GP Kernels for Cross-Spectrum Analysis

Authored by:

Lawrence Carin
David E. Carlson
Kafui Dzirasa
Kyle R. Ulrich

Abstract

Multi-output Gaussian processes provide a convenient framework for multi-task problems. An illustrative and motivating example of a multi-task problem is multi-region electrophysiological time-series data, where experimentalists are interested in both power and phase coherence between channels. Recently, Wilson and Adams (2013) proposed the spectral mixture (SM) kernel to model the spectral density of a single task in a Gaussian process framework. In this paper, we develop a novel covariance kernel for multiple outputs, called the cross-spectral mixture (CSM) kernel. This new, flexible kernel represents both the power and phase relationship between multiple observation channels. We demonstrate the expressive capabilities of the CSM kernel through implementation of a Bayesian hidden Markov model, where the emission distribution is a multi-output Gaussian process with a CSM covariance kernel. Results are presented for measured multi-region electrophysiological data.

1 Paper Body

Gaussian process (GP) models have become an important component of the machine learning literature. They have provided a basis for non-linear multivariate regression and classification tasks, and have enjoyed much success in a wide variety of applications [16]. A GP places a prior distribution over latent functions, rather than model parameters. In the sense that these functions are defined for any number of sample points and sample positions, as well as any general functional form, GPs are nonparametric. The properties of the latent functions are defined by a positive definite covariance kernel that controls the covariance between the function at any two sample points. Recently, the spectral mixture (SM) kernel was proposed by Wilson and Adams [24] to model a spectral density with a scale-location mixture of Gaussians. This flexible and interpretable class of kernels is capable of recovering any composition of stationary kernels [27, 9, 13]. The SM kernel has been used for GP regression of a scalar output (i.e.,

single function, or observation ?task?), achieving impressive results in extrapolating atmospheric CO2 concentrations [24]; image inpainting [25]; and feature extraction from electrophysiological signals [21]. However, the SM kernel is not defined for multiple outputs (multiple correlated functions). Multioutput GPs intersect with the field of multi-task learning [4], where solving similar problems jointly allows for the transfer of statistical strength between problems, improving learning performance when compared to learning all tasks individually. In this paper, we consider neuroscience applications where low-frequency (¡ 200 Hz) extracellular potentials are simultaneously recorded from implanted electrodes in multiple brain regions of a mouse [6]. These signals are known as local field potentials (LFPs) and are often highly correlated between channels. Inferring and understanding that interdependence is biologically significant. 1

A multi-output GP can be thought of as a standard GP (all observations are jointly normal) where the covariance kernel is a function of both the input space and the output space (see [2] and references therein for a comprehensive review); here ?input space? means the points at which the functions are sampled (e.g., time), and the ?output space? may correspond to different brain regions. A particular positive definite form of this multi-output covariance kernel is the sum of separable (SoS) kernels, or the linear model of coregionalization (LMC) in the geostatistics literature [10], where a separable kernel is represented by the product of separate kernels for the input and output spaces. While extending the SM kernel to the multi-output setting via the LMC framework (i.e., the SMLMC kernel) provides a powerful modeling framework, the SM-LMC kernel does not intuitively represent the data. Specifically, the SM-LMC kernel encodes the cross-amplitude spectrum (square root of the cross power spectral density) between every pair of channels, but provides no crossphase information. Together, the cross-amplitude and cross-phase spectra form the cross-spectrum, defined as the Fourier transform of the cross-covariance between the pair of channels. Motivated by the desire to encode the full cross-spectra into the covariance kernel, we design a novel kernel termed the cross-spectral mixture (CSM) kernel, which provides an intuitive representation of the power and phase dependencies between multiple outputs. The need for embedding the full cross-spectrum into the covariance kernel is illustrated by a recent surge in neuroscience research discovering that LFP interdependencies between regions exhibit phase synchrony patterns that are dependent on frequency band [11, 17, 18]. The remainder of the paper is organized as follows. Section 2 provides a summary of GP regression models for vector-valued data, and Section 3 introduces the SM, SM-LMC, and novel CSM covariance kernels. In Section 4, the CSM kernel is incorporated in a Bayesian hidden Markov model (HMM) [14] with a GP emission distribution as a demonstration of its utility in hierarchical modeling. Section 5 provides details on inverting the Bayesian HMM with variational inference, as well as details on a fast, novel GP fitting process that approximates the CSM kernel by its representation in the spectral domain. Section 6 analyzes the performance of this approximation and presents results for the CSM kernel in the neuroscience application, considering measured multi-region LFP data from the brain of a mouse. We conclude in Section 7 by discussing how this

novel kernel can trivially be extended to any time-series application where GPs and the cross-spectrum are of interest.

