathbf{B}$; this situation corresponds for instance to the observations being distorted by the sensors. This situation

[1] An equivalent approach based on the Whittle approximation is proposed in [73]. Here, the likelihood of the ith independent stationary random processes $\mathbf{s}_i(l)$, where $i \in \{1, \ldots, n\}$, is approximated by modeling the DFT, $\tilde{\mathbf{s}}_i(k)$ for $k = (1, \ldots, m)$, as series of m independent Gaussians. These Gaussians are assigned zero mean, and variances equal to the power at the kth frequency. Independence is then attained when the inter-signal correlation is zero between the frequency components $\tilde{\mathbf{x}}_i(k)$ and $\tilde{\mathbf{x}}_j(k)$, $i \neq j$, of the unmixed signals at each k. In addition, methods designed specifically for separating signals generated using Gaussian processes, or via generalised ARMA models (that is, with non-Gaussian noise), are proposed in [71].

is analysed in [82, 1], in which it is shown that a fixed source density model, of the kind described in Section 2, performs exceptionally badly: rather, it is necessary to estimate the densities of the sources when computing the contrast, which is accomplished using a kernel density estimate (the details of this step differ in the two studies cited). A comparison of these methods with the KMI would therefore be of interest. Various efforts have also been made to solve the more general case

$$\mathbf{t} = f(\mathbf{s}).$$

This problem requires constraints on f to be determinate (and even then, it is generally the case that each source s_i can only be recovered up to a nonlinear distortion; this is the analogue of the scaling indeterminacy (Theorem 2.1.2) in the linear mixing case). An example of indeterminacy when $f(\cdot)$ is unconstrained is given in [51]: we apply the Darmois decomposition to the random variable $\mathbf{t}$ to obtain

$$\begin{aligned} \mathsf{x}_1 &= \mathsf{F}_{\mathsf{t}_1}(\mathsf{t}_1), \\ \mathsf{x}_2 &= \mathsf{F}_{\mathsf{t}_2}(\mathsf{t}_2|\mathsf{t}_1), \\ \mathsf{x}_3 &= \mathsf{F}_{\mathsf{t}_3}(\mathsf{t}_3|\mathsf{t}_1, \mathsf{t}_2), \end{aligned}$$

and so on. The components of the output $\mathbf{x}$ are then independent, and each has a uniform distribution in $[0, 1]$. A constraint proposed in [51], which applies when there are two observations ($n = 2$), is to require $f(\cdot)$ to be a conformal mapping[2] of the complex variable $t_1 + \imath t_2$, which removes the ambiguity in most cases. A different approach is proposed in [45], where it is shown that enforcing temporal decorrelation over a single time step is sufficient to test whether the recovered independent processes are simply the result of a Darmois decomposition. While this does not rule out other transforms that return independent signals unrelated to the sources, it suggests that time dependencies have a crucial rôle to play in general nonlinear mixing. In the scheme suggested in [39], demixing is achieved by mapping the observations to a reproducing kernel Hilbert space, finding a low dimensional basis in the feature space which approximately spans the subspace formed by the observations[3], and enforcing the second order temporal decorrelation of projections onto this basis (the last step is much like that in [11]). It remains unclear whether guarantees exist to ensure the recovered independent sources are identical (aside from the nonlinear distortion indeterminacy described above) to the original sources; indeed, multiple signals are recovered, and the correct ones are recognised based on their expected statistical properties. It should also be noted that the algorithm does not work when the time dependency between the samples is ignored, thus lending support to the hypothesis in [45]. The applicability of the KMI is less clear than in the case of post-nonlinear mixtures, although this might follow from a better understanding of [39] and its relation to our work.

7.4 Models for dependent random variables

The KGV is used in [8] to model the source distribution $\mathbf{f}_\mathbf{s}$ as a tree structured graphical model, which incorporates dependencies between components (unlike the independence assumption made in ICA); this can also be thought of as a method for estimating the high dimensional multivariate density $\mathbf{f}_\mathbf{t}$ using only densities of lower dimension and linear mixing. If $\mathcal{T}$ is an undirected spanning tree on the vertices $\{1, \ldots, m\}$, then the source model $\widehat{\mathbf{f}}_\mathbf{s}$ is assumed to factorise in $\mathcal{T}$. Writing $\mathbf{x} = \mathbf{W}\mathbf{t}$ as the unmixed signal, the minimum possible loss due to this encoding (with respect to both $\mathbf{W}$ and to the choice of $\mathcal{T}$) is expressed as

$$\min_{\mathbf{f}_\mathsf{x}} D_{\mathrm{KL}}\left(\mathbf{f}_\mathbf{x} || \widehat{\mathbf{f}}_\mathbf{s}\right) = I(\mathbf{x}) - \sum_{(i,j)\in\mathcal{T}} I(\mathsf{x}_i, \mathsf{x}_j). \tag{7.4.1}$$

Two methods for attaining this minimum are compared in [8]. In the first approach, the KGV is used to replace all the terms in (7.4.1), and the resulting expression is minimised with respect to $\mathcal{T}$

[2] That is, the derivative must always exist and be non-zero.

[3] This basis generally has a very low dimension, irrespective of the number of observations.

and $\mathbf{W}$. In the second case, the decomposition (2.2.9) in Section 2.2.3 is applied to $I(\mathbf{x})$, and both the pairwise mutual information terms and the entropies are estimated using Parzen windows; this expression is then minimised. Although the second method generally performs better, the KGV is also very effective. This gives an indication of the tightness of the upper bound on the mutual information provided by the KGV, since minimising (7.4.1) in the KGV setting involves maximising *upper bounds* on the pairwise mutual information terms. Given that the KMI is in theory a tighter upper bound than the KGV, it would be interesting to compare its performance with the KGV in this setting.

Appendix A

Proofs and Definitions

A.1 Standard linear algebra results

These results are mainly taken from [43, 40], with some results also from [68, 81, 16]. Proofs are rarely provided, although certain proofs are included when they yield insight into the main body of the discussion.

A.1.1 Miscellaneous definitions

Definition A.1.1 (Trace). The trace of an $n \times n$ matrix $\mathbf{A}$ is the sum of its diagonal elements $a_{i,i}$;

$$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} a_{i,i}$$

Definition A.1.2 (Rank). The rank of a matrix $\mathbf{A}$ is the maximum number of independent columns *and* rows in $\mathbf{A}$.

Definition A.1.3 (Subspaces of a matrix). Consider an $m \times n$ matrix $\mathbf{A}$. The four fundamental subspaces of $\mathbf{A}$ are:

- *Column space:* the space spanned by the columns of $\mathbf{A}$
- *Row space:* the space spanned by the rows of $\mathbf{A}$
- *Left nullspace:* the orthogonal complement to the *column space* of $\mathbf{A}$
- *Nullspace:* the orthogonal complement to the *row space* of $\mathbf{A}$

Theorem A.1.4 (Rank of a product of matrices). *The* column space *of* $\mathbf{AB}$ *is spanned by the column space of* $\mathbf{A}$*, and the row space of* $\mathbf{AB}$ *is spanned by the row space of* $\mathbf{B}$*. Consequently,*

$$\operatorname{rank}(\mathbf{AB}) = \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\}.$$

A.1.2 Matrix inner products, projections

Definition A.1.5 (Matrix inner product). Given matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ in the space $\mathcal{V}$ of $m \times n$ matrices (or some subspace thereof), the inner product is any function $f(\mathbf{A}, \mathbf{B})$ for which the axioms of an inner product space are satisfied, namely

- $\langle \mathbf{A}, \mathbf{B} \rangle = \langle \mathbf{B}, \mathbf{A} \rangle$
- $\langle \alpha \mathbf{A} + \beta \mathbf{B}, \mathbf{C} \rangle = \alpha \langle \mathbf{A}, \mathbf{C} \rangle + \beta \langle \mathbf{B}, \mathbf{C} \rangle$
- $\langle \mathbf{A}, \mathbf{A} \rangle \geq 0$, with equality if and only if $\mathbf{A} = \mathbf{0}$.

One such function is $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)$.

Definition A.1.6 (Matrix norm). Given an $m \times n$ matrix $\mathbf{A}$, it follows from the definition of the matrix inner product that the matrix norm can be written

$$\|\mathbf{A}\| = \langle \mathbf{A}, \mathbf{A} \rangle^{1/2} = \left(\operatorname{tr}\left(\mathbf{A}^\top \mathbf{A}\right)\right)^{1/2}.$$

Definition A.1.7 (Angle between two matrices). It is possible to define an angle between two $m \times n$ matrices $\mathbf{A}, \mathbf{B}$, as

$$\cos(\theta) = \frac{\langle \mathbf{A}, \mathbf{B} \rangle}{\|\mathbf{A}\| \, \|\mathbf{B}\|}$$

We emphasise that the Cauchy-Schwarz inequality holds for the inner product and norm defined above.

Theorem A.1.8 (Projection of a vector onto a column space). *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix, for which* $m \geq n$*, and let* $\mathbf{b}$ *be an* m *dimensional vector in* $\mathbb{R}^m$ *(we do* not *require, however, that* $\operatorname{rank}(\mathbf{A}) = n$*). Writing as* $\mathbf{p} = \mathbf{A}\boldsymbol{\gamma}$ *the least squares projection of* $\mathbf{b}$ *onto the* n *columns of* $\mathbf{A}$*, where* $\boldsymbol{\gamma}$ *is an* n *dimensional vector, then*

$$\boldsymbol{\gamma} = \left(\mathbf{A}^\top \mathbf{A}\right)^- \mathbf{A}^\top \mathbf{b}$$

and

$$\mathbf{p} = \mathbf{A}\left(\mathbf{A}^\top \mathbf{A}\right)^- \mathbf{A}^\top \mathbf{b}, \tag{A.1.1}$$

where we used the pseudoinverse in Definition A.1.17.

Proof. When $\mathbf{p}$ is a least squares estimate of $\mathbf{b}$ on the subspace spanned by the columns of $\mathbf{A}$, it follows that $\mathbf{b} - \mathbf{p}$ is orthogonal to the columns of $\mathbf{A}$. In other words,

$$\mathbf{0} = \mathbf{A}^\top (\mathbf{b} - \mathbf{p}) = \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\boldsymbol{\gamma}),$$

or

$$\mathbf{A}^\top \mathbf{A}\boldsymbol{\gamma} = \mathbf{A}^\top \mathbf{b}.$$

It follows that $\boldsymbol{\gamma} = \left(\mathbf{A}^\top \mathbf{A}\right)^- \mathbf{A}^\top \mathbf{b}$. Then, using the definition $\mathbf{p} = \mathbf{A}\boldsymbol{\gamma}$, we obtain (A.1.1). □

A.1.3 Properties of the determinant

Theorem A.1.9 (Determinant of a product of matrices). *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $n \times n$ *matrices, then*

$$\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B})$$

Theorem A.1.10 (Determinant of a matrix of which a submatrix is 0). *If* $\mathbf{M}$ *is defined such that*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \text{or} \quad \mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix},$$

then

$$|\mathbf{M}| = |\mathbf{A}| \, |\mathbf{D}|.$$

Theorem A.1.11 (Determinant of a matrix containing unit submatrices). *If* $\mathbf{B}$ *and* $\mathbf{C}^\top$ *are* $m \times n$ *matrices, then*

$$\left|\begin{bmatrix} \mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{I} \end{bmatrix}\right| = |\mathbf{I} - \mathbf{BC}| = |\mathbf{I} - \mathbf{CB}|$$

Theorem A.1.12 (Determinant of a scaled matrix). *If* $\mathbf{A}$ *is an* $n \times n$ *matrix, then*

$$|k\mathbf{A}| = k^n |\mathbf{A}|$$

Definition A.1.13 (Singular matrix). An $n \times n$ matrix $\mathbf{A}$ is singular if and only if its determinant is zero.

Theorem A.1.14 (Determinant of matrix with less than full rank). *If* $\mathbf{A}$ *is an* $n \times n$ *matrix, and* $\text{rank}(\mathbf{A}) < n$*, then* $|\mathbf{A}| = 0$.

Theorem A.1.15 (Determinant of inverse). *If* $\mathbf{A}$ *is an invertible matrix, then*

$$\frac{1}{|\mathbf{A}|} = \left|\mathbf{A}^{-1}\right|.$$

Theorem A.1.16 (Determinant of a partitioned matrix). *If*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix},$$

and assuming the existence of $\mathbf{A}^{-1}$ *and* $\mathbf{D}^{-1}$*, then*

$$|\mathbf{M}| = |\mathbf{A}| \left|\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}\right| = |\mathbf{D}| \left|\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}\right|.$$

A.1.4 Properties of the matrix inverse

Definition A.1.17 (Matrix pseudoinverse). Let $\mathbf{A}$ be any $m \times n$ matrix, and let $\mathbf{G}$ be an $n \times m$ matrix. Then $\mathbf{G}$ is called the *pseudoinverse* or *generalised inverse* of $\mathbf{A}$ when

$$\mathbf{AGA} = \mathbf{A}.$$

When $\mathbf{A}$ is square with full rank, then $\mathbf{G} = \mathbf{A}^{-1}$ is the only such pseudoinverse; otherwise, an infinite number of such pseudoinverses exist (moreover, every matrix has at least one pseudoinverse). The pseudoinverse of $\mathbf{A}$ is commonly written $\mathbf{A}^-$.

Theorem A.1.18 (Pseudoinverse as a solution to a linear system). *Let* $\mathbf{A}$ *be any* $m \times n$ *matrix, and let* $\mathbf{B}$ *be an* $m \times p$ *matrix for which the linear system* $\mathbf{AX} = \mathbf{B}$ *is consistent. Then* $\mathbf{GB}$ *is a solution to* $\mathbf{AX} = \mathbf{B}$ *if and only if* $\mathbf{G} = \mathbf{A}^-$.

