Gaussian Process Regression for Continuous Emotion Recognition with Global Temporal Invariance

Mia Atcheson M@ATCHESON.ORG
Vidhyasaharan Sethu V.SETHU@UNSW.EDU.AU
Julien Epps J.EPPS@UNSW.EDU.AU
School of Electrical Engineering and Telecommunications, University of New South Wales

Abstract

Continuous emotion recognition (CER) is a task which requires the prediction of time series emotional parameter outputs corresponding to query time series inputs given training data in the form of matched pairs of input and output time series. In order to address this task, it is important to be able to model not only relationships between points in the input and output spaces, but also temporal relationships between points within the output space. Gaussian process regression (GPR) is an inference technique which has desirable properties for CER, including its ability to produce predictive distributions over the outputs rather than only point estimates. However, GPR is generally applied to pointwise prediction or interpolation tasks, rather than to predictions of entire functional outputs. We propose a covariance structure that is able to incorporate both input-output and temporal information to produce predictions that take into account the functional nature of CER data. We demonstrate the application of this method to simulated data, and to the AVEC2016 CER task, showing that GPR with this covariance structure is able to make predictions of emotional arousal from audio with over twice the accuracy of a straightforward pointwise application of GPR in the input feature space, and is furthermore able to produce predictions with accuracy approaching that of a competitive CER system using only very general component covariance models.

1. Introduction

Observations of temporally varying phenomena can often give rise to data which are most naturally expressed as functions of time. If we are to carry out learning tasks which require the prediction of functional outputs, or which must generalize from functional inputs, it is necessary to develop inference systems which are able to model not only relationships between input and output observations at a particular moment in time, but also the temporal interrelationships between such data.

Continuous emotion recognition (CER) is a task with a functional prediction structure: CER systems aim to describe the emotional content of a communication as a continuous function of time mapping into a real vector valued emotional parameter space, which describes the emotional content at a particular moment by decomposing it into a small set of emotional parameters. This contrasts with earlier models that predicted discrete emotional labels, or assigned static emotional values to entire utterances rather than continuously as a function of time. Wöllmer et al. (2008); Gunes and Schuller (2013); Grimm et al. (2007).

Successful and widely used emotional parameters include *arousal*, describing the level of activity or excitement associated with an emotion, *valence*, describing the positive or negative evaluation associated with an emotion, and *dominance*, describing the level of social dominance or submissiveness conveyed by an emotional communication Wu et al. (2010). In particular, an arousal/valence

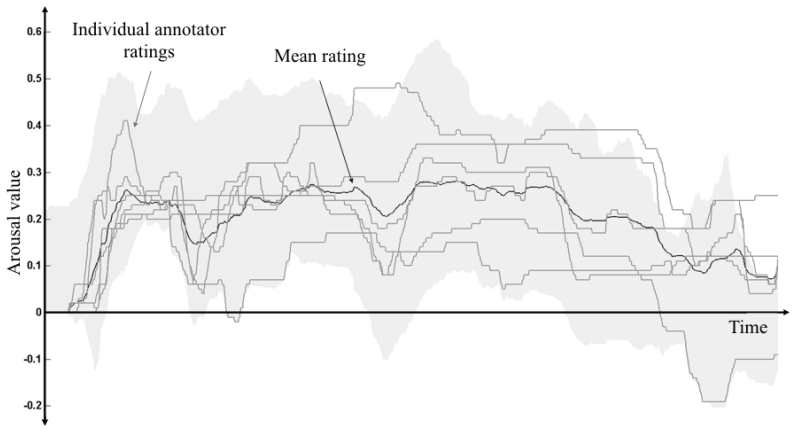

Figure 1: Example of a set of emotional parameter annotations. The horizontal axis denotes time, the grey lines the ratings submitted by individual human annotators, and the black line their mean. The shaded grey area represents a two standard deviation region around the mean. This uncertainty in estimating the emotional parameter values suggests that methods which are able to predict or train from distributions over such outputs may be more appropriate than methods which produce only point estimates.

emotional parameter space is often used in the CER setting. Due to the nature of emotion, these parameter values cannot be observed directly to produce training data, but must be estimated by humans; an example of a set of annotations for the arousal parameter over a short period of time is shown in Figure 1. However, each human annotator may not agree as to the correct emotional parameter value at any given point in time, so it is not possible to obtain a known-correct ground truth for training using this annotation method. This suggests that probabilistic methods which are able to predict and train to distributions over these parameters may be more suitable than methods which are able to model only pointwise outputs. Furthermore, we do not expect the nature of the relationships between observations such as speech recordings and the corresponding emotional parameter values to change significantly with time, nor do we expect the time indices of points residing in separate recordings to be meaningfully related. As such, methods which are able to model local, relative temporal structure without presuming a global temporal structure may be suitable for CER.

In practice, bidirectional long short-term memory neural networks (BLSTMs), a type of neural network architecture which allows for the modelling of temporal relationships using non-decaying memory cells, have been used successfully for CER Wöllmer et al. (2013); Pei et al. (2015). However, as a neural network method, BLSTM does not use an easily-interpretable model structure, so it is much less straightforward to incorporate *a priori* information into the model than with GPR covariance functions, and the trained network and its weights may be difficult for humans to meaningfully interpret. Furthermore, BLSTM also is not a probabilistic method, and provides only point predictions of output points, rather than the full posterior distribution produced by GPR.