2

Review of Multi-Output Gaussian Process Regression

A multi-output regression task estimates samples from C output channels, y n = [yn1 , . . . , ynC]T corresponding to the n-th input point xn (e.g., the n-th temporal sample). An unobserved latent function f (x) = [f1 (x), . . . , fC (x)]T is responsible for generating the observations, such that y n ? N (f (xn), H ?1), where H = diag(?1 , . . . , ?C) is the precision of additive Gaussian noise. A GP prior on the latent function is formalized by f (x) ? GP(m(x), K(x, x0)) for arbitrary input x, where the mean function m(x) ? RC is set to equal 0 without loss of generality, and 0 the covariance function (K(x, x0))c,c0 = k c,c (x, x0) = cov(fc (x), fc0 (x0)) creates dependencies 0 between observations at input points x and x , as observed on channels c and c0 . In general, the input space x could be vector valued, but for simplicity we here assume it to be scalar, consistent with our motivating neuroscience application in which x corresponds to time. A convenient representation for multi-output kernel functions is to separate the kernel into the product of a kernel for the input space and a kernel for the interactions between the outputs. This is known as a separable kernel. A sum of separable kernels (SoS) representation [2] is given by 0

k c,c (x, x0) =

Q X

bq (c, c0)kq (x, x0),

or

q=1

K(x, x0) =

Q X

B q kq (x, x0),

(1)

q=1

where kq (x, x0) is the input space kernel for component q, bq (c, c0) is the q-th output interaction kernel, and B q ? RC?C is a positive semi-definite output kernel matrix. Note that we have a discrete set of C output spaces, c ? {1, . . . , C}, where the input space x is continuous, and discretely sampled arbitrarily in experiments. The SoS formulation is also known as the linear model of coregionalization (LMC) [10] and B q is termed the coregionalization matrix. When Q = 1, the LMC reduces to the intrinsic coregionalization model (ICM) [2], and when rank(B q) is restricted to equal 1, the LMC reduces to the semiparametric latent factor model (SLFM) [19]. 2

Any finite number of latent functional evaluations f = [f1 (x), . . . , fC (x)]T at locations x = [x1 , . . . , xN]T has a multivariate normal distribution N (f ; 0, K), such that K is formed through the block partitioning ? 1,1 ? k (x, x) ? ? ? k 1,C (x, x) Q ? ? X .. .. .. = B q ? kq (x, x), (2) K=? ? . . . k C,1 (x, x) ? ? ?

k C,C (x, x)

q=1

0

where each k c,c (x, x) is an N ? N matrix and ? symbolizes the Kronecker product. A vector-valued dataset consists of observations y = vec([y 1 , . . . , y N]T) ? RCN at the respective locations x = [x1 , . . . , xN]T such that the first N elements of y are from channel 1 up to the last N elements belonging to channel C. Since both the likelihood p(y—f , x) and distribution over latent functions p(f —x) are Gaussian, the marginal likelihood is conveniently represented by Z p(y—x) = p(y—f , x)p(f —x)df = N (0, ?), ? = K + H ?1 ? I N , (3) where all possible functions f have been marginalized out. Each input-space covariance kernel is defined by a set of hyperparameters, ?. This conditioning was removed for notational simplicity, but will henceforth be included in the notation. For example, if the squared exponential kernel is used, then kSE (x, x0 ; ?) = exp(? 12 ——x ? x0 ——2 /‘2), defined by a single hyperparameter ? = {‘}. To fit a GP to the dataset, the hyperparameters are typically chosen to maximize the marginal likelihood in (3) via gradient ascent.

3

Expressive Kernels in the Spectral Domain

This section first introduces the spectral mixture (SM) kernel [24] as well as a multi-output extension of the SM kernel within the LMC framework. While the SM-LMC model is capable of representing complex spectral relationships between channels, it does not intuitively model the cross-phase spectrum between channels. We propose a novel kernel known as the cross-spectral mixture (CSM) kernel that provides both the cross-amplitude and cross-phase spectra of multi-channel observations. Detailed derivations of each of these kernels is found in the Supplemental Material. 3.1

The Spectral Mixture Kernel

A spectral Gaussian (SG) kernel is defined by an amplitude spectrum with a single Gaussian distribution reflected about the origin, 1 SSG (?; ?) = [N (?; ??, ?) + N (?; ?, ?)] , (4) 2 where ? = {?, ?} are the kernel hyperparameters, ? represents the peak frequency, and the variance ? is a scale parameter that controls the spread of the spectrum around ?. This spectrum is a function of angular frequency. The Fourier transform of (4) results in the stationary, positive definite autocovariance function 1 kSG (? ; ?) = exp(? ?? 2) cos(??), (5) 2 0 where stationarity implies dependence on input domain differences k(? ; ?) = k(x, ? x ; ?) with ? = 0 x ? x . The SG kernel may also be derived by considering a latent signal f (x) = 2 cos(?(x + ?)) with frequency uncertainty ? ? N (?, ?) and phase offset ??. The kernel is the auto-covariance function for f (x), such that kSG (? ; ?) = cov(f (x), f (x+?)). When computing the auto-covariance, the frequency ? is marginalized out, providing the kernel in (5) that includes all frequencies in the spectral domain with probability 1. A weighted, linear combination of SG kernels gives the spectral mixture (SM) kernel [24], kSM (? ; ?) =