Theorem A.1.19 (Inverse of a matrix product). *If* $\mathbf{A}$, $\mathbf{B}$ *are invertible matrices, then*

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Theorem A.1.20 (Inverse of a partitioned matrix). *If*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix},$$

then assuming the existence of $\mathbf{A}^{-1}$*,*

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BECA}^{-1} & -\mathbf{A}^{-1}\mathbf{BE} \\ -\mathbf{ECA}^{-1} & \mathbf{E} \end{bmatrix},$$

where $\mathbf{E} = (\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}$*. Assuming the existence of* $\mathbf{D}^{-1}$*,*

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{F} & -\mathbf{FBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CF} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CFBD}^{-1} \end{bmatrix},$$

where $\mathbf{F} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}$.

Theorem A.1.21 (Sherman-Morrison-Woodbury formulae). *If* $\mathbf{I} + \mathbf{V}^\top \mathbf{A}^{-1}\mathbf{U}$ *is nonsingular, then*

$$\left(\mathbf{A} + \mathbf{U}\mathbf{V}^\top\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I} + \mathbf{V}^\top\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^\top\mathbf{A}^{-1}.$$

If $\mathbf{v}^\top\mathbf{A}\mathbf{u} \neq -1$*, then*

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^\top\right)^{-1} = \mathbf{A}^{-1} - \frac{1}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}.$$

A.1.5 Eigenvalues and eigenvectors

Definition A.1.22 (Eigenvalues and eigenvectors). If $\mathbf{A}$ is a *square* matrix of dimension n, and

$$\mathbf{A}\mathbf{z} = \lambda\mathbf{z},$$

then λ is the eigenvalue of $\mathbf{A}$ corresponding to the eigenvector $\mathbf{u}$, where $\mathbf{u} \neq \mathbf{0}$. As a consequence of the above definition, the eigenvalues of a diagonal matrix are the elements of the diagonal.

Theorem A.1.23 (Characteristic polynomial). *The roots* λ_i *of the eigenvalue problem*

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

are the same as the roots of the characteristic polynomial

$$\left|\mathbf{A} - \lambda\mathbf{I}\right|.$$

Theorem A.1.24 (Similar matrices). *Given the* $m \times m$ *matrices* $\mathbf{A}, \mathbf{M}$*, where* $\mathbf{M}$ *is invertible, then the matrices* $\mathbf{A}$ *and* $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$ *are said to be* similar*. It follows that* $\mathbf{A}$ *and* $\mathbf{B}$ *have the same eigenvalue spectrum* $\boldsymbol{\lambda}$*.*

Proof. This proof is taken from Strang [81]. We simply apply the definition of the eigenvalue decomposition;

$$\begin{aligned}
\mathbf{A}\mathbf{u} &= \lambda\mathbf{u} \\
\mathbf{M}\mathbf{B}\mathbf{M}^{-1}\mathbf{u} &= \lambda\mathbf{u} \\
\mathbf{B}\left(\mathbf{M}^{-1}\mathbf{u}\right) &= \lambda\left(\mathbf{M}^{-1}\mathbf{u}\right).
\end{aligned}$$

□

Theorem A.1.25 (The singular value decomposition). *Any* $m \times n$ *matrix* $\mathbf{A}$ *of rank* s *may be written*

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{R}^\top,$$

where

- $\mathbf{Q}$ *is an* $m \times s$ *matrix, of which the columns are the eigenvectors of* $\mathbf{A}\mathbf{A}^\top$ *with non-zero eigenvalues, and* $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$*,*
- $\mathbf{R}$ *is an* $n \times s$ *matrix, of which the columns are the eigenvectors of* $\mathbf{A}^\top\mathbf{A}$ *with non-zero eigenvalues, and* $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$*,*
- Λ *is an* $s \times s$ *diagonal matrix, containing the singular values,*
- *both* $\mathbf{A}\mathbf{A}^\top$ *and* $\mathbf{A}^\top\mathbf{A}$ *have the same set of non-zero eigenvalues, obtained by squaring the singular values.*

A.1.6 Properties of symmetric matrices

Theorem A.1.26 (Eigenvalues of symmetric matrix). *The eigenvalues of a symmetric matrix are* real.

Theorem A.1.27 (Determinant of symmetric matrix). *Any symmetric $n \times n$ matrix $\mathbf{A}$ with eigenvalues $\lambda_1, \cdots, \lambda_n$ has the property*

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$$

Theorem A.1.28 (Trace of symmetric matrix). *Any symmetric $n \times n$ matrix $\mathbf{A}$ with eigenvalues $\lambda_1, \cdots, \lambda_n$ has the property*

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

Theorem A.1.29 (Spectral decomposition theorem). *Any symmetric $n \times n$ matrix $\mathbf{A}$ with eigenvalues $\lambda_1, \cdots, \lambda_n$ can be written*

$$\mathbf{A} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top,$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, $\mathbf{E}\mathbf{E}^\top = \mathbf{I}$, and $\boldsymbol{\lambda}$ is the vector of eigenvalues of $\mathbf{A}$.

A.1.7 Properties of positive (semi)definite matrices

Definition A.1.30 (Positive definite/semidefinite matrices). A *symmetric $n \times n$* matrix $\mathbf{A}$, with individual entries $a_{i,j}$, is positive definite if and only if

$$\boldsymbol{\alpha}^\top \mathbf{A} \boldsymbol{\alpha} \geq 0 \tag{A.1.2}$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^n$, with equality only when $\boldsymbol{\alpha} = \mathbf{0}$. The matrix $\mathbf{A}$ is positive semidefinite if and only if (A.1.2) is non-negative for all $\boldsymbol{\alpha} \in \mathbb{R}^n$. If $\mathbf{A}$ is complex and conjugate symmetric, so that $a_{i,j} = a_{j,i}^*$, then $\mathbf{A}$ is positive definite if and only if

$$\left(\boldsymbol{\alpha}^\top\right)^* \mathbf{A} \boldsymbol{\alpha} \geq 0,$$

for all $\boldsymbol{\alpha} \in \mathbb{C}^n$, with equality only when $\boldsymbol{\alpha} = \mathbf{0}$.

Theorem A.1.31 (Eigenvalues of a positive definite matrix). *If a matrix is positive definite, then its eigenvalues are* positive. *If a matrix is positive semidefinite, then its eigenvalues are* non-negative.

Theorem A.1.32 (Cholesky decomposition). *If $\mathbf{A}$ is a positive definite or positive semidefinite matrix, it may be decomposed as $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$, where $\mathbf{B}$ is upper triangular, using the standard LU decomposition procedure. This is known as the Cholesky decomposition.*

A.1.8 Derivatives

Theorem A.1.33 (Derivative of a linear function). *Given an $m \times n$ matrix $\mathbf{A}$ and an $m \times 1$ vector $\mathbf{b}$, then the derivative of $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ with respect to the $n \times 1$ dimensional vector $\mathbf{x}$ is*

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}^\top.$$

Theorem A.1.34 (Derivative of a quadratic form). *Let $\mathbf{A}$ be a symmetric $m \times m$ matrix, and $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ a scalar valued function. Then the derivative of this function with respect to the $m \times 1$ dimensional vector $\mathbf{x}$ is*

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

A.2 Normalised covariance: equivalent eigenvalue problem

A.2.1 Solution unconstrained

In this section, which constitutes a proof of Theorem 3.2.1, we find the vectors $\boldsymbol{\alpha}_i \in \mathcal{F}_\mathcal{X} : \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_i \leq 1$ and $\boldsymbol{\beta}_i \in \mathcal{F}_\mathcal{Y} : \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_i \leq 1$ onto which random vectors $\mathbf{x}$ and $\mathbf{y}$ respectively project, such that the covariance γ_i between these projections is a stationary point with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$. These results are well established, and may be found for instance in [87]. The Lagrangian for this problem is written

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \xi) &= \mathsf{E}_{\mathbf{x},\mathbf{y}}(\boldsymbol{\alpha}^\top \mathbf{x}\mathbf{y}^\top \boldsymbol{\beta}) - \mathsf{E}_\mathbf{x}(\boldsymbol{\alpha}^\top \mathbf{x})\mathsf{E}_\mathbf{y}(\mathbf{y}^\top \boldsymbol{\beta}) - \lambda\left(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - 1\right) - \xi\left(\boldsymbol{\beta}^\top \boldsymbol{\beta} - 1\right) \\ &= \boldsymbol{\alpha}^\top \mathbf{C}_{xy} \boldsymbol{\beta} - \lambda\left(\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - 1\right) - \xi\left(\boldsymbol{\beta}^\top \boldsymbol{\beta} - 1\right). \end{aligned}$$

We wish for the stationary points of the covariance to occur at the zero derivatives of the above expression: according to [61, Section 7.3], this is true as long as Slater's condition holds on the feasible region (that is, the feasible region is required to have an interior point, which is true of the constraints $\boldsymbol{\alpha}^\top \boldsymbol{\alpha} \leq 1$ and $\boldsymbol{\beta}^\top \boldsymbol{\beta} \leq 1$). We therefore set the derivatives with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \xi$ to zero, to compute the saddle points. This yields

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\alpha}} &= \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda\boldsymbol{\alpha} = 0, \\ \frac{\partial L}{\partial \boldsymbol{\beta}} &= \mathbf{C}_{xy}^\top \boldsymbol{\alpha} - 2\xi\boldsymbol{\beta} = 0. \end{aligned}$$

We now demonstrate that $\lambda = \xi$, by premultiplying the first expression by $\boldsymbol{\alpha}^\top$ and the second expression by $\boldsymbol{\beta}^\top$;

$$\begin{aligned} \boldsymbol{\alpha}^\top \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha} &= \boldsymbol{\alpha}^\top \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda = 0, \\ \boldsymbol{\beta}^\top \mathbf{C}_{xy}^\top \boldsymbol{\alpha} - 2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta} &= \boldsymbol{\beta}^\top \mathbf{C}_{xy}^\top \boldsymbol{\alpha} - 2\xi = 0. \end{aligned}$$

We therefore write $\gamma = 2\lambda = 2\xi$, and the solution becomes a single eigenvalue equation,

$$\begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \gamma \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}, \tag{A.2.1}$$

which completes the proof.

A.2.2 An alternative form of the unconstrained solution

In this section, we derive the form of the eigenvalue solution presented in Lemma 3.3.1. The initial form of the kernel covariance in (3.2.5) may be written

$$\max_{\mathbf{c},\mathbf{d}} \left(\mathbf{c}^\top \widetilde{\mathbf{K}}_{mm}^{(x)} \widetilde{\mathbf{K}}_{mm}^{(y)} \mathbf{d}\right) : \mathbf{c}^\top \widetilde{\mathbf{K}}_{mm}^{(x)} \mathbf{c} \leq 1, \mathbf{d}^\top \widetilde{\mathbf{K}}_{mm}^{(y)} \mathbf{d} \leq 1.$$

Introducing two new variables

$$\breve{\boldsymbol{\alpha}} = \left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2} \mathbf{c}, \qquad \breve{\boldsymbol{\beta}} = \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2} \mathbf{d},$$

this becomes

$$\max_{\breve{\boldsymbol{\alpha}},\breve{\boldsymbol{\beta}}} \left(\breve{\boldsymbol{\alpha}} \left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2} \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2} \breve{\boldsymbol{\beta}}\right) : \breve{\boldsymbol{\alpha}}^\top \breve{\boldsymbol{\alpha}} \leq 1, \breve{\boldsymbol{\beta}}^\top \breve{\boldsymbol{\beta}} \leq 1.$$

The Lagrangian is written

$$L\left(\breve{\boldsymbol{\alpha}}, \breve{\boldsymbol{\beta}}, \lambda, \gamma\right) = \breve{\boldsymbol{\alpha}}^\top \left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2} \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2} \breve{\boldsymbol{\beta}} - \lambda\left(\breve{\boldsymbol{\alpha}}^\top \breve{\boldsymbol{\alpha}} - 1\right) - \xi\left(\breve{\boldsymbol{\beta}}^\top \breve{\boldsymbol{\beta}} - 1\right).$$

Using Slater's condition [61, Section 7.3], we find that the zero derivatives of the above are the stationary points of $\breve{\boldsymbol{\alpha}}\left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\breve{\boldsymbol{\beta}}$ with respect to the constrained $\breve{\boldsymbol{\alpha}}, \breve{\boldsymbol{\beta}}$. As in the previous section, we get the system of equations

$$\begin{cases} \left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\breve{\boldsymbol{\beta}} &= \gamma\breve{\boldsymbol{\alpha}}, \\ \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2}\breve{\boldsymbol{\alpha}} &= \gamma\breve{\boldsymbol{\beta}}. \end{cases}$$

Substituting the first of these into the second, we obtain

$$\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^{1/2}\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\breve{\boldsymbol{\beta}} = \gamma^2\breve{\boldsymbol{\beta}}. \tag{A.2.2}$$

In other words,

$$\gamma^2 = \left\|\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\widetilde{\mathbf{K}}_{mm}^{(x)}\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\right\| = \left\|\widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)}\right\|,$$

where the final equality is found by premultiplying (A.2.2) with $\left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}$ and making the change of variables $\breve{\boldsymbol{\beta}} = \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^{1/2}\breve{\boldsymbol{\beta}}$ (the norm here represents the largest eigenvalue of the matrix).