Q X

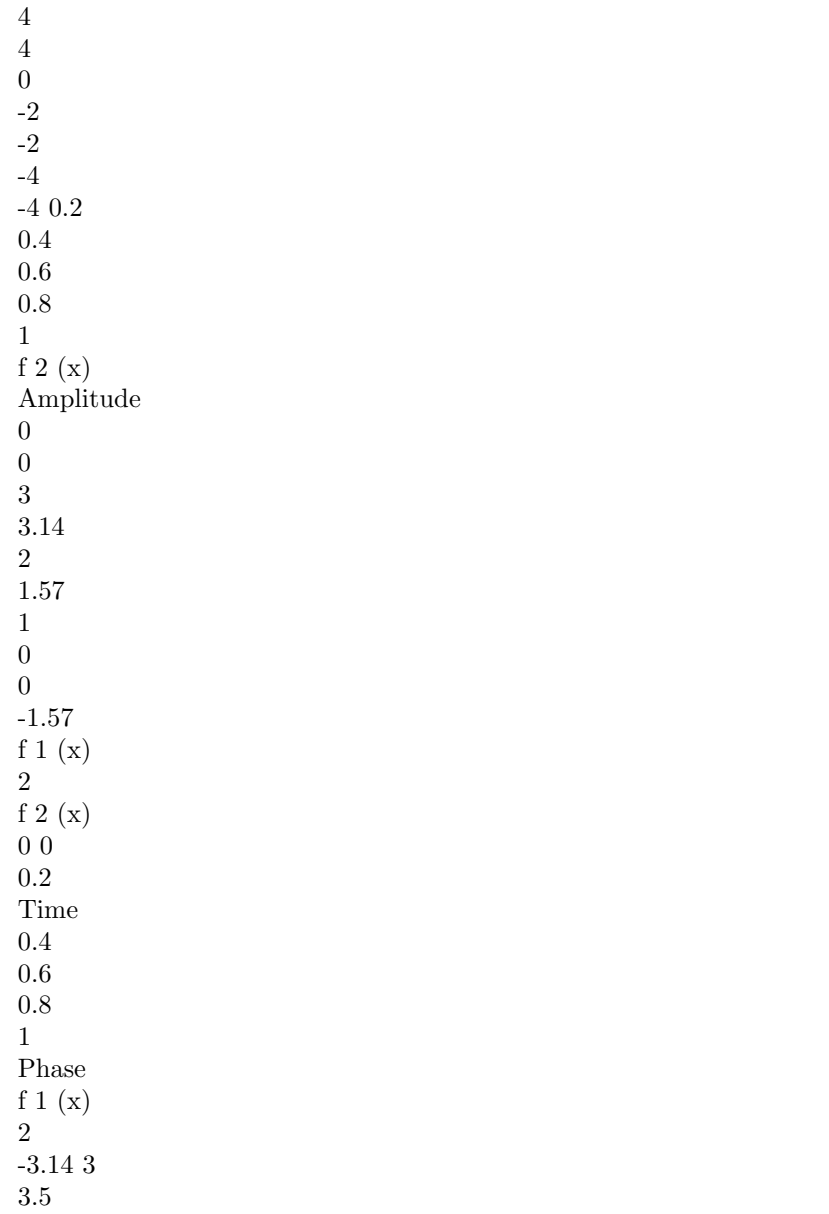
aq kSG (? ; ? q),

SSM (?; ?) =

q=1

Q X
aq SSG (?; ? q),
(6)
q=1

where ? q = {aq , ?q , ?q } and ? = {? q } has 3Q degrees of freedom. The SM kernel may be derived as the Fourier transform of the spectral density SSM (?; ?) or as the auto-covariance of latent funcPQ p tions f (x) = q=1 2aq cos(?q (x + ?q)) with uncertainty in angular frequency ?q ? N (?q , ?q). 3



4
Time
4.5
5
5.5
6
Frequency

Figure 1: Latent functions drawn for two channels f1 (x) (blue) and f2 (x) (red) using the CSM kernel (left) and rank-1 SM-LMC kernel (center). The functions are comprised of two SG components centered at 4 and 5 Hz. For the CSM kernel, we set the phase shift ?c0 ,2 = ?. Right: the cross-amplitude (purple) and cross-phase (green) spectra between f1 (x) and f2 (x) are shown for the CSM kernel (solid) and SM-LMC kernel (dashed). The ability to tune phase relationships is beneficial for kernel design and interpretation.

The moniker for the SM kernel in (6) reflects the mixture of Gaussian components that define the spectral density of the kernel. The SM kernel is able to represent any stationary covariance kernel given large enough Q; to name a few, this includes any combination of squared exponential, Mat‘ern, rational quadratic, or periodic kernels [9, 16, 24]. 3.2

The Cross-Spectral Mixture Kernel

A multi-output version of the SM kernel uses the SG kernel directly within the LMC framework: Q X K SM-LMC (? ; ?) = B q kSG (? ; ? q), (7) q=1

where Q SG kernels are shared among the outputs via the coregionalization matrices {B q }Q q=1 . A generalized, non-stationary version of this SM-LMC kernel was proposed in [23] using the Gaussian process regression network (GPRN) [26]. The marginal distribution for any single channel is simply a Gaussian process with a SM covariance kernel. While this formulation is capable of providing a full cross-amplitude spectrum between two channels, it contains no information about a crossP phase spectrum. Specifically, each channel is merely a weighted sum of q Rq latent functions where Rq = rank(B q). Whereas these functions are shared exactly across channels, our novel CSM kernel shares phase-shifted versions of these latent functions across channels. Definition 3.1. The cross-spectral mixture (CSM) kernel takes the form Rq q Q X X

1 arcq arc0 q exp(? ?q ? 2) cos ?q ? + ?rc0 q ? ?rcq , (8) 2 q=1 r=1 PQ Rq Q }q=1 has 2Q + q=1 Rq (2C ? 1) degrees of freedom, where ? = {?q , ?q , {arq , ?rq , ?r1q , 0}r=1 and arcq and ?rcq respectively represent the amplitude and shift in the input space for latent functions associated with channel c. In the LMC framework, the CSM kernel is (Q) Rq X X K CSM (? ; ?) = Re Bq e kSG (? ; ? q) , Bq = ? r (? r)? , 0

c,c kCSM (? ; ?) =

q

q=1

q

r=1

p 1 r r e kSG (? ; ? q) = exp(? ?q ? 2 + j?q ?), ?cq = arcq exp(?j?cq), 2 r where e kSG (?, ? q) is phasor notation of the SG kernel, B q is rank-Rq , {?cq

} are complex scalar r r coefficients encoding amplitude and phase, and ?cq , ?q ?cq is an alternative phase representation. ? We use complex notation where j = ?1, Re{?} returns the real component of its argument, and ? ? represents the complex conjugate of ?. Both the CSM and SM-LMC kernels force the marginal distribution of data from a single channel to be a Gaussian process with a SM covariance kernel. The CSM kernel is derived in the Supplemental Material by considering functions represented by phase-shifted sinusoidal signals, PQ PRq p r iid fc (x) = 2acq cos(?qr (x + ?rcq)), where each ?qr ? N (?q , ?q). Computing the r=1 q=1 cross-covariance function cov(fc (x), fc0 (x + ?)) provides the CSM kernel. A comparison between draws from Gaussian processes with CSM and SM-LMC kernels is shown in Figure 1. The utility of the CSM kernel is clearly illustrated by its ability to encode phase 4

information, as well as its powerful functional form of the full cross-spectrum (both amplitude 0 (?) are obtained by repand phase). The amplitude function Ac,c0 (?) and phase function ?c,cP 0 resenting the cross-spectrum in phasor notation, i.e., ?c,c0 (?; ?) = q (B q)c,c SSG (?; ? q) = Ac,c0 (?) exp(j?c,c0 (?)). Interestingly, while the CSM and SM-LMC kernels have identical marginal amplitude spectra for shared {?q , ?q , aq }, their cross-amplitude spectra differ due to the inherent destructive interference of the CSM kernel (see Figure 1, right).