A.2.3 Solution restricted to a specific basis

In this section, we prove Theorem 4.2.3 by deriving the stationary directions of the normalised covariance, in the case where these are restricted to be linear combinations of a specific set of basis vectors. In other words,

$$\widehat{\boldsymbol{\alpha}} = \mathbf{Q}\hat{\mathbf{c}}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{R}\hat{\mathbf{d}}, \tag{A.2.3}$$

where the columns of $\mathbf{Q}, \mathbf{R}$ are vectors in the respective feature spaces $\mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}$ (hats are used to distinguish these solutions from the unconstrained case in Section 3.2). To our knowledge, this result is new: constrained solutions in the context of the *canonical correlation* are discussed in [55], in which the projection directions are restricted to certain linear combinations of the observed samples[1] $\boldsymbol{x}, \boldsymbol{y}$; and in [29], although the constraints in the latter case are inequalities on the coefficients, and do not represent a restriction to a specific basis. The constrained empirical estimate of (3.2.1) in Theorem 3.2.1 becomes[2]

$$\operatorname{cov}\left(\widehat{\boldsymbol{\alpha}}^\top\mathbf{x}, \widehat{\boldsymbol{\beta}}^\top\mathbf{y}\right) = \hat{\mathbf{c}}^\top\mathbf{Q}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top\mathbf{R}\hat{\mathbf{d}}, : \widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{\alpha}} \leq 1, \widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\beta}} \leq 1,$$

where we replace $\mathbf{C}_{xy}$ with its empirical estimate $\mathbf{X}\mathbf{H}\mathbf{Y}^\top$, and the centered sample matrices $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}$ are defined in (3.1.8). The Lagrangian is written

$$L\left(\hat{\mathbf{c}}, \hat{\mathbf{d}}, \lambda, \xi\right) = \hat{\mathbf{c}}^\top\mathbf{Q}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top\mathbf{R}\hat{\mathbf{d}} - \lambda\left(\hat{\mathbf{c}}^\top\mathbf{Q}^\top\mathbf{Q}\hat{\mathbf{c}} - 1\right) - \xi\left(\hat{\mathbf{d}}^\top\mathbf{R}^\top\mathbf{R}\hat{\mathbf{d}} - 1\right).$$

The derivation of the solution then takes the same form as that used in Appendix A.2.1, and we obtain

$$\begin{bmatrix} \mathbf{0} & \mathbf{Q}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top\mathbf{R} \\ \mathbf{R}^\top\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^\top\mathbf{Q} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{d}} \end{bmatrix} = \tilde{\gamma}\begin{bmatrix} \mathbf{Q}^\top\mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^\top\mathbf{R} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{d}} \end{bmatrix},$$

where the generalised eigenvalue is written $\tilde{\gamma}$ for consistency with (4.2.11) in Section 4.2.2. This may be rearranged as

$$\begin{bmatrix} \mathbf{0} & \left(\mathbf{Q}^\top\mathbf{Q}\right)^-\mathbf{Q}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top\mathbf{R} \\ \left(\mathbf{R}^\top\mathbf{R}\right)^-\mathbf{R}^\top\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^\top\mathbf{Q} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{d}} \end{bmatrix} = \tilde{\gamma}\begin{bmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{d}} \end{bmatrix}.$$

[1] The samples being mapped to their respective feature spaces.

[2] We neglect the constant factor $(\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})^{\frac{1}{2}}$ required in Section 4.2.2; this has no effect on the reasoning in this appendix.

We may premultiply by $\begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$ without changing the values of $\tilde{\gamma}$, since the solution $\begin{bmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{d}} \end{bmatrix}$ above is in the row space of this matrix for any non-zero $\tilde{\gamma}$. Doing this, and substituting the definitions of $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ in (A.2.3), we obtain

$$\begin{bmatrix} \mathbf{P}_Q & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_R \end{bmatrix} \begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top \\ \widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \tilde{\gamma} \begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix}, \tag{A.2.4}$$

where we define the projection operators

$$\mathbf{P}_Q = \mathbf{Q}\left(\mathbf{Q}^\top\mathbf{Q}\right)^{-}\mathbf{Q}^\top, \quad \mathbf{P}_R = \mathbf{R}\left(\mathbf{R}^\top\mathbf{R}\right)^{-}\mathbf{R}^\top.$$

A.3 Canonical correlation: definition and properties

A.3.1 Derivation of the projection directions

In this section, we derive the projection directions for the canonical correlation. The derivation is a standard procedure, and is taken from [13, 55, 37]. The random variables x and y are defined respectively on $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, and we would like to find vectors $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$, $i \in (1, \ldots, s)$ onto which the inputs and outputs respectively project, such that the correlation ρ_i between these projections is a stationary point with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ (the stationary points are given the index i). The *linear variates* $\mathsf{a}_i, \mathsf{b}_i$ are written $\mathsf{a}_i = \boldsymbol{\alpha}_i^\top\left(\mathsf{x} - \mathsf{E}_\mathsf{x}\left(\mathsf{x}\right)\right)$ and $\mathsf{b}_i = \boldsymbol{\beta}_i^\top\left(\mathsf{y} - \mathsf{E}_\mathsf{y}\left(\mathsf{y}\right)\right)$, and the correlation between these variates is

$$\rho_i = \operatorname{corr}\left(\mathsf{a}_i, \mathsf{b}_i\right) = \frac{\boldsymbol{\alpha}_i^\top \mathbf{C}_{xy}\boldsymbol{\beta}_i}{\sqrt{\left(\boldsymbol{\alpha}_i^\top \mathbf{C}_{xx}\boldsymbol{\alpha}_i\right)\left(\boldsymbol{\beta}_i^\top \mathbf{C}_{yy}\boldsymbol{\beta}\right)}}, \tag{A.3.1}$$

where the covariance submatrices are defined in (3.1.1), (3.1.2), and (3.1.3), and it is assumed that $\boldsymbol{\alpha}_i$ is not in the nullspace of $\mathbf{C}_{xx}$, and $\boldsymbol{\beta}_i$ is not in the nullspace of $\mathbf{C}_{yy}$ (these constraints are important when empirical estimates of the covariance matrices are used). We note from the above that $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ can be multiplied by arbitrary constants, and these relations still hold. Let us then normalise $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, such that

$$\boldsymbol{\alpha}^\top \mathbf{C}_{xx}\boldsymbol{\alpha} \leq 1 \quad \text{and} \quad \boldsymbol{\beta}^\top \mathbf{C}_{yy}\boldsymbol{\beta} \leq 1.$$

We then obtain the equivalent Lagrangian

$$L\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \xi\right) = \boldsymbol{\alpha}^\top \mathbf{C}_{xy}\boldsymbol{\beta} - \lambda\left(\boldsymbol{\alpha}^\top \mathbf{C}_{xx}\boldsymbol{\alpha} - 1\right) - \xi\left(\boldsymbol{\beta}^\top \mathbf{C}_{yy}\boldsymbol{\beta} - 1\right).$$

The zero derivatives of the above are the stationary points of $\boldsymbol{\alpha}^\top \mathbf{C}_{xx}\boldsymbol{\beta}$ with respect to the constrained $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ according to Slater's condition [61, Section 7.3]. In other words,

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\alpha}} &= \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda\mathbf{C}_{xx}\boldsymbol{\alpha} = 0, \\ \frac{\partial L}{\partial \boldsymbol{\beta}} &= \mathbf{C}_{xy}^\top\boldsymbol{\alpha} - 2\xi\mathbf{C}_{yy}\boldsymbol{\beta} = 0. \end{aligned}$$

As in Appendix A.2, we can show that $\lambda = \xi$, by premultiplying the first expression by $\boldsymbol{\alpha}^\top$, and the second expression by $\boldsymbol{\beta}^\top$;

$$\begin{aligned} \boldsymbol{\alpha}^\top \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda\boldsymbol{\alpha}^\top \mathbf{C}_{xx}\boldsymbol{\alpha} &= \boldsymbol{\alpha}^\top \mathbf{C}_{xy}\boldsymbol{\beta} - 2\lambda = 0, \\ \boldsymbol{\beta}^\top \mathbf{C}_{xy}^\top\boldsymbol{\alpha} - 2\xi\boldsymbol{\beta}^\top \mathbf{C}_{yy}\boldsymbol{\beta} &= \boldsymbol{\beta}^\top \mathbf{C}_{xy}^\top\boldsymbol{\alpha} - 2\xi = 0. \end{aligned}$$

We therefore write $\rho = 2\lambda = 2\xi$, and the solution becomes a single generalised eigenvalue equation,

$$\begin{bmatrix} 0 & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{A.3.2}$$

An alternative form is

$$\begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = (1+\rho_i) \begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{A.3.3}$$

A.3.2 Properties of the canonical correlation

In this section, we find the number of canonical correlations between two random vectors, and describe the properties of these correlations, as summarised from [55, 37]. We begin with the expression in (A.3.2), where we assume that $\mathbf{C}_{xx}$, $\mathbf{C}_{yy}$, with respective dimensions[3] n_x, n_y, have full rank; and that $\mathbf{C}_{xy}$ has rank $s = \min\{n_x, n_y\}$. We expand out this solution to obtain

$$\begin{bmatrix} \mathbf{C}_{xx}^{-1} & 0 \\ 0 & \mathbf{C}_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{xy}\boldsymbol{\alpha}_i \\ \mathbf{C}_{xy}^\top\boldsymbol{\beta}_i \end{bmatrix} = \rho_i \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix},$$

which becomes

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\boldsymbol{\beta}_i &= \rho_i\boldsymbol{\alpha}_i \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\boldsymbol{\alpha}_i &= \rho_i\boldsymbol{\beta}_i \end{cases}$$

or

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\boldsymbol{\alpha}_i &= \rho_i^2\boldsymbol{\alpha}_i \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\boldsymbol{\beta}_i &= \rho_i^2\boldsymbol{\beta}_i \end{cases}.$$

Using Theorem A.1.4, and applying the spectral decomposition theorem (Theorem A.1.29), reveals that the number of non-zero eigenvalues ρ_i^2 is $s = \min(n_x, n_y)$. Next, taking the first of these expressions, we find

$$\begin{aligned} \rho_i^2\boldsymbol{\alpha}_i &= \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\boldsymbol{\alpha}_i \\ &= \left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xx}^{-1/2}\right)\mathbf{C}_{xy}\left(\mathbf{C}_{yy}^{-1/2}\mathbf{C}_{yy}^{-1/2}\right)\mathbf{C}_{xy}^\top\left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xx}^{1/2}\right)\boldsymbol{\alpha}_i, \end{aligned}$$

and thus

$$\rho_i^2\mathbf{C}_{xx}^{1/2}\boldsymbol{\alpha}_i = \left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}\right)\left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}\right)^\top\mathbf{C}_{xx}^{1/2}\boldsymbol{\alpha}_i.$$

Similarly,

$$\rho_i^2\mathbf{C}_{yy}^{1/2}\boldsymbol{\beta}_i = \left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}\right)^\top\left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}\right)\mathbf{C}_{yy}^{1/2}\boldsymbol{\beta}_i.$$

We may make our notation more compact by defining

$$\mathcal{A} := \begin{bmatrix} \boldsymbol{\alpha}_1 & \dots & \boldsymbol{\alpha}_s \end{bmatrix} \quad \mathcal{B} := \begin{bmatrix} \boldsymbol{\beta}_1 & \dots & \boldsymbol{\beta}_s \end{bmatrix}, \quad \text{and} \quad \mathbf{B} := \mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}.$$

Thus, using the singular value decomposition (Theorem A.1.25), the columns of $\mathbf{C}_{xx}^{1/2}\mathcal{A}$, which are the eigenvectors of $\mathbf{B}\mathbf{B}^\top$, satisfy

$$\left(\mathbf{C}_{xx}^{1/2}\mathcal{A}\right)^\top\mathbf{C}_{xx}^{1/2}\mathcal{A} = \mathcal{A}^\top\mathbf{C}_{xx}\mathcal{A} = \mathbf{I}_s,$$

and the columns of $\mathbf{C}_{yy}^{1/2}\mathcal{B}$, which are the eigenvectors of $\mathbf{B}^\top\mathbf{B}$, satisfy

$$\left(\mathbf{C}_{yy}^{1/2}\mathcal{B}\right)^\top\mathbf{C}_{yy}^{1/2}\mathcal{B} = \mathcal{B}^\top\mathbf{C}_{yy}\mathcal{B} = \mathbf{I}_s.$$

Finally, given that the singular value decomposition of $\mathbf{B}$ satisfies

$$\mathbf{B} = \left(\mathbf{C}_{xx}^{1/2}\mathcal{A}\right)\operatorname{diag}\left(\begin{bmatrix} \rho_1 & \dots & \rho_s \end{bmatrix}\right)\left(\mathbf{C}_{yy}^{1/2}\mathcal{B}\right)^\top,$$

we have

$$\mathcal{A}^\top\mathbf{C}_{xy}\mathcal{B} = \left(\mathbf{C}_{xx}^{1/2}\mathcal{A}\right)^\top\left(\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}\right)\left(\mathbf{C}_{yy}^{1/2}\mathcal{B}\right) = \operatorname{diag}\left(\begin{bmatrix} \rho_1 & \dots & \rho_s \end{bmatrix}\right).$$

[3] The dimensions follow from our simplifying assumption that $\mathcal{F}_\mathcal{X}$ and $\mathcal{X}$ coincide, as do $\mathcal{F}_\mathcal{Y}$ and $\mathcal{Y}$.

A.3.3 A geometric interpretation, incorporating the sample

The following interpretation of the kernel canonical correlation is taken from [55], and is included since it is crucial in understanding why kernel CCA requires the regularisation used in [7, 62] (as seen in Appendix A.6.1). We begin with a sample $((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m))$ of size m, where the points are mapped to their respective feature spaces: $\mathbf{x}_i \in \mathcal{F}_\mathcal{X}$ and $\mathbf{y}_i \in \mathcal{F}_\mathcal{Y}$. We use the data matrices defined in (3.1.5), and the empirical covariances in (3.1.6).