4

Multi-Channel HMM Analysis

Neuroscientists are interested in examining how the network structure of the brain changes as animals undergo a task, or various levels of arousal [15]. The LFP signal is a modality that allows researchers to explore this network structure. In the model provided in this section, we cluster segments of the LFP signal into discrete ?brain states? [21]. Each brain state is represented by a unique cross-spectrum provided by the CSM kernel. The use of the full cross-spectrum to define brain states is supported by previous work discovering that 1) the power spectral density of LFP signals indicate various levels of arousal states in mice [7, 21], and 2) frequency-dependent phase synchrony patterns change as animals undergo different conditions in a task [11, 17, 18] (see Figure 2). The vector-valued observations from C channels are segmented into W contiguous, non-overlapping windows. The windows are common across channels, such that the C-channel data for window w w T w w ? {1, . . . , W } are represented by y w n = [yn1 , . . . , ynC] at sample location xn . Given data, each window consists of Nw temporal samples, but the model is defined for any set of sample locations. We model the observations {y w n } as emissions from a hidden Markov model (HMM) with L hidden, discrete states. State assignments are represented by latent variables ?w ? {1, . . . , L} for each window w ? {1, . . . , W }. In general, L is a set upper bound of the number of states (brain states [21], or ?clusters?), but the model can shrink down and infer the number of states needed to fit the data. This is achieved by defining the dynamics of the latent states according to a Bayesian HMM [14]: ?1 ? Categorical(?0),

?w ? Categorical(??w?1) ?w ? 2,

?0 , ¿ ? Dirichlet(?),

where the initial state assignment is drawn from a categorical distribution with probability vector ?0 and all subsequent states assignments are drawn from the transition vector ??w?1 . Here, ¿h is the probability of transitioning from state ‘ to state h. The vectors {?0 , ?1 , . . . , ?L } are independently drawn from symmetric Dirichlet distributions centered around ? = [1/L, . . . , 1/L] to impose sparsity on transition probabilities. In effect, this allows the model to learn the number of states needed for the data (i.e., fewer than L) [3]. Each cluster ‘ ? {1, . . . , L} is assigned GP parameters ? ‘ . The latent cluster assignment ?w for window w indicates which set of GP parameters control the emission distribution of the HMM: ?1 w yw n ? N (f w (xn), H ?w),

f w (x) ? GP(0, K(x, x0 ; ? ?w)),

(9)

0

c,c where (K(x, x0 ; ? ‘))c,c0 = kCSM (x, x0 ; ? ‘) is the CSM kernel, and the cluster-dependent precision H ?w = diag(? ?w) generates independent Gaussian observation noise. In this way, each window w is modeled as a stochastic process with a multi-channel cross-spectrum defined by ? ?w . Raw LFP Data

Cross-Amplitude Spectrum

Cross-Phase Spectrum 1

BLA IL Cortex

DELTA Waves

THETA Waves ALPHA Waves

BETA Waves

Lag (rad)

Potential

Amplitude

0.5 0 -0.5 -1 -1.5

0.1

0.2

0.3

0.4

Time (sec)

0.5

0.6

0.7

0.8

0

2

4

6

8

10

Frequency (Hz)

12

14

16
0
2
4
6
8
10
12
14
16
Frequency (Hz)

Figure 2: A short segment of LFP data recorded from the basolateral amygdala and infralimbic cortex is shown on the left. The cross-amplitude and phase spectra are produced using Welch?s averaged periodogram method [22] for several consecutive 5 second windows of LFP data. Frequency dependent phase synchrony lags are consistently present in the cross-phase spectrum, motivating the CSM kernel. This frequency dependency aligns with preconcieved notions of bands, or brain waves (e.g., 8-12 Hz alpha waves).

5
5
Inference

w T A convenient notation vectorizes all observations within a window, y w = vec([y w 1 , . . . , y Nw]), w where vec(A) is the vectorization of matrix A; i.e., the first Nw elements of y are observations from channel 1, up to the last Nw elements of y w belonging to channel C. Because samples are obtained on an evenly spaced temporal grid, we fix Nw = N and align relative sample locations within a window to an oracle xw = x = [x1 , . . . , xN]T for all w.

The model in Section 4 generates the set of observations Y = {y w }W w=1 at aligned sample locations W x given kernel hyperparameters ? = {? ‘ , ? ‘ }L and model variables ? = {{¿ }L ‘=1 ‘=0 , {?w }w=1 }. The latent variables ? are inverted using mean-field variational inference [3], obtaining an approxiQL mate posterior distribution q(?) = q(?1:W) ‘=0 Dir(¿ ; ¿). The approximate posterior is chosen to minimize the KL divergence to the true posterior distribution p(?—Y , ?, x) using the standard variational EM method detailed in Chapter 3 of [3]. During each iteration of the variational EM algorithm, the kernel hyperparameters ? are chosen to PW PL maximize the expected marginal log-likelihood Q = w=1 ‘=1 q(?w = ‘) log N (y w ; 0, ¿)via gradient ascent, where q(?w = ‘) is the marginal posterior probability that window w is assigned e ‘ } is the CSM kernel matrix for state ‘ with the complex form to brain state ‘, and ¿ = Re{? P ‘ e e ¿ = q B q ? kSG (x, x; ? ‘) + H ?1 ‘ ? I N . Performing gradient ascent requires the derivatives P ?1 w ?1 ?¿ ?Q 1 T w,‘ tr((¿w ¿w ? ¿) ??j) where ¿w = ¿ y [16]. A na??ve implementation of ??j = 2 this gradient requires the inversion of ¿ , which has complexity O(N 3 C 3) and storage requirements O(N 2 C 2) since a simple method to invert a sum of Kronecker products does not exist. A common trick for GPs with evenly spaced samples (e.g., a temporal grid) is to use the discrete Fourier transform (DFT)

to approximate the inverse of ¿ by viewing this as an approximately circulant matrix [5, 12]. These methods can speed up inference because circulant matrices are diagonalizable by the DFT coefficient matrix. Adjusting these methods to the multi-output formulation, we show how the DFT of the marginal covariance matrices retains the cross-spectrum information. Proposition 5.1. Let y w ? N (0, ??w) represent the marginal likelihood of circularly-symmetric [8] real-valued observations in window w, and denote the concatenation of the DFT of each channel as z w = (I C ? U)? y w where U is the N ? N unitary DFT matrix. Then, z w is shown in the Supplemental Material to have the complex normal distribution [8]: z w ? CN (0, 2S ?w),

S ‘ = ? ?1

Q X

B ‘q ? W ‘q + H ?1 ‘ ? IN ,

(10)

q=1

where ? = xi+1 ? xi for all i = 2, . . . , N , and W ‘q ? diag([SSG (?; ? ‘q), 0]) is approximately diagonal. The spectral density SSG (?; ?) = [SSG (?1 ; ?), . . . , SSG (?b N +1 c ; ?)] is found via (4) at 2

N

2? angular frequencies ? = N , and 0 = [0, . . . , 0] is a row vector of N 2?1 zeros. ? 0, 1, . . . , 2 The hyperparameters of the CSM kernels ? may now be optimized from the expected marginal log-likelihood of Z = {z w }W w=1 instead of Y . Conceptually, the only difference during the fitting process is that, with the latter, derivatives of the covariance kernel are used, while, with the former, derivatives of the power spectral density are used. Computationally, this method improves the na??ve O(N 3 C 3) complexity of fitting the standard CSM kernel to O(N C 3) complexity. Memory requirements are also reduced from O(N 2 C 2) to O(N C 2). The reason for this improvement is that S ‘ is now represented as N independent C ? C blocks, reducing the inversion of S ‘ to inverting a permuted block-diagonal matrix.

6

Experiments

Section 6.1 demonstrates the performance of the CSM kernel and the accuracy of the DFT approximation In Section 6.2, the DFT approximation for the CSM kernel is used in a Bayesian HMM framework to cluster time-varying multi-channel LFP data based on the full cross-spectrum; the HMM states here correspond to states of the brain during LFP recording. 6

Table 1: The mean and standard deviation of the difference between the AIC value of a given model and the AIC value of the rank-2 CSM model. Lower values are better.

0.035

KL Divergence

0.03 0.025

7 = 0.5 Hz 7 = 1 Hz 7 = 3 Hz

0.02 0.015 0.01 0.005 0 0

1

2
3
4
5
6
7
8
Series Length (seconds)

Figure 3: Time-series data is drawn from a Gaussian process with a known CSM covariance kernel, where the domain restricted to a fixed number of seconds. A Gaussian process is then fitted to this data using the DFT approximation. The KL-divergence of the fitted marginal likelihood from the true marginal likelihood is shown.

6.1

Rank

Model

? AIC

1 1 1 2 2 2 3 3 3

SE-LMC SM-LMC CSM SE-LMC SM-LMC CSM SE-LMC SM-LMC CSM

4770 (993) 512 (190) 109 (110) 5180 (1120) 325 (167) 0 (0) 5550 (1240) 412 (184) 204 (71.7)

Performance and Inference Analysis

The performance of the CSM kernel is compared to the SM-LMC kernel and SE-LMC (squared exponential) kernel. Each of these models allow Q=20, and the rank of the coregionalization matrices is varied from rank-1 to rank-3. For a given rank, the CSM kernel always obtains the largest marginal likelihood for a window of LFP data, and the marginal likelihood always increases for increasing rank. To penalize the number of kernel parameters (e.g., a rank-3, Q=20 CSM kernel for 7 channels has 827 free parameters to optimize), the Akaike information criterion (AIC) is used for model selection [1]. For this reason, we do not test rank greater than 3. Table 1 shows that a rank-2 CSM kernel is selected using this criterion, followed by a rank-1 CSM kernel. To show the rank-2 CSM kernel is consistently selected as the preferred model we report means and standard deviations of AIC value differences across 30 different randomly selected 3-second windows of LFP data. Next, we provide numerical results for the conditions required when using the DFT approximation in (10). This allows for one to define details of a particular application in order to determine if the DFT approximation to the CSM kernel is appropriate. A CSM kernel is defined for two outputs with a single Gaussian component, Q = 1. The mean frequency and variance for this component are set to push the limits of the application. For example, with LFP data, low frequency content is of interest, namely greater than 1 Hz; therefore, we test values of ? e1 ? { 12 , 1, 3} Hz. We anticipate 2 variances at these frequencies to be around ?e1 = 1 Hz . A conversion to angular frequency gives ?1 = 2?e ?1 and ?1 = 4? 2 ?e1 . The covariance matrix ? in (3) is formed using these parameters, a fixed noise variance, and N observations on a time grid with sampling rate of 200