We now recall the equations describing the solution to the canonical correlation problem: from (A.3.2), these are

$$\left\{ \begin{array}{rcl} \mathbf{C}_{xy}\boldsymbol{\beta}_i &=& \rho_i \mathbf{C}_{xx}\boldsymbol{\alpha}_i \\ \mathbf{C}_{xy}^\top \boldsymbol{\alpha}_i &=& \rho_i \mathbf{C}_{yy}\boldsymbol{\beta}_i \end{array} \right. .$$

If we explicitly incorporate the empirical expressions for the covariance matrices from (3.1.6), we get

$$\left\{ \begin{array}{rcl} \widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_i &=& \rho_i \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_i \\ \widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_i &=& \rho_i \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_i \end{array} \right. ,$$

where $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}$ are the centered sample matrices, as described in (3.1.8). Making the substitutions $\mathbf{a}_i = \widetilde{\mathbf{X}}^\top \boldsymbol{\alpha}_i$ and $\mathbf{b}_i = \widetilde{\mathbf{Y}}^\top \boldsymbol{\beta}_i$ for the linear variates, we obtain

$$\left\{ \begin{array}{rcl} \widetilde{\mathbf{X}}\mathbf{b}_i &=& \rho_i \widetilde{\mathbf{X}}\mathbf{a}_i, \\ \widetilde{\mathbf{Y}}\mathbf{a}_i &=& \rho_i \widetilde{\mathbf{Y}}\mathbf{b}_i, \end{array} \right. .$$

Now since $\mathbf{a}_i$ is in the span of the columns of $\widetilde{\mathbf{X}}^\top$, we may write $\mathbf{a}_i = \widetilde{\mathbf{X}}^\top \boldsymbol{\gamma}_a$ (all steps are repeated in the $\mathbf{b}_i$ case). Thus

$$\left\{ \begin{array}{rcl} \widetilde{\mathbf{X}}\mathbf{b}_i &=& \rho_i \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top \boldsymbol{\gamma}_a \\ \widetilde{\mathbf{Y}}\mathbf{a}_i &=& \rho_i \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^\top \boldsymbol{\gamma}_b \end{array} \right. \quad \text{and hence} \quad \left\{ \begin{array}{rcl} \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)^- \widetilde{\mathbf{X}}\mathbf{b}_i &=& \rho_i \boldsymbol{\gamma}_a \\ \left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^\top\right)^- \widetilde{\mathbf{Y}}\mathbf{a}_i &=& \rho_i \boldsymbol{\gamma}_b \end{array} \right. ,$$

where we use the pseudoinverse to account for the fact that $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ may not have full rank. Premultiplying by $\widetilde{\mathbf{X}}^\top$ and $\widetilde{\mathbf{Y}}^\top$ respectively gives

$$\left\{ \begin{array}{rcl} \widetilde{\mathbf{X}}^\top \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)^- \widetilde{\mathbf{X}}\mathbf{b}_i &=& \rho_i \mathbf{a}_i \\ \widetilde{\mathbf{Y}}^\top \left(\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^\top\right)^- \widetilde{\mathbf{Y}}\mathbf{a}_i &=& \rho_i \mathbf{b}_i \end{array} \right. .$$

Comparing with Theorem A.1.8, we observe that $\rho_i \mathbf{a}_i$ is the projection of $\mathbf{b}_i$ onto the column space of $\widetilde{\mathbf{X}}^\top$, and $\rho_i \mathbf{b}_i$ is the projection of $\mathbf{a}_i$ onto the column space of $\widetilde{\mathbf{Y}}^\top$.

A.3.4 Link with the Gaussian mutual information

In this section, we demonstrate the link between the canonical correlation coefficients in Theorem 3.4.1 and the Gaussian mutual information in (4.1.1). The proof is an expansion on the discussion in [7, Appendix A]. Beginning with (3.4.4), we note that $\begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix}$ is positive definite, and can be written

$$\begin{bmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{bmatrix} = \mathbf{S}\mathbf{S} \qquad \text{where } \mathbf{S} = \begin{bmatrix} \mathbf{C}_{xx}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{yy}^{1/2} \end{bmatrix} \text{ and } \mathbf{C}_{xx} = \mathbf{C}_{xx}^{1/2}\mathbf{C}_{xx}^{1/2},$$

and the matrix square roots are also symmetric. Making this replacement in (3.4.5), writing $\left[\begin{array}{cc} \breve{\boldsymbol{\alpha}}_i^\top & \breve{\boldsymbol{\beta}}_i^\top \end{array}\right]^\top = \mathbf{S}\left[\begin{array}{cc} \boldsymbol{\alpha}_i^\top & \boldsymbol{\beta}_i^\top \end{array}\right]^\top$, and assuming $\mathbf{S}$ has full rank, yields

$$\begin{aligned} \left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\left[\begin{array}{c} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{array}\right] &= (1+\rho_i)\,\mathbf{S}\mathbf{S}\left[\begin{array}{c} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{array}\right] \\ \left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\mathbf{S}^{-1}\left[\begin{array}{c} \breve{\boldsymbol{\alpha}}_i \\ \breve{\boldsymbol{\beta}}_i \end{array}\right] &= (1+\rho_i)\mathbf{S}\left[\begin{array}{c} \breve{\boldsymbol{\alpha}}_i \\ \breve{\boldsymbol{\beta}}_i \end{array}\right] \\ \mathbf{S}^{-1}\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\mathbf{S}^{-1}\left[\begin{array}{c} \breve{\boldsymbol{\alpha}}_i \\ \breve{\boldsymbol{\beta}}_i \end{array}\right] &= (1+\rho_i)\left[\begin{array}{c} \breve{\boldsymbol{\alpha}}_i \\ \breve{\boldsymbol{\beta}}_i \end{array}\right]. \end{aligned}$$

This decomposition is necessary to preserve the symmetry of the left hand matrix, which in turn guarantees the eigenvalues are real. Note that the eigenvalues $1+\rho_i$ for which we are trying to solve have not been changed by this procedure; moreover, when the covariance matrices have full rank, there are $\min(l_x, l_y)$ pairs $\pm\rho_i$ of eigenvalues. Finally, we note that since $\mathbf{S}^{-1}\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\mathbf{S}^{-1}$ is symmetric, the determinant may be written as the product of the eigenvalues,

$$\left|\mathbf{S}^{-1}\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\mathbf{S}^{-1}\right| = \prod_i (1+\rho_i)(1-\rho_i) = \prod_i \left(1-\rho_i^2\right),$$

in accordance with Theorem A.1.27. Using this result, we write the ratio of determinants in (4.1.1) as

$$\begin{aligned} \frac{\left|\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\right|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} &= \frac{\left|\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\right|}{\left|\left[\begin{array}{cc} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{array}\right]\right|} = \frac{\left|\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\right|}{|\mathbf{S}\mathbf{S}|} \\ &= \left|\mathbf{S}^{-1}\left[\begin{array}{cc} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{array}\right]\mathbf{S}^{-1}\right| = \prod_i \left(1-\rho_i^2\right). \end{aligned}$$

A.4 Approximate mutual information between discretised distributions

In this section, we derive the approximate mutual information near independence for two discrete random variables. The proof is taken from [7, Appendix B]. We define the joint distribution $\mathbf{P}_{\hat{\mathrm{x}},\hat{\mathrm{y}}}(i,j)$, where $i \in \{1,\ldots,l_x\}$ and $j \in \{1,\ldots,l_y\}$, and the associated marginal distributions $\mathbf{P}_{\hat{\mathrm{x}}}(i)$ and $\mathbf{P}_{\hat{\mathrm{y}}}(j)$. Writing $\mathbf{P}_{\hat{\mathrm{x}},\hat{\mathrm{y}}}(i,j) = \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,(1+\epsilon_{i,j})$ for an appropriate choice of $\epsilon_{i,j}$, where $\epsilon_{i,j}$ is small near independence, we find

$$\begin{aligned} I(\hat{\mathrm{x}};\hat{\mathrm{y}}) &= \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}},\hat{\mathrm{y}}}(i,j)\log\left(\frac{\mathbf{P}_{\hat{\mathrm{x}},\hat{\mathrm{y}}}(i,j)}{\mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)}\right) \\ &= \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,(1+\epsilon_{i,j})\log(1+\epsilon_{i,j}) \\ &\approx \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\left(\epsilon_{i,j}+\epsilon_{i,j}^2/2\right) \\ &= \frac{1}{2}\sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,\epsilon_{i,j}^2 \end{aligned}$$

using $(1+\epsilon)\log(1+\epsilon) \approx \epsilon + \epsilon^2/2$ for small ϵ, and

$$\begin{aligned}\sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}}}(i)\,\mathsf{P}_{\hat{\mathsf{y}}}(j)\,\epsilon_{i,j} &= \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}}}(i)\,\mathsf{P}_{\hat{\mathsf{y}}}(j)\,(1+\epsilon_{i,j}) - \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}}}(i)\,\mathsf{P}_{\hat{\mathsf{y}}}(j) \\ &= \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) - \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathsf{P}_{\hat{\mathsf{x}}}(i)\,\mathsf{P}_{\hat{\mathsf{y}}}(j) \\ &= 1-1=0.\end{aligned}$$

A.5 Approximate mutual information between 2 Gaussians

A.5.1 Ratio of determinants for the Gaussian mutual information

The following result is stated without proof in [7, Appendix B]. Recall that the mutual information between x_G and y_G is

$$I(\mathsf{x}_G;\mathsf{y}_G) = -\frac{1}{2}\log\left(\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|}\right) = -\frac{1}{2}\log\left(\frac{\left|\begin{bmatrix}\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top & \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\end{bmatrix}\right|}{\left|\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right|\left|\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right|}\right). \quad \text{(A.5.1)}$$

We make the expansion

$$\begin{aligned}\mathbf{P}_{xy} &= \mathbf{p}_x\mathbf{p}_y^\top + \mathbf{D}_x\boldsymbol{\epsilon}\mathbf{D}_y & \text{(A.5.2)}\\ &= \mathbf{D}_x\left(\mathbf{1}_{l_x}\mathbf{1}_{l_y}^\top + \boldsymbol{\epsilon}\right)\mathbf{D}_y & \text{(A.5.3)}\end{aligned}$$

where $\boldsymbol{\epsilon}$ is the matrix which quantifies the departure from independence; this means $(\mathbf{P}_{xy})_{i,j} = (\mathbf{p}_x)_i(\mathbf{p}_y)_j(1+(\boldsymbol{\epsilon})_{i,j})$. Then we use Theorems A.1.15 and A.1.16 to obtain

$$\begin{aligned}\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} &= \frac{\left|\begin{bmatrix}\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top & \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\end{bmatrix}\right|}{\left|\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right|\left|\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right|} \\ &= \frac{\left|\left(\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right) - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\right|}{\left|\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right|} \\ &= \left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right)^{-1}\right| \\ &= \left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right)^{-1}\right|\end{aligned}$$

In the reasoning above, we glossed over the fact that both $\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top$ and $\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top$ have rank at most $l_x - 1$ and $l_y - 1$ respectively, and are not invertible[4]. To see this, we make the expansions

$$\begin{aligned}\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top &= \mathbf{D}_x\left(\mathbf{I}_{l_x} - \mathbf{1}_{l_x}\mathbf{p}_x^\top\right) = \mathbf{D}_x\mathbf{E}_x, \\ \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top &= \mathbf{D}_y\left(\mathbf{I}_{l_y} - \mathbf{1}_{l_y}\mathbf{p}_y^\top\right) = \mathbf{D}_y\mathbf{E}_y,\end{aligned}$$

where $\mathbf{E}_x := \mathbf{I}_{l_x} - \mathbf{1}_{l_x}\mathbf{p}_x^\top$ and $\mathbf{E}_y := \mathbf{I}_{l_y} - \mathbf{1}_{l_y}\mathbf{p}_y^\top$. Recall that normalisation requires

$$\mathbf{p}_x^\top\mathbf{1}_{l_x} = 1, \qquad \mathbf{p}_y^\top\mathbf{1}_{l_y} = 1.$$

[4] This is *not* to say that the ratio of determinants in (A.5.1) is undefined: rather, we must define it as a limit using matrices of full rank.

Consequently, the final column of both $\mathbf{E}_x$ and $\mathbf{E}_y$ may be written as the negative sum of the remaining columns.

We may get around this problem by adding a small diagonal term $\xi\mathbf{I}$ to both $\mathbf{E}_x$ and $\mathbf{E}_y$, so that

$$\tilde{\mathbf{E}}_x := (1+\xi)\,\mathbf{I}_{l_x} - \mathbf{1}_{l_x}\mathbf{p}_x^\top \qquad \tilde{\mathbf{E}}_y := (1+\xi)\,\mathbf{I}_{l_y} - \mathbf{1}_{l_y}\mathbf{p}_y^\top.$$

The resulting matrices are invertible, and may be made arbitrarily close to $\mathbf{E}_x$ and $\mathbf{E}_y$ as ξ drops to zero. We may therefore use this limiting case to determine the ratio of determinants in (A.5.1). The inverses of $\mathbf{D}_x\tilde{\mathbf{E}}_x$ and $\mathbf{D}_y\tilde{\mathbf{E}}_y$ may be greatly simplified, using the results

$$\mathbf{D}_x^{-1}\mathbf{p}_x = \mathbf{1}_{l_x} \quad \text{and} \quad \mathbf{p}_x^\top\mathbf{D}_x^{-1}\mathbf{p}_x = \mathbf{p}_x^\top\mathbf{1}_{l_x} = 1,$$

(with analogous results for $\mathbf{D}_y$ and $\mathbf{p}_y$) along with Theorem A.1.21 to expand out the inverses;

$$\begin{aligned}
\left(\mathbf{D}_x\tilde{\mathbf{E}}_x\right)^{-1} &= \left(\mathbf{D}_x\,(1+\xi) - \mathbf{p}_x\mathbf{p}_x^\top\right)^{-1} \\
&= (1+\xi)^{-1}\,\mathbf{D}_x^{-1} + \frac{(1+\xi)^{-2}\,\mathbf{D}_x^{-1}\mathbf{p}_x\mathbf{p}_x^\top\mathbf{D}_x^{-1}}{1-(1+\xi)^{-1}\,\mathbf{p}_x^\top\mathbf{D}_x^{-1}\mathbf{p}_x} \\
&= (1+\xi)^{-1}\,\mathbf{D}_x^{-1} + \frac{\mathbf{D}_x^{-1}\mathbf{p}_x\mathbf{p}_x^\top\mathbf{D}_x^{-1}}{(1+\xi)-\mathbf{p}_x^\top\mathbf{D}_x^{-1}\mathbf{p}_x} \\
&= (1+\xi)^{-1}\,\mathbf{D}_x^{-1} + \frac{\mathbf{D}_x^{-1}\mathbf{p}_x\mathbf{p}_x^\top\mathbf{D}_x^{-1}}{\xi} \\
&= (1+\xi)^{-1}\,\mathbf{D}_x^{-1} + \xi^{-1}\mathbf{1}_{l_x}\mathbf{1}_{l_x}^\top.
\end{aligned}$$