Hz. Data y are drawn from the marginal likelihood with covariance ?. ? The KL A new CSM kernel is fit to y using the DFT approximation, providing an estimate ?. divergence of the fitted marginal likelihood from the true marginal likelihood is " # 1 —?— ?1 ? ? , KL(p(y—?)——p(y—?)) = log ? N + tr(? ?) ? 2 —?— where — ? — and tr(?) are the determinant and trace operators, respectively. Computing 1 ? KL(p(y— ?)——p(y—?)) for various values of ? e and N provides the results in Figure 3. This plot 1 N shows that the DFT approximation struggles to resolve low frequency components unless the series length is sufficiently long. Due to the approximation error, when using the DFT approximation on LFP data we a priori filter out frequencies below 1.5 Hz and perform analyses with a series length of 3 seconds. This ensures the DFT approximation represents the true covariance matrix. The following application of the CSM kernel uses these settings. 6.2

Including the CSM Kernel in a Bayesian Hierarchical Model

We analyze 12 hours of LFP data of a mouse transitioning between different stages of sleep [7, 21]. Observations were recorded simultaneously from 4 channels [6], high-pass filtered at 1.5 Hz, and subsampled to 200 Hz. Using 3 second windows provides N = 600 and W = 14, 400. The HMM was implemented with the number of kernel components Q = 15 and the number of states L = 7. 7

1.57
2
Amplitude
0
1
Phase
3
?1.57
0
?3.14
6
Amplitude
2.5
3.14
BasalAmy
5
3.14
DLS
1.5
1
1.57
5 3
0
1
Phase
Amplitude
6

0.5
?1.57
0 0
0
15
3
5 3
0
1
Phase
1.57
2
1
Phase
?1.57 ?3.14
6
DHipp
3.14 1.57
5 3
0
1
0
?1
Phase
Amplitude
10
Frequency (Hz)
3.14
DMS
0
Amplitude
5
?3.14
6
?2
?1.57
0 0
5
10
15 0
5
Frequency
10
15 0
Frequency
5

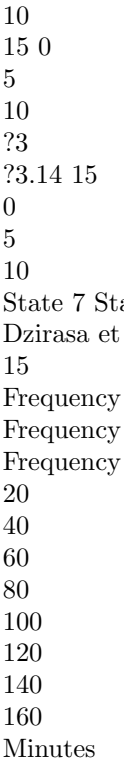


Figure 4: A subset of results from the Bayesian HMM analysis of brain states. In the upper left, the full crossspectrum for an arbitrary state (state 7) is plotted. In the upper right, the amplitude (top) and phase (bottom) functions for the cross-spectrum between the Dorsomedial Striatum (DMS) and Hippocampus (DHipp) are shown for all seven states. On the bottom, the maximum likelihood state assignments are shown and compared to the state assignments from [7]. The same colors between the CSM state assignments and the phase and amplitude functions correspond to the same state. These colors are alligned to the [7] states, but there is no explicit relationship between the colors of the two state sequences.

This was chosen because sleep staging tasks categorize as many as seven states: various levels of rapid eye movement, slow wave sleep, and wake [20]. Although rigorous model selection on L is necessary to draw scientific conclusions from the results, the purpose of this experiment is to illustrate the utility of the CSM kernel in this application. An illustrative subset of the results are shown in Figure 4. The full cross-spectrum is shown for a single state (state 7), and the cross-spectrum between the Dorsomedial Striatum and the Dorsal Hippocampus are shown for all states. Furthermore, we show the progression of these brain state assignments over 3 hours and compare them to states from the method of [7], where statistics of the Hippocampus spectral density were clustered in an ad hoc fashion. To the best of our knowledge, this method represents the most relevant and accurate results for sleep staging from LFP signals

in the neuroscience literature. From these results, it is apparent that our clusters pick up sub-states of [7]. For instance, states 3, 6, and 7 all appear with high probability when the method from [7] infers state 3. Observing the cross-phase function of sub-state 7 reveals striking differences from other states in the theta wave (4-7 Hz) and the alpha wave (8-15 Hz). This cross-phase function is nearly identical for states 2 and 5, implying that significant differences in the cross-amplitude spectrum may have played a role in identifying the difference between these two brain states. Many more of these interesting details exist due to the expressive nature of the CSM kernel. As a full interpretation of the cross-spectrum results is not the focus of this work, we contend that the CSM kernel has the potential to have a tremendous impact in fields such as neuroscience, where the dynamics of cross-spectrum relationships of LFP signals are of great interest.