We know, however, that

$$\begin{aligned}
\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{1}_{l_y}\mathbf{1}_{l_y}^\top &= \begin{bmatrix} \sum_{j=1}^{l_y}(\mathbf{P}_{xy})_{1,j} & \cdots & \sum_{j=1}^{l_y}(\mathbf{P}_{xy})_{1,j} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{l_y}(\mathbf{P}_{xy})_{l_x,j} & \cdots & \sum_{j=1}^{l_y}(\mathbf{P}_{xy})_{l_x,j} \end{bmatrix} \\
&\quad - \begin{bmatrix} \sum_{j=1}^{l_y}\left(\mathbf{p}_x\mathbf{p}_y^\top\right)_{1,j} & \cdots & \sum_{j=1}^{l_y}\left(\mathbf{p}_x\mathbf{p}_y^\top\right)_{1,j} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{l_y}\left(\mathbf{p}_x\mathbf{p}_y^\top\right)_{l_x,j} & \cdots & \sum_{j=1}^{l_y}\left(\mathbf{p}_x\mathbf{p}_y^\top\right)_{l_x,j} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{p}_x & \cdots & \mathbf{p}_x \end{bmatrix} - \begin{bmatrix} \mathbf{p}_x & \cdots & \mathbf{p}_x \end{bmatrix} = \mathbf{0},
\end{aligned}$$

with an analogous result for $\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{1}_{l_x}\mathbf{1}_{l_x}^\top$. Therefore,

$$\begin{aligned}
\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x\tilde{\mathbf{E}}_x\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{D}_y\tilde{\mathbf{E}}_y\right)^{-1} &= (1+\xi)^{-2}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top \\
&\quad \mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right)^{-1}\right| &= \\
\lim_{\xi\to 0}\left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\mathbf{D}_x\tilde{\mathbf{E}}_x\right)^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{D}_y\tilde{\mathbf{E}}_y\right)^{-1}\right| &= \\
\lim_{\xi\to 0}\left|\mathbf{I}_{l_y} - (1+\xi)^{-2}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}\right| &= \\
\left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}\right|.
\end{aligned}$$

A.5.2 Approximation to the mutual information near independence

The following result is proved in [7, Appendix B]. We wish to approximate the ratio of determinants

$$\begin{aligned}\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}||\mathbf{C}_{yy}|} &= \left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top \mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}\right| \\ &= \left|\mathbf{I}_{l_y} - \mathbf{D}_y^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top \mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\right|\end{aligned}$$

near independence. We again start with the expansion in (A.5.2), from which it follows that

$$\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top = \mathbf{D}_x\boldsymbol{\epsilon}\mathbf{D}_y.$$

Thus

$$\begin{aligned}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top \mathbf{D}_x^{-1}\left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1} &= \left(\mathbf{D}_x\boldsymbol{\epsilon}\mathbf{D}_y\right)^\top \mathbf{D}_x^{-1}\left(\mathbf{D}_x\boldsymbol{\epsilon}\mathbf{D}_y\right)\mathbf{D}_y^{-1} \\ &= \mathbf{D}_y\boldsymbol{\epsilon}^\top\mathbf{D}_x\boldsymbol{\epsilon}.\end{aligned}$$

Next, we write

$$\begin{aligned}\left|\mathbf{I}_{l_y} - \mathbf{D}_y\boldsymbol{\epsilon}^\top\mathbf{D}_x\boldsymbol{\epsilon}\right| &= \left|\mathbf{I}_{l_y} - \mathbf{D}_y^{1/2}\mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x^{1/2}\mathbf{D}_x^{1/2}\boldsymbol{\epsilon}\right| \\ &= \left|\mathbf{D}_y^{1/2}\right|\left|\mathbf{D}_y^{-1/2} - \mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x^{1/2}\mathbf{D}_x^{1/2}\boldsymbol{\epsilon}\right| \\ &= \left|\mathbf{I}_{l_y} - \mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x^{1/2}\mathbf{D}_x^{1/2}\boldsymbol{\epsilon}\mathbf{D}_y^{1/2}\right|.\end{aligned}$$

Now since $\mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x^{1/2}\mathbf{D}_x^{1/2}\boldsymbol{\epsilon}\mathbf{D}_y^{1/2}$ is symmetric, we may write it as

$$\mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x^{1/2}\mathbf{D}_x^{1/2}\boldsymbol{\epsilon}\mathbf{D}_y^{1/2} = \mathbf{E}\Lambda\mathbf{E}^\top,$$

where $\mathbf{E}\mathbf{E}^\top = \mathbf{I}$, using Theorem A.1.29. Thus

$$\begin{array}{rcccl}-\frac{1}{2}\log\left(\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}||\mathbf{C}_{yy}|}\right) & = & -\frac{1}{2}\log\left(\left|\mathbf{I} - \mathbf{E}\Lambda\mathbf{E}^\top\right|\right) & = & -\frac{1}{2}\log\left(|\mathbf{E}|\,|\mathbf{I} - \Lambda|\,\left|\mathbf{E}^\top\right|\right) \\ & = & -\frac{1}{2}\log\left(\prod_j(1-\lambda_j)\right) & = & -\frac{1}{2}\sum_j\log(1-\lambda_j) \\ & = & -\frac{1}{2}\sum_j\left(-\lambda_j - \frac{\lambda_j^2}{2} - \frac{\lambda_j^3}{3} - \ldots\right) & \approx & \frac{1}{2}\sum_j\lambda_j \\ & = & \frac{1}{2}\mathrm{tr}\left(\mathbf{D}_y^{1/2}\boldsymbol{\epsilon}^\top\mathbf{D}_x\boldsymbol{\epsilon}\mathbf{D}_y^{1/2}\right) & = & \sum_{i=1}^{l_x}\sum_{j=1}^{l_y}(\boldsymbol{\epsilon})_{i,j}^2\,(\mathbf{p}_x)_i\,(\mathbf{p}_y)_j,\end{array}$$

where we use Theorem A.1.28 in the penultimate step.

A.6 Discussion of Bach and Jordan's derivation of the KGV

A.6.1 Computation of the unregularised kernel canonical correlations

In this section, we prove Lemma 3.4.5, which is used to show a regularised empirical estimate for the kernel canonical correlates is needed when the associated RKHSs have high dimension. We begin with (3.4.6), which we restate below for reference;

$$\begin{bmatrix}\mathbf{0} & \widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \\ \widetilde{\mathbf{K}}_{mm}^{(y)}\widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{c}_i \\ \mathbf{d}_i\end{bmatrix} = \rho_i\begin{bmatrix}\left(\widetilde{\mathbf{K}}_{mm}^{(x)}\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(\widetilde{\mathbf{K}}_{mm}^{(y)}\right)^2\end{bmatrix}\begin{bmatrix}\mathbf{c}_i \\ \mathbf{d}_i\end{bmatrix}.$$

This is equivalent to

$$\begin{bmatrix}\mathbf{0} & \left(\mathbf{K}_{mm}^{(x)-}\right)^2\widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \\ \left(\mathbf{K}_{mm}^{(y)-}\right)^2\widetilde{\mathbf{K}}_{mm}^{(y)}\widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{c}_i \\ \mathbf{d}_i\end{bmatrix} = \rho_i\begin{bmatrix}\mathbf{c}_i \\ \mathbf{d}_i\end{bmatrix},$$

where we use the pseudoinverses since the Gram matrices do not have full rank. If we recall that $\mathbf{H}$ is the centering matrix, then the solutions ρ_i correspond to the solutions of

$$\begin{aligned}
0 &= \begin{vmatrix} -\rho\mathbf{I} & \left(\mathbf{K}_{mm}^{(x)-}\right)^2 \widetilde{\mathbf{K}}_{mm}^{(x)} \widetilde{\mathbf{K}}_{mm}^{(y)} \\ \left(\mathbf{K}_{mm}^{(y)-}\right)^2 \widetilde{\mathbf{K}}_{mm}^{(y)} \widetilde{\mathbf{K}}_{mm}^{(x)} & -\rho\mathbf{I} \end{vmatrix} \\
&= |\rho\mathbf{I}| \left| \rho\mathbf{I} - \frac{1}{\rho} \left(\mathbf{K}_{mm}^{(y)-}\right)^2 \widetilde{\mathbf{K}}_{mm}^{(y)} \widetilde{\mathbf{K}}_{mm}^{(x)} \left(\mathbf{K}_{mm}^{(x)-}\right)^2 \widetilde{\mathbf{K}}_{mm}^{(x)} \widetilde{\mathbf{K}}_{mm}^{(y)} \right| \\
&= |\rho\mathbf{I}| \left| \rho\mathbf{I} - \frac{1}{\rho}\mathbf{H} \right| \\
&= \rho^m \frac{\left(\rho^2 - 1\right)^{m-1}}{\rho^{m-2}},
\end{aligned}$$

which has $m-1$ roots $+1$, $m-1$ roots -1, and 2 roots 0. To avoid this problem, a regularised empirical estimate is used, as discussed in Bach and Jordan [7].

A.6.2 Further discussion of KGV proof

In this section, we describe some possible problems in the derivation of the kernel generalised variance in [7, Appendix B]. We assume that $\mathcal{X}$ and $\mathcal{Y}$ are both bounded intervals on $\mathbb{R}$. Recall from Definition 3.4.4 that the kernel canonical correlations can be written

$$\rho = \operatorname{corr}\left(f(\mathsf{x}), g(\mathsf{y})\right) \tag{A.6.1}$$

$$= \frac{\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(f(\mathsf{x})g(\mathsf{y})\right) - \mathbf{E}_{\mathsf{x}}\left(f(\mathsf{x})\right)\mathbf{E}_{\mathsf{y}}\left(g(\mathsf{y})\right)}{\sqrt{\mathbf{E}_{\mathsf{x}}\left(f^2(\mathsf{x})\right) - \mathbf{E}_{\mathsf{x}}^2\left(f(\mathsf{x})\right)}\sqrt{\mathbf{E}_{\mathsf{y}}\left(g^2(\mathsf{y})\right) - \mathbf{E}_{\mathsf{y}}^2\left(g(\mathsf{y})\right)}}, \tag{A.6.2}$$

where $f \in \mathcal{F}_{\mathcal{X}}$ and $g \in \mathcal{F}_{\mathcal{Y}}$. We approximate f, g using the expansions

$$f(\mathsf{x}) \approx \sum_{i=1}^{l_x} \hat{c}_i k(\mathsf{x} - q_i) = \hat{\mathbf{c}}^\top \mathbf{k}_l^{(\mathsf{x})}, \tag{A.6.3}$$

$$g(\mathsf{y}) \approx \sum_{j=1}^{l_y} \hat{d}_j k(\mathsf{y} - y_j) = \hat{\mathbf{d}}^\top \mathbf{k}_l^{(\mathsf{y})}, \tag{A.6.4}$$

where[5]

$$\mathbf{k}_l^{(\mathsf{x})} = \begin{bmatrix} \langle \mathsf{x}, \mathbf{q}_1 \rangle & \cdots & \langle \mathsf{x}, \mathbf{q}_{l_x} \rangle \end{bmatrix}^\top, \quad \mathbf{k}_l^{(\mathsf{y})} = \begin{bmatrix} \langle \mathsf{y}, \mathbf{r}_1 \rangle & \cdots & \langle \mathsf{y}, \mathbf{r}_{l_y} \rangle \end{bmatrix}^\top, \tag{A.6.5}$$

and the grids $\boldsymbol{q}$ and $\boldsymbol{r}$ over $\mathcal{X}$ and $\mathcal{Y}$ are defined in Section 4.1.2. Substituting these approximations in (A.6.2), we get

$$\begin{aligned}
\rho \approx \hat{\rho} := \; & \hat{\mathbf{c}}^\top \left(\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\mathbf{k}_l^{(\mathsf{x})} \left(\mathbf{k}_l^{(\mathsf{y})} \right)^\top \right) - \mathbf{E}_{\mathsf{x}} \left(\mathbf{k}_l^{(\mathsf{x})} \right) \mathbf{E}_{\mathsf{y}} \left(\mathbf{k}_l^{(\mathsf{y})} \right)^\top \right) \hat{\mathbf{d}} \\
& \times \left[\hat{\mathbf{c}}^\top \left(\mathbf{E}_{\mathsf{x}} \left(\mathbf{k}_l^{(\mathsf{x})} \left(\mathbf{k}_l^{(\mathsf{x})} \right)^\top \right) - \mathbf{E}_{\mathsf{x}} \left(\mathbf{k}_l^{(\mathsf{x})} \right) \mathbf{E}_{\mathsf{x}} \left(\mathbf{k}_l^{(\mathsf{x})} \right)^\top \right) \hat{\mathbf{c}} \right]^{-1/2} \\
& \times \left[\hat{\mathbf{d}}^\top \left(\mathbf{E}_{\mathsf{y}} \left(\mathbf{k}_l^{(\mathsf{y})} \left(\mathbf{k}_l^{(\mathsf{y})} \right)^\top \right) - \mathbf{E}_{\mathsf{y}} \left(\mathbf{k}_l^{(\mathsf{y})} \right) \mathbf{E}_{\mathsf{y}} \left(\mathbf{k}_l^{(\mathsf{y})} \right)^\top \right) \hat{\mathbf{d}} \right]^{-1/2}.
\end{aligned}$$

Replacing the random vectors x, y with the observations $\mathbf{X}, \mathbf{Y}$, we obtain (4.2.21) in Section 4.2.4, which is then used to derive the KGV. It is still not clear in what manner the population expression

[5] Note that the superscripts $(\mathsf{x}), (\mathsf{y})$ used below are *sans serif*, which indicates that x, y in the inner products are random vectors, and *not* the sample $\boldsymbol{x}, \boldsymbol{y}$.

$\hat{\rho}$ above relates to the Gaussian approximation to the discretised mutual information ((4.1.13) in Section 4.1.3): we now address this problem.