7

Conclusion

This work introduces the cross-spectral mixture kernel as an expressive kernel capable of extracting patterns for multi-channel observations. Combined with the powerful nonparametric representation of a Gaussian process, the CSM kernel expresses a functional form for every pairwise cross-spectrum between channels. This is a novel approach that merges Gaussian processes in the machine learning community to standard signal processing techniques. We believe the CSM kernel has the potential to impact a broad array of disciplines since the kernel can trivially be extended to any time-series application where Gaussian processes and the cross-spectrum are of interest. Acknowledgments The research reported here was funded in part by ARO, DARPA, DOE, NGA and ONR. 8

2 References

[1] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716?723, 1974. [2] M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. Foundations and Trends in Machine Learning, 4(3):195?266, 2012. [3] M. J. Beal. Variational Algorithms for Approximate Bayesian Inference. PhD thesis, University College London. [4] R. Caruana. Multitask learning. Machine Learning, 28(1):41?75, 1997. [5] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. SIAM Journal on Scientific Computing, 18(4):1088?1107, 1997. [6] K. Dzirasa, R. Fuentes, S. Kumar, J. M. Potes, and M. A. L. Nicolelis. Chronic in vivo multi-circuit neurophysiological recordings in mice. Journal of Neuroscience Methods, 195(1):36?46, 2011. [7] K. Dzirasa, S. Ribeiro, R. Costa, L. M. Santos, S. C. Lin, A. Grosmark, T. D. Sotnikova, R. R. Gainetdinov, M. G. Caron, and M. A. L. Nicolelis. Dopaminergic control of sleep?wake states. The Journal of Neuroscience, 26(41):10577?10589, 2006. [8] R. G. Gallager. Principles of digital communication. pages 229?232, 2008. [9] M. G?onen and E. Alpaydn. Multiple

kernel learning algorithms. JMLR, 12:2211?2268, 2011. [10] P. Goovaerts. Geostatistics for Natural Resources Evaluation. Oxford University Press, 1997. [11] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone. High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science, 324(5931):1207?1210, 2009. [12] M. L?azaro-Gredilla, J. Qui?nonero Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. JMLR, (11):1865?1881, 2010. [13] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. AAAI, 2014. [14] D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, 1997. [15] D. Pfaff, A. Ribeiro, J. Matthews, and L. Kow. Concepts and mechanisms of generalized central nervous system arousal. ANYAS, 2008. [16] C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. 2006. [17] P. Sauseng and W. Klimesch. What does phase information of oscillatory brain activity tell us about cognitive processes? Neuroscience and Biobehavioral Reviews, 32:1001?1013, 2008. [18] C. M. Sweeney-Reed, T. Zaehle, J. Voges, F. C. Schmitt, L. Buentjen, K. Kopitzki, C. Esslinger, H. Hinrichs, H. J. Heinze, R. T. Knight, and A. Richardson-Klavehn. Corticothalamic phase synchrony and cross-frequency coupling predict human memory formation. eLIFE, 2014. [19] Y. W. Teh, M. Seeger, and M. I. Jordan. Semiparametric latent factor models. AISTATS, 10:333?340, 2005. [20] M. A. Tucker, Y. Hirota, E. J. Wamsley, H. Lau, A. Chaklader, and W. Fishbein. A daytime nap containing solely non-REM sleep enhances declarative but not procedural memory. Neurobiology of Learning and Memory, 86(2):241?7, 2006. [21] K. Ulrich, D. E. Carlson, W. Lian, J. S. Borg, K. Dzirasa, and L. Carin. Analysis of brain states from multi-region LFP time-series. NIPS, 2014. [22] P. D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics, 15(2):70?73, 1967. [23] A. G. Wilson. Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes. PhD thesis, University of Cambridge, 2014. [24] A. G. Wilson and R. P. Adams. Gaussian process kernels for pattern discovery and extrapolation. ICML, 2013. [25] A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham. Fast kernel learning for multidimensional pattern extrapolation. NIPS, 2014. [26] A. G. Wilson and D. A. Knowles. Gaussian process regression networks. ICML, 2012. ? la carte ? learning fast kernels. AISTATS, 2015. [27] Z. Yang, A. J. Smola, L. Song, and A. G. Wilson. A

9