We begin by restating the argument of the logarithm in (4.1.13) in the form found in Appendix A.5.1; in other words,

$$\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} = \frac{\left|\begin{bmatrix} \mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top & \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top \end{bmatrix}\right|}{\left|\mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top\right|\left|\mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top\right|}$$

Comparing with our expression for $\hat{\rho}$, we observe that the link between the Gaussian approximation to the discrete mutual information and the KGV could be shown by demonstrating

$$\mathbf{P}_{xy} \overset{?}{\approx} \mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right), \quad \mathbf{D}_x \overset{?}{\approx} \mathsf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{x})}\right)^\top\right), \quad \mathbf{p}_x \overset{?}{\approx} \mathsf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right) \tag{A.6.6}$$

under appropriate conditions, with similar results for the terms in y (certain constants related to the grid spacing are neglected here: see below). We consider the case where both kernels are Gaussian; that is,

$$\begin{aligned} \langle \mathbf{x}, \mathbf{q}_i \rangle = k(x - q_i) &= \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-q_i)^2}{2\sigma_x^2}\right), \\ \langle \mathbf{y}, \mathbf{r}_j \rangle = k(y - r_j) &= \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(y-r_i)^2}{2\sigma_y^2}\right), \end{aligned}$$

bearing in mind that the impulse function is a limiting case [14];

$$\delta_{q_i}(x) = \lim_{\sigma_x \to 0} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-q_i)^2}{2\sigma_x^2}\right) := \lim_{\sigma_x \to 0} k\left(x - q_i\right).$$

To compute the covariance structure of the vectors in (A.6.5), we require expressions for the expectations

$$\mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right), \qquad \mathsf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right), \qquad \mathsf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{x})}\right)^\top\right)$$
$$\mathsf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right), \quad \mathsf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\right)$$

The expectation of individual entries in the matrix $\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top$ is

$$\begin{aligned} \mathsf{E}_{\mathsf{x},\mathsf{y}}\left[\langle \mathbf{x}, \mathbf{q}_i \rangle \langle \mathbf{y}, \mathbf{r}_j \rangle\right] &= \int_{\mathcal{X}}\int_{\mathcal{Y}} k(x - q_i)k(y - r_j)\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy \\ &= \left[k(x)k(y) \star \mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)\right](q_i, r_j), \end{aligned}$$

which is the convolution of the product of kernels with the underlying (unknown) density $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)$ of the random variables x,y in input space, evaluated at q_i, r_j. Since the kernels are normalised, then the above expectation is also a *probability density* $\mathbf{f}^k_{\mathsf{x},\mathsf{y}}(x,y)$, where the superscript k indicates that it represents the density $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)$ smoothed by $k(x)k(y)$. Similarly,

$$\begin{aligned} \mathsf{E}_{\mathsf{x}}\left[\langle \mathbf{x}, \mathbf{q}_i \rangle \langle \mathbf{x}, \mathbf{q}_j \rangle\right] &= \int_{\mathcal{X}} k(x - q_i)k(x - q_j)\mathbf{f}_{\mathsf{x}}(x)dx \\ &\approx \begin{cases} \left[k^2(x) \star \mathbf{f}_{\mathsf{x}}(x)\right](q_i) & i = j \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

where the above assumes $\sigma_x \ll \Delta_x \ll 1$. Note, however, that

$$k^2(x-q_i) = \frac{1}{2\pi\sigma_x^2}\exp\left(-\frac{(x-q_i)^2}{\sigma_x^2}\right) \quad \text{(A.6.7)}$$

$$= \frac{1}{2\sigma_x\sqrt{\pi}} \times \frac{1}{\sqrt{\pi\sigma_x^2}}\exp\left(-\frac{(x-q_i)^2}{\sigma_x^2}\right), \quad \text{(A.6.8)}$$

and thus $k^2(x)$ is *not* a probability density (the integral over $\mathbb{R}$ is equal to $\frac{1}{2\sigma_x\sqrt{\pi}}$). Finally,

$$\begin{aligned}\mathbf{E}_{\mathsf{x}}\left[\langle \mathbf{x}, \mathbf{q}_i\rangle\right] &= \int_{\mathbb{R}} k(x-q_i)\mathbf{f}_{\mathsf{x}}(x)dx \\ &= \left[k(x) \star \mathbf{f}_{\mathsf{x}}(x)\right](q_i).\end{aligned}$$

In the light of these observations, it might seem that the relations in (A.6.6) ought to hold in the limit as $\Delta_x, \Delta_y \to 0$ and $\sigma_x, \sigma_y \to 0$, so long as $\sigma_x \ll \Delta_x$ and $\sigma_y \ll \Delta_y$: the grid size must be small to allow us to make the approximations

$$\mathbf{P}_{\hat{\mathsf{x}}}(i) = \int_{q_i}^{q_i+\Delta_x} \mathbf{f}_{\mathsf{x}}(x)\,dx \approx \Delta_x \mathbf{f}_{\mathsf{x}}(q_i)$$

and

$$\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) = \int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y} \mathbf{f}_{\mathsf{x},\mathsf{y}}(xy)dxdy \approx \Delta_x\Delta_y \mathbf{f}_{\mathsf{x},\mathsf{y}}(q_i, r_j),$$

and the kernel size is made small so that the kernel functions approach delta functions (although the *squared* kernel functions do not do so). Ignoring for the moment the problem of dealing with the factors Δ_x, Δ_y (which are in any case assumed to decrease much more slowly than the kernel sizes, and do not impact upon the limiting argument below), we can write population expression for the kernel generalised variance, in the limit of small kernel size, as

$$\begin{aligned}&\lim_{\sigma_x,\sigma_y\to 0} \mathcal{N}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}\right) \\ =\ &\lim_{\sigma_x,\sigma_y\to 0} -\frac{1}{2}\log\left(\left|\mathbf{I} - \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right) - \mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right)^\top\right.\right. \\ &\quad \times \left(\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{x})}\right)^\top\right) - \mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right)\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right)^\top\right)^{-1} \\ &\quad \times \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right) - \mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right) \\ &\quad \left.\left. \times \left(\mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right) - \mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right)^{-1}\right|\right) \\ \approx\ &\lim_{\sigma_x,\sigma_y\to 0} -\frac{1}{2}\log\left(\left|\mathbf{I} - \mathbf{0}\left(\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^{(\mathsf{x})}\left(\mathbf{k}_l^{(\mathsf{x})}\right)^\top\right) - \mathbf{0}\right)^{-1}\mathbf{0}\left(\mathbf{E}_{\mathsf{y}}\left(\mathbf{k}_l^{(\mathsf{y})}\left(\mathbf{k}_l^{(\mathsf{y})}\right)^\top\right) - \mathbf{0}\right)^{-1}\right|\right) \\ =\ &0,\end{aligned}$$

where the penultimate line retains only those terms that remain large in the limit as σ_x, σ_y approach zero (in other words, the smaller terms are set to zero), and we use the expression for the squared kernel in (A.6.8). This problem reveals the need to enforce the *opposite assumption* to that made above, namely that $\sigma_x \gg \Delta_x$ and $\sigma_y \gg \Delta_y$ (see Section 4.2).

A.7 Some miscellaneous proofs

A.7.1 Effect on norm of taking sums of rows

In this section, we prove Theorem 4.2.2; namely, that if $\mathbf{B}$ is a symmetric $n \times n$ matrix with positive elements $b_{i,j}$, and $\mathbf{c}$ an arbitrary $n \times 1$ vector with elements c_i, then

$$\mathbf{c}^\top \mathrm{diag}\left(\mathbf{B}\mathbf{1}_n\right)\mathbf{c} \geq \mathbf{c}^\top \mathbf{B}\mathbf{c}.$$

We first give the expansion

$$\begin{aligned}\mathbf{c}^\top \mathrm{diag}\left(\mathbf{B}\mathbf{1}_n\right)\mathbf{c} &= \sum_{i=1}^{n} c_i^2 \left(\sum_{j=1}^{m} b_{i,j}\right) \\ &= \sum_{i=1}^{n} c_i^2 b_{i,i} + \sum_{i=1}^{n}\sum_{j=i+1}^{n} b_{i,j}\left(c_i^2 + c_j^2\right).\end{aligned}$$

Next, we expand

$$\mathbf{c}^\top \mathbf{B}\mathbf{c} = \sum_{i=1}^{n} c_i^2 b_{i,i} + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} c_i c_j b_{i,j}.$$

For any $c_i c_j \in \mathbb{R}$,

$$\begin{aligned}c_i^2 + c_j^2 &= 2c_i c_j + \left(c_i - c_j\right)^2 \\ &\geq 2c_i c_j.\end{aligned}$$

By comparing terms with equal $b_{i,j}$, and bearing in mind that $b_{i,j} \geq 0$, we complete the proof.

A.7.2 The centered kernel matrix is singular

In this section, we show that $\widetilde{\mathbf{K}}_{mm}^{(x)}$ is singular. To see why, we first make use of Theorem A.1.9, which yields

$$\left|\widetilde{\mathbf{K}}_{mm}^{(x)}\right| = |\mathbf{H}|\left|\mathbf{K}_{mm}^{(x)}\right||\mathbf{H}|,$$

and $|\mathbf{H}| = \left|\mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right| = 0$ using Theorem A.1.14, since its rank is less than m (any column of $\mathbf{H}$ can be expressed as the negative of the sum of the remaining columns).

A.7.3 Proof that centering matrix is idempotent

We define

$$\mathbf{H} = \mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top.$$

Then

$$\begin{aligned}\mathbf{H}\mathbf{H} &= \left(\mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right)\left(\mathbf{I} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right) \\ &= \mathbf{I} - \frac{2}{m}\mathbf{1}_m\mathbf{1}_m^\top + \frac{1}{m^2}\mathbf{1}_m\left(\mathbf{1}_m^\top\mathbf{1}_m\right)\mathbf{1}_m^\top \\ &= \mathbf{I} - \frac{2}{m}\mathbf{1}_m\mathbf{1}_m^\top + \frac{m}{m^2}\mathbf{1}_m\mathbf{1}_m^\top \\ &= \mathbf{H}.\end{aligned}$$

A.8 Basic results in information theory

The following material is taken from [25, 42]. We mostly omit the proofs, since the intent is simply to give an overview of useful results and definitions.

A.8.1 Information theory in discrete spaces

Any signal (for instance, a piece of music or a film) may be interpreted as the output of a random process. This output can be characterised by the entropy, a quantity introduced by Shannon to describe the lower bound on how far the signal can be compressed.

Definition A.8.1 (Entropy). The entropy of a discrete random variable $\hat{\mathsf{x}}$ with possible values $i \in \{1, \ldots, n\}$ and probabilities $\mathsf{P}_{\hat{\mathsf{x}}}(i)$ is

$$H(\hat{\mathsf{x}}) = -\sum_{i=1}^{n} \mathsf{P}_{\hat{\mathsf{x}}}(i) \log_2 \left(\mathsf{P}_{\hat{\mathsf{x}}}(i)\right) = \mathsf{E}_{\hat{\mathsf{x}}} \left(-\log_2 \left(\mathsf{P}_{\hat{\mathsf{x}}}(\hat{\mathsf{x}})\right)\right).$$

This can be interpreted as the *average* number of bits required to describe the random variable, or as the *average uncertainty* in the random variable. It is measured in *bits*, if the logarithm is base 2, or *nats,* if it is base e (in the subsequent discussion, we omit the subscript of the log, since the units are not important in our context).

The joint entropy is a measure of the average number of bits required to represent several random variables, and the conditional entropy is the entropy of a random variable given one or more random variables (for instance, the average number of bits needed to represent $\hat{\mathsf{x}}$ if $\hat{\mathsf{y}}$ is known). Thus, given a discrete random variable $\hat{\mathsf{x}}$ with possible values $i \in \{1, \ldots, n\} := \mathcal{X}$ and probabilities $\mathsf{P}_{\hat{\mathsf{x}}}(i)$, and a discrete random variable $\hat{\mathsf{y}}$ with possible values $j \in \{1, \ldots, m\} := \mathcal{Y}$ and probabilities $\mathsf{P}_{\hat{\mathsf{y}}}(j)$, the following definitions apply.

Definition A.8.2 (Joint entropy for two random variables).

$$H\left(\hat{\mathsf{x}}, \hat{\mathsf{y}}\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) \log \left(\mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)\right).$$

Definition A.8.3 (Conditional entropy for two random variables).

$$H\left(\hat{\mathsf{x}}|\hat{\mathsf{y}}\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \mathsf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) \log \left(\mathsf{P}_{\hat{\mathsf{x}}}(i|\hat{\mathsf{y}} = j)\right),$$

It is important to note that the mean in the conditional entropy is taken using the joint distribution, and not the conditional distribution.

A simple result of the above definitions is

$$H(\hat{\mathsf{x}}, \hat{\mathsf{y}}) = H(\hat{\mathsf{x}}) + H(\hat{\mathsf{y}}|\hat{\mathsf{x}}) = H(\hat{\mathsf{y}}) + H(\hat{\mathsf{x}}|\hat{\mathsf{y}}).$$

Next, we define the *relative entropy*, or the *Kullback Leibler divergence.*

Definition A.8.4 (Kullback Leibler divergence). Given two probability measures $\mathsf{P}_{\hat{\mathsf{x}}}, \mathsf{Q}_{\hat{\mathsf{x}}}$ defined on a finite set $\mathcal{X}$,

$$D_{\mathrm{KL}}\left(\mathsf{P}_{\hat{\mathsf{x}}} || \mathsf{Q}_{\hat{\mathsf{x}}}\right) = \sum_{i \in \mathcal{X}} \mathsf{P}_{\hat{\mathsf{x}}}(i) \log \left(\frac{\mathsf{P}_{\hat{\mathsf{x}}}(i)}{\mathsf{Q}_{\hat{\mathsf{x}}}(i)}\right).$$

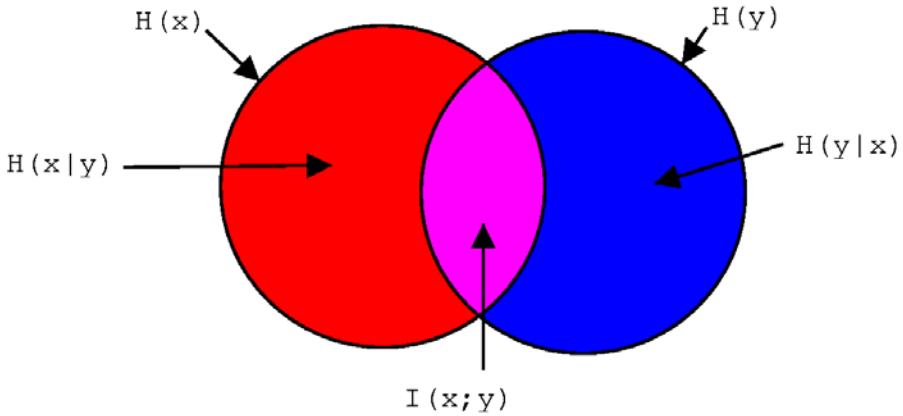

Figure A.8.1: Relations between entropies and related quantities for random variables x, y. $H(\mathsf{x})$ is represented by the entire area of the circle indicated by the relevant arrow; $H(\mathsf{x}|\mathsf{y})$ is the area shaded in red.

This has a number of interpretations. First, it is the expected log likelihood ratio for data generated according to the distribution $\mathbf{P}_{\hat{\mathsf{x}}}$, given a hypothesis $\mathbf{Q}_{\hat{\mathsf{x}}}$. It also describes the average number of *extra* bits that must be sent when a set of code words designed for the distribution $\mathbf{Q}_{\hat{\mathsf{x}}}$ is used to transmit data generated by the distribution $\mathbf{P}_{\hat{\mathsf{x}}}$, as compared with the average number of bits needed when the code words are designed for $\mathbf{P}_{\hat{\mathsf{x}}}$, so as to achieve an almost zero probability of error. A crucial property of the KL divergence is given in the following theorem.

Theorem A.8.5 (Positivity of the KL divergence). $D_{\mathrm{KL}}\left(\mathbf{P}_{\hat{\mathsf{x}}}||\mathbf{Q}_{\hat{\mathsf{x}}}\right) \geq 0$, *and* $D_{\mathrm{KL}}\left(\mathbf{P}_{\hat{\mathsf{x}}}||\mathbf{Q}_{\hat{\mathsf{x}}}\right) = 0$ *if and only if* $\mathbf{P}_{\hat{\mathsf{x}}} = \mathbf{Q}_{\hat{\mathsf{x}}}$.

Definition A.8.6 (Mutual information for two random variables). The *mutual information* takes the following, equivalent forms:

$$\begin{aligned} I\left(\hat{\mathsf{x}};\hat{\mathsf{y}}\right) &:= H(\hat{\mathsf{x}}) - H(\hat{\mathsf{x}}|\hat{\mathsf{y}}) \\ &= H(\hat{\mathsf{y}}) - H(\hat{\mathsf{y}}|\hat{\mathsf{x}}) \\ &= H(\hat{\mathsf{x}}) + H(\hat{\mathsf{y}}) - H(\hat{\mathsf{x}},\hat{\mathsf{y}}) \\ &= D_{\mathrm{KL}}\left(\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}||\mathbf{P}_{\hat{\mathsf{x}}}\mathbf{P}_{\hat{\mathsf{y}}}\right) \\ &= \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} \mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)\log\left(\frac{\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)}{\mathbf{P}_{\hat{\mathsf{x}}}(i)\mathbf{P}_{\hat{\mathsf{y}}}(j)}\right) \end{aligned}$$

The mutual information represents the *reduction* in the average number of bits needed to represent $\hat{\mathsf{x}}$ after being given $\hat{\mathsf{y}}$ (that is, the saving we make when representing $\hat{\mathsf{x}}$ given knowledge of $\hat{\mathsf{y}}$, relative to when we do not know $\hat{\mathsf{y}}$). In other words, this can be interpreted as the average number of bits needed to transmit the information that $\hat{\mathsf{x}}$ and $\hat{\mathsf{y}}$ have in common. Alternatively, it represents (via the KL divergence) the penalty incurred in assuming that $\hat{\mathsf{x}}$ and $\hat{\mathsf{y}}$ are independent. A summary of these various relations is given in Figure A.8.1.

A.8.2 Information theory in continuous spaces

Basic definitions

We now present quantities analogous to those defined above in the case of discrete spaces, but for continuous spaces. We begin by defining the differential entropy.

Definition A.8.7 (Differential entropy). Given a continuous random variable x defined on a support set $\mathcal{X} \subset \mathbb{R}$, with density $\mathbf{f}_{\mathsf{x}}$, then the differential entropy is defined as

$$h(\mathbf{f}_{\mathsf{x}}) = h(\mathsf{x}) := -\int_{\mathcal{X}} \mathbf{f}_{\mathsf{x}}(x) \log \mathbf{f}_{\mathsf{x}}(x) dx.$$

This is distinguished from the entropy by the lower case "h".

One interpretation of the above is that it represents the logarithm of the equivalent side length of the smallest set that contains most of the probability; for further detail, see [25, Chapter 9]. We now describe the link between differential entropy and the entropy definition for a discrete random variable (Definition A.8.1).

Theorem A.8.8 (Entropy and differential entropy). *We discretise the support $\mathcal{X} \subset \mathbb{R}$ of the continuous random variable x using n intervals of size Δ. Let $\hat{\mathsf{x}}$ be a discrete random variable, taking value $i \in \{1, \ldots, n\}$ with probability $\mathbf{P}_{\hat{\mathsf{x}}}(i)$ when $\mathsf{x} \in [i\Delta, (i+1)\Delta)$. Then the mean value theorem requires that there exist some $x_i \in [i\Delta, (i+1)\Delta)$ such that*

$$\mathbf{P}_{\hat{\mathsf{x}}}(i) = \int_{i\Delta}^{(i+1)\Delta} \mathbf{f}_{\mathsf{x}}(x) dx = \mathbf{f}_{\mathsf{x}}(x_i)\Delta.$$

Consequently, it can be shown that

$$\lim_{\Delta \to 0} \left(H(\hat{\mathsf{x}}) + \log \Delta\right) = h(\mathsf{x}).$$

By analogy with the joint and conditional entropies in the previous section, we define similar differential quantities below.

Definition A.8.9 (Joint differential entropy). Given the random vector $\mathbf{x} = [\mathsf{x}_1, \ldots, \mathsf{x}_n]^\top$ with density $\mathbf{f}_{\mathbf{x}}$ defined on $\mathcal{X} \subset \mathbb{R}^n$,

$$h(\mathbf{x}) = -\int_{\mathcal{X}} \mathbf{f}_{\mathbf{x}}(\mathbf{x}) \log \left(\mathbf{f}_{\mathbf{x}}(\mathbf{x})\right) d\mathbf{x}.$$

Definition A.8.10 (Conditional differential entropy). If random vectors $\mathbf{x}, \mathbf{y}$, defined respectively on $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$, have joint density $\mathbf{f}_{\mathbf{x},\mathbf{y}}$, then

$$h(\mathbf{x}|\mathbf{y}) = -\int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{f}_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) \log \left(\mathbf{f}_{\mathbf{x}}(\mathbf{x}|\mathbf{y})\right) d\mathbf{x} d\mathbf{y}.$$

Finally, the KL divergence between two continuous random variables, and the mutual information, are defined in a manner consistent with the previous section.

Definition A.8.11 (KL divergence in terms of densities). Given two densities $\mathbf{f}_{\mathbf{x}}$, $\mathbf{g}_{\mathbf{x}}$ defined on $\mathcal{X} \subset \mathbb{R}^n$, then

$$D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x}} || \mathbf{g}_{\mathbf{x}}) = \int_{\mathcal{X}} \mathbf{f}_{\mathbf{x}}(\mathbf{x}) \log \left(\frac{\mathbf{f}_{\mathbf{x}}(\mathbf{x})}{\mathbf{g}_{\mathbf{x}}(\mathbf{x})}\right) d\mathbf{x}.$$

The following theorem gives some useful properties of the KL divergence.

Theorem A.8.12 (Properties of the KL divergence on continuous spaces). *The KL divergence between densities $\mathbf{f}_{\mathbf{x}}$, $\mathbf{g}_{\mathbf{x}}$ has the properties:*

- $D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x}} || \mathbf{g}_{\mathbf{x}}) \geq 0$ *with equality if and only if* $\mathbf{f}_{\mathbf{x}} = \mathbf{g}_{\mathbf{x}}$ *almost everywhere,*
- $D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x}} || \mathbf{g}_{\mathbf{x}})$ *is invariant with respect to the invertible transform $f(\mathbf{x})$ of $\mathbf{x}$; i.e.* $D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x}} || \mathbf{g}_{\mathbf{x}}) = D_{\mathrm{KL}}(\mathbf{f}_{f(\mathbf{x})} || \mathbf{g}_{f(\mathbf{x})}) = D_{\mathrm{KL}}(\mathbf{f}_{f^{-1}(\mathbf{x})} || \mathbf{g}_{f^{-1}(\mathbf{x})})$,

- $D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x}}||\mathbf{g}_{\mathbf{x}})$ *is invariant with respect to random permutation of the components of* $\mathbf{x}$

Definition A.8.13 (Mutual information between continuous random vectors). Given two random vectors $\mathbf{x}, \mathbf{y}$, defined respectively on $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$, and with joint density $\mathbf{f}_{\mathbf{x},\mathbf{y}}$, the mutual information is

$$\begin{aligned} I(\mathbf{x},\mathbf{y}) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbf{f}_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) \log\left(\frac{\mathbf{f}_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{\mathbf{f}_{\mathbf{x}}(\mathbf{x})\mathbf{f}_{\mathbf{y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \\ &= h(\mathbf{x}) + h(\mathbf{y}) - h(\mathbf{x},\mathbf{y}) \\ &= h(\mathbf{x}) - h(\mathbf{x}|\mathbf{y}) \\ &= h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) \\ &= D_{\mathrm{KL}}(\mathbf{f}_{\mathbf{x},\mathbf{y}}||\mathbf{f}_{\mathbf{x}}\mathbf{f}_{\mathbf{y}}). \end{aligned}$$

In this case, convention dictates that we not use a lowercase "I". Relations with the differential entropies and KL divergence are analogous to the discrete case. A nice consequence of the above definition is that the mutual information between continuous random variables is defined as the limit of the discrete mutual information as the discretisation parameter Δ approaches zero, without the need for the $\log\Delta$ term. Thus, if $\hat{\mathrm{x}}$ and $\hat{\mathrm{y}}$ are discretised random variables in the sense of Theorem A.8.8, then

$$\lim_{\Delta\to 0}\left(I(\hat{\mathrm{x}},\hat{\mathrm{y}})\right) = I(\mathrm{x},\mathrm{y}).$$

The Gaussian case

The joint differential entropy of the Gaussian distribution is given by the following theorem.

Theorem A.8.14 (Differential entropy of Gaussian random variables). *Given a random vector* $\mathbf{x} = [\mathrm{x}_1,\ldots,\mathrm{x}_n]^\top$ *with Gaussian probability density,*

$$\mathbf{f}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

then the joint differential entropy is

$$h(\mathbf{x}) = \frac{1}{2}\log(2\pi e)^n |\mathbf{C}|.$$

This is a useful result, since the Gaussian distribution can easily be shown to have the maximum differential entropy of *any* distribution with the same covariance matrix, as stated in the next theorem.

Theorem A.8.15 (Gaussian random variables have highest differential entropy). *If the random vector* $\mathbf{x} = [\mathrm{x}_1,\ldots,\mathrm{x}_n]^\top$ *has zero mean and covariance matrix* $\mathbf{C}$*, then*

$$h(\mathbf{x}) \le \frac{1}{2}\log(2\pi e)^n |\mathbf{C}|,$$

with equality if and only if $\mathbf{f}_{\mathbf{x}}(\mathbf{x})$ *is a Gaussian density.*

On the basis of the above theorem, we may use the Gaussian differential entropy to upper bound the entropy of *any* discrete random variable. This involves finding a continuous (piecewise constant) random variable with the same entropy as the discrete variable, and using Theorem A.8.15 to bound this continuous variable.

Theorem A.8.16 (Upper bound on the entropy using a Gaussian approximation). *Consider the discrete random variable* $\hat{\mathrm{x}}$ *, with possible values* $i \in \{1,\ldots,n\}$ *and probabilities* $\mathbf{P}_{\hat{\mathrm{x}}}(i)$*. Then*

$$H(\hat{\mathrm{x}}) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{n}\mathbf{P}_{\hat{\mathrm{x}}}(i)i^2 - \left(\sum_{i=1}^{n}\mathbf{P}_{\hat{\mathrm{x}}}(i)i\right)^2 + \frac{1}{12}\right).$$

Finally, we investigate the effect on entropy of applying invertible transforms to random variables. This is a classic result; see for instance [49].

Theorem A.8.17 (Effect of an invertible transform on entropy). *Consider two random vectors* $\mathbf{x}, \mathbf{y}$*, related by an invertible transform* $\mathbf{y} = f(\mathbf{x})$*, and let* Jf *be the Jacobian of* f*. Then*

$$h(\mathbf{y}) = h(\mathbf{x}) + \mathsf{E}_{\mathbf{x}}\left(\log\left|\det\left(Jf(\mathbf{x})\right)\right|\right).$$

Proof. The densities $\mathbf{f}_{\mathbf{y}}(\mathbf{y})$ and $\mathbf{f}_{\mathbf{x}}(\mathbf{x})$ are related by

$$\mathbf{f}_{\mathbf{y}}(\mathbf{y}) = \frac{\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)}{\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|}.$$

Making this replacement in the definition of differential entropy, we find

$$\begin{aligned}
h(\mathbf{y}) &= -\int \mathbf{f}_{\mathbf{y}}(\mathbf{y}) \log\left(\mathbf{f}_{\mathbf{y}}(\mathbf{y})\right) d\mathbf{y} \\
&= -\int \frac{\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)}{\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|} \log\left(\frac{\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)}{\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|}\right) d\mathbf{y} \\
&= -\int \frac{\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)}{\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|} \log\left(\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)\right) d\mathbf{y} \\
&\quad + \int \frac{\mathbf{f}_{\mathbf{x}}\left(f^{-1}(\mathbf{y})\right)}{\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|} \log\left(\left|\det\left(Jf\left(f^{-1}(\mathbf{y})\right)\right)\right|\right) d\mathbf{y}
\end{aligned}$$

We then make the substitution

$$f^{-1}(\mathbf{y}) = \mathbf{x} \quad \text{and} \quad d\mathbf{x} = d\mathbf{y}\left|\det Jf(\mathbf{x})\right|^{-1}$$

to complete the proof. □

A.9 Cumulants, characteristic functions, and the Gram-Charlier expansion

In this section, we introduce cumulants, Hermite polynomials, and the Gram-Charlier expansion, using material taken from [42, 49]. We first define the characteristic function of a random vector.

Definition A.9.1 (Characteristic function). The characteristic function of a random vector $\mathbf{x} \in \mathbb{R}^n$ is

$$\varphi(\boldsymbol{\omega}) = \mathsf{E}_{\mathbf{x}}\left(\exp\left(\imath\boldsymbol{\omega}^{\top}\mathbf{x}\right)\right),$$

where $\boldsymbol{\omega} \in \mathbb{R}^n$.

The *cumulant generating function* is the logarithm of the characteristic function. In the case where x is univariate, this has the Taylor expansion

$$\phi(\omega) = \ln\left(\varphi(\omega)\right) = \sum_{k=0}^{\infty} \kappa_k \frac{(\imath\omega)^k}{k!},$$

where κ_k are the cumulants. The first four cumulants are given below as functions of the relevant moments;

$$\begin{aligned}
\kappa_1 &= \mathsf{E}_{\mathsf{x}}(\mathsf{x}), \\
\kappa_2 &= \mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right) - \left(\mathsf{E}_{\mathsf{x}}(\mathsf{x})\right)^2, \\
\kappa_3 &= \mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^3\right) - 3\mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right)\mathsf{E}_{\mathsf{x}}(\mathsf{x}) + 2\left(\mathsf{E}_{\mathsf{x}}(\mathsf{x})\right)^3, \\
\kappa_4 &= \mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^4\right) - 4\mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^3\right)\mathsf{E}_{\mathsf{x}}(\mathsf{x}) - 3\left(\mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right)\right)^2 + 12\mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right)\left(\mathsf{E}_{\mathsf{x}}(\mathsf{x})\right)^2 - 6\left(\mathsf{E}_{\mathsf{x}}(\mathsf{x})\right)^4.
\end{aligned}$$

The final expression is known as the *kurtosis*. In the zero mean case, this becomes

$$\kappa_4 = \mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^4\right) - 3\left(\mathsf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right)\right)^2.$$

A well known application of the kurtosis is in measuring non-Gaussianity, since the kurtosis is zero for Gaussian distributions[6]. More generally, the kurtosis is a measure of the "flatness" of a distribution: when the kurtosis is negative, the distribution is called *sub-Gaussian*, and tends to be less peaked and possess shorter tails than the Gaussian (for instance, the uniform distribution); when the kurtosis is positive, the distribution is called *super-Gaussian*, and is more peaked with long tails (e.g. the Laplacian distribution). We next define the Hermite polynomials.

Definition A.9.2 (The Hermite polynomials). The Hermite polynomials $H_k(x)$ are defined using the recursion

$$H_{k+1}(x) = xH_k(x) - kH_{k-1}(x).$$

These are biorthogonal with the mth order derivatives of the Gaussian distribution.

We now use these definitions to specify approximations of arbitrary densities, by expanding around the Gaussian density. We consider here the Edgeworth and Gram-Charlier expansions, which were respectively proposed in [24] and [4] for use in computing ICA contrast functions.

Definition A.9.3 (Edgeworth expansion around the Gaussian density). Let $\mathbf{f}_{\mathsf{x}}(x)$ be a probability density with zero mean and unit variance, and $\mathbf{g}_{\mathsf{x}}(x)$ a Gaussian random variable with the same mean and variance. Then we may expand $\mathbf{f}_{\mathsf{x}}(x)$ as

$$\mathbf{f}_{\mathsf{x}}(x) = \mathbf{g}_{\mathsf{x}}(x)\left(1 + \frac{\kappa_3}{3!}H_3(x) + \frac{\kappa_4}{4!}H_4(x) + \frac{10\kappa_3^2}{6!}H_6(x) + \ldots\right)$$

Definition A.9.4 (Gram-Charlier expansion around the Gaussian density). Let $\mathbf{f}_{\mathsf{x}}(x)$ be a probability density with zero mean and unit variance, and $\mathbf{g}_{\mathsf{x}}(x)$ a Gaussian random variable with the same mean and variance. Then we may expand $\mathbf{f}_{\mathsf{x}}(x)$ as

$$\mathbf{f}_{\mathsf{x}}(x) = \mathbf{g}_{\mathsf{x}}(x)\left(1 + \sum_{k=3}^{\infty} c_k H_k(x)\right),$$

where the coefficients c_k are functions of the cumulants: the first four non-zero coefficients are

$$\begin{array}{rclcrcl} c_3 & = & \frac{\kappa_3}{6}, & & c_4 & = & \frac{\kappa_4}{24}, \\ c_5 & = & \frac{\kappa_5}{120}, & & c_6 & = & \frac{1}{720}\left(\kappa_6 + 10\kappa_3^2\right). \end{array}$$

The Gram-Charlier approximation is obtained by computing the power series expansion of the cumulant generating function, and then taking an inverse Fourier transform. When $\mathbf{f}_{\mathsf{x}}(x)$ is close to Gaussian, only the low order terms in both expansions need be retained. Note, however, that care must be taken when truncating the Gram-Charlier expansion, so as to discard terms of similar magnitude, whereas the terms in the Edgeworth expansion decrease uniformly.

[6] There exist non-Gaussian distributions with zero kurtosis, however, so this is simply a heuristic.

Bibliography

[1] S. Achard, D.-T. Pham, and C. Jutten. Blind source separation in post-nonlinear mixtures. In *3rd International Conference on ICA and BSS*, 2001.

[2] Alijah Ahmed. *Signal Separation.* PhD thesis, Department of Engineering, University of Cambridge, 2000.

[3] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[4] S.-I. Amari, A. Cichoki, and Yang H. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press, 1996.

[5] F. Bach and M. Jordan. Kernel independent component analysis - (matlab code, version 1.1). `http://www.cs.berkeley.edu/~fbach/kernel-ica/index.htm`

[6] F. Bach and M. Jordan. Kernel independent component analysis. Technical Report UCB/CSD-01-1166, University of California Berkeley, 2001.

[7] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[8] F. Bach and M. Jordan. Tree-dependent component analysis. In *Uncertainty in Artificial Intelligence*, volume 18, 2002.

[9] C. Barker. Mutual information for gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970.

[10] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[11] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[12] A. Belouchrani and M. G. Amin. Blind source separation based on time-frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, 1998.

[13] M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations, 1997.

[14] R. N. Bracewell. *The Fourier Transform and its Applications.* McGraw Hill, New York, 1986.

[15] L. Breiman and Friedman J. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

[16] Mike Brookes. The matrix reference manual. `http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html`

[17] V. D. Calhoun, T. Adali, V. B. McGinty, J. J. Peckar, T. D. Watson, and G. D. Pearlson. fmri activation in a visual-perception task: Network of areas detected using the general linear model and independent components analysis. *Neuroimage*, 14(5):1080–1088, 2001.

[18] J.-F. Cardoso. Blind separation of real signals with jade (matlab code, version 1.5). `ftp://tsi.enst.fr/pub/jfc/Algo/Jade/jadeR.m`

[19] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

[20] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90(8):2009–2026, 1998.

[21] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[22] J.-F. Cardoso. Three easy routes to independent component analysis. In *International Conference on Indpendent Component Analysis and Signal Separation*, volume 3, San Diego, California, 2001. Institute for Neural Computation.

[23] A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing.* John Wiley and Sons, New York, 2002.

[24] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[25] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[26] N. Cristianini, A. Elisseeff, and J. Shawe-Taylor. On optimizing kernel alignment. Technical Report NC2-TR-2001-087, NeuroCOLT, `http://www.neurocolt.com`, 2001.

[27] N. Cristianini, J. Shawe-Taylor, and J. Kandola. Spectral kernel methods for clustering. In *NIPS*, volume 14, Cambridge, MA, 2002. MIT Press.

[28] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Stat.*, 21:2–8, 1953.

[29] D. Das and P. Sen. Restricted canonical correlations. *Linear Algebra and its Applications*, 210:29–47, 1994.

[30] M. Davy, A. Gretton, A. Doucet, and P. J. W. Rayner. Optimized support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12):442–445, 2002.

[31] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, New York, second edition, 2001.

[32] R. Durrett. *Probability: Theory and Examples.* Duxbury Press, Belmont, California, 1996.

[33] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[34] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

[35] Hugo Gävert, Jarmo Hurri, Jaakko Särelä, and Aapo Hyvärinen. FastICA (matlab code, version 2.1). `http://www.cis.hut.fi/projects/ica/fastica/`

[36] M. Girolami. An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[37] M. Greenacre. *Theory and Applications of Correspondence Analysis.* Academic Press, London, 1984.

[38] A. Gretton, M. Davy, A. Doucet, and P. J. W. Rayner. Nonstationary signal classification using support vector machines. In *IEEE Workshop on Statistical Signal Processing Proceedings*, pages 305–308. IEEE Signal Processing Society, 2001.

[39] S. Harmeling, A. Ziehe, M. Kawanabe, and K.-M. Müller. Kernel-based nonlinear blind source separation. Technical Report 1/2002, Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin Germany, 2002.

[40] D. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, New York, 1997.

[41] T. Hastie and R. Tibshirani. Independent components analysis through product density estimation. In *NIPS*, volume 15, Cambridge, MA, 2002. MIT Press.

[42] S. Haykin. *Neural Networks : A Comprehensive Foundation*. Macmillan, New York, 2nd edition, 1998.

[43] R. Herbrich. *Learning Kernel Classifers: Theory and Algorithms*. MIT press, Cambridge, MA, 2002.

[44] James Hopgood. *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments*. PhD thesis, Department of Engineering, University of Cambridge, 2001.

[45] S. Hosseni and C. Jutten. On the separability of nonlinear mixtures of temporally correlated sources. *IEEE Signal Processing Letters*, 10(2):43–46, 2003.

[46] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Proc. IEEE Neural Networks for Signal Processing Workshop*, pages 388–397, 1997.

[47] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *NIPS*, volume 10, pages 273–279, Cambridge, MA, 1998. MIT Press.

[48] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[49] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

[50] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[51] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[52] A. Hyvärinen and M. Plumbley. Optimization with orthogonality constraints: a modified gradient method. Unpublished note, 2002.

[53] C. Jutten and J. Hérault. Blind separation of sources i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[54] J. Karvanen, J. Eriksson, and V. Koivunen. Adaptive score functions for maximum likelihood ICA. *Journal of VLSI Signal Processing Systems*, 32:83–92, 2002.

[55] Malte Kuss. Kernel multivariate analysis. Master's thesis, Technical University of Berlin, 2001.

[56] P. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.

[57] T.-W. Lee, M. Girolami, A. Bell, and T. Sejnowski. A unifying framework for independent component analysis. *Computers and Mathematics with Applications*, 39:1–21, 2000.

[58] T.-W. Lee, M. Girolami, and T. Sejnowski. The extended Infomax algorithm for mixed sub-gaussian and super-gaussian sources (matlab code, version 1.0). `http://www.cnl.salk.edu/~tewon/ICA/Code/ext_ica_download.html`

[59] T.-W. Lee, M. Girolami, and T. Sejnowski. Independent component analysis using an extended Infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):409–433, 1999.

[60] D. Mackay. Maximum likelihood and covariant algorithms for independent component analysis. Technical report, Cavendish Laboratory, University of Cambridge, 1999.

[61] O. Mangasarian. *Nonlinear Programming*. SIAM, Philadelphia, 1994.

[62] Thomas Melzer, Michale Reiter, and Horst Bischof. Kernel canonical correlation analysis. Technical Report PRIP-TR-65, Pattern Recognition and Image Processing Group, TU Wien, 2001.

[63] S. Mendelson. Goemetric methods in the analysis of Glivenko-Cantelli classes. In D. Helmbold and R. Williamson, editors, *Proceedings of COLT*, pages 256–272, 2001.

[64] E. Miller and J. Fisher III. Independent components analysis by direct entropy minimization. Technical Report UCB/CSD-3-1221, Computer Science Division, University of California Berkeley, 2003.

[65] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, New York, 1991.

[66] B. Pearlmutter. Music samples to illustrate the context-sensitive generalisation of ICA. `http://www.cs.unm.edu/~bap/demos.html`

[67] B. Pearlmutter and L. Parra. A context-sensitive generalisation of ICA. In *Proc. ICONIP*, 1996.

[68] C. E. Pearson, editor. *Handbook of Applied Mathematics*. Van Nostrand Reinhold Company, New York, 1983.

[69] D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[70] D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 2002. Submitted.

[71] D.-T. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory*, 48(7):1935–1946, 2002.

[72] D.-T. Pham and J.-F. Cardoso. Blind separartion of instantaneous mixtures of non-stationary sources. *IEEE Transactions on Signal Processing*, 49(9):1837–1848, 2001.

[73] D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

[74] D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

[75] A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

[76] R. Rosipal and L. Trejo. Kernel partial least squares regression in reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 1(2):97–123, 2001.

[77] B. Schölkopf, R. Herbrich, and A. Smola. A generalised representer theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2001.

[78] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[79] B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999.

[80] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[81] G. Strang. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, New York, third edition edition, 1988.

[82] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

[83] T. van Gestel, J. Suykens, J. de Brabanter, B. de Moor, and J. Vanderwalle. Kernel canonical correlation annalysis and least squares support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*. Springer Verlag, 2001.

[84] O. Vasicek. A test for normality based on the sample entropy. *Journal of the Royal Statistical Society, Series B*, 38(1):54–59, 1976.

[85] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*, volume 15, Cambridge, MA, 2002. MIT Press.

[86] N. Vlassis and Y. Motomura. Efficient source adaptivity in independent component analysis. *IEEE Transactions on Neural Networks*, 12(3):559–566, 2001.

[87] S. Wold, H. Ruhe, H. Wold, and W. J. Dunne III. The collinarity problem in linear regression. the partial least squares (pls) approach to the generalized inverse. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.

[88] H. H. Yang and S.-I. Amari. Adaptive on-line learning algorithms for blind separation – maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997.

[89] H. H. Yang, S.-I. Amari, and A. Cichocki. Information theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–300, 1998.

[90] X.-L. Zhu and X.-D. Zhang. Adaptive RLS algorithm for blind source separation using a natural gradient. *IEEE Signal Processing Letters*, 9(12):432–435, 2002.