Another method which has shown some success in CER is output-associative RVM (OA-RVM) Nicolaou et al. (2012), which is a modification to the relevance vector machine method to account for relationships between output points within a fixed temporal window, allowing it to perform

regularization as well as replication for functional prediction tasks. Additionally, support vector machines have been used for this task, but this is not a functional method, and does not itself account for temporal relationships Valstar et al. (2016). In order to apply SVR to this problem, the data must first be segmented into temporal windows which are processed individually, which requires an *a priori* choice of window length upon which the performance of the predictions may be strongly dependent.

Gaussian process regression (GPR) is an inference technique which allows the estimation of a posterior distribution over test outputs given a set of training pairs of input and output points. It is a fully Bayesian method, allowing it to deal directly with distributions over outputs. While GPR has generally been applied to interpolation or to the prediction of non-functional values, it permits the specification of a detailed covariance structure, so it may be applied to functional prediction tasks such as CER if a suitable covariance model is constructed.

While GPR has generally been applied to pointwise prediction, some attempts have been made to apply the technique to limited classes of functional prediction tasks. In Shi et al. (2005), an approach is developed which uses a mixture of Gaussian processes mediated by a latent indicator variable, and this model is applied to a task in which the center of mass of the body of a paraplegic patient in the process of performing an electrically-assisted standing-up maneuver is predicted from the forces and torques measured at the arms of the chair, under the patient's feet *etc*. In Shi et al. (2007), the model is modified to include a separate estimate of the mean structure of the predicted process, with GPR used to model the covariance structure. In Wang and Shi (2014), the approach is generalized to admit non-Gaussian output variables. However, this line of approach is not generally suitable for modelling functional relationships with only local temporal structure such as those encountered in CER, as it relies on an alignment between the time indices of separate series of observations, which is not generally applicable to CER.

In Lian (2010), a more general functional GPR model was developed, using a covariance function composed of the product of an isotropic covariance function in the input space with an isotropic covariance function in time. This model was demonstrated to produce successful results in a multiple-step-ahead time series prediction task, but due to the multiplicative covariance structure, when the time-based component tends to zero due to the temporal distance between points, the whole covariance function is forced to be small, such that points which are distant in time will have low covariances even if they are nearby in the input space. This precludes modelling of input-space relationships between temporally distant points, confining the applicability of this method to extrapolation or interpolation tasks in which only temporally local points are considered in each predictive distribution. However, for CER, we wish to be able to predict the emotional parameter values corresponding to recordings for which we have no access to emotional annotations, in which case neither interpolation or extrapolation is possible.

In this paper, we propose a covariance structure which can be used to apply GPR to functional inference tasks such as CER that require the modelling of both temporal relationships between points nearby in time, and non-temporal relationships between points at any temporal distance, using separate covariance functions which are combined additively. This approach is then demonstrated by applying it to both simulated data and a real-world CER task.

2. Continuous Prediction and Global Temporal Invariance

Gaussian process regression (GPR) is a Bayesian technique for supervised learning. A Gaussian process is simply a collection of random variables of which any finite subset has a jointly Gaussian distribution. GPR uses Gaussian processes as distributions over a function space, allowing Bayesian inference to be carried out to produce a posterior distribution over functional test points, conditioned on the known values at some training points. This technique has a number of desirable properties. As a fully-Bayesian inference method, it provides a joint posterior distribution over the test outputs, rather than merely a point estimate. Furthermore, there is no requirement for observations to be regular in time or space; samples can be handled at arbitrary coordinates.

When used in GPR, a Gaussian process is specified by its *mean function* $m(\boldsymbol{x})$ and *covariance function* (or *kernel function*) $\kappa(\boldsymbol{x}, \boldsymbol{x'})$. The mean function specifies the mean value of the output as a function of the corresponding input, while the covariance function specifies the covariance between any pair of output values $y = f(\boldsymbol{x})$ and $y' = f(\boldsymbol{x'})$ in terms of their corresponding *input* values $\boldsymbol{x}$ and $\boldsymbol{x'}$ Because the covariance function is defined in terms of the input values, which are known for both training and test points, we can write the joint distribution over a set of training output points Y corresponding to a matrix of input points X, and a set of test output points Y_* (which we intend to predict) corresponding to a matrix of test inputs X_*, even though the values Y_* are unknown.

GPR is a particularly suitable approach for affective computing problems such as CER due to its ability to predict distributions over the outputs rather than simply making a point prediction of each output value. Not only does this provide a measure of confidence in each prediction, which can be obtained by examining the variance of the output distribution, but it also allows the parameters of the system to be trained using a distribution over each training output rather than a single known ground truth value. Because it is not possible to directly observe emotional parameters, training data for CER tasks must be obtained by human annotation. However, each human annotator may not agree as to the correct emotional parameter value at any given point in time, so it is not possible to obtain a known-correct ground truth for training using this annotation method.

An important property of the relationship between the inputs and outputs in a CER setting is *global temporal invariance*: we do not expect the nature of the relationship between input and output points to depend on their absolute time indices, but we do wish to model relative temporal relationships within the same time series. In CER, this condition corresponds to the practical fact that the start point $t = 0$ of each recording is arbitrary: we would not expect the observation at n seconds after the start of a given recording to have any particular relationship with the observation at n seconds after the start of another recording made at a different time, nor would we expect the relationship between the inputs and outputs to change if we were to shift the time indices of each observation by a constant amount.

3. Proposed Covariance Structure

We wish to construct a covariance structure which will allow GPR to be applied to tasks such as CER where training data consist of pairs of corresponding input and output time series, and at test time, we wish to generate a predictive distribution over an output time series given an input query time series. Our proposed approach to applying GPR to this class of problem is to combine two separate covariance functions: one, κ_x, to represent relationships between input and output points, and one, κ_t, to represent temporal relationships amongst output points, each of which can be independently specified according to the structure of the problem. A weighted sum is then taken over the kernels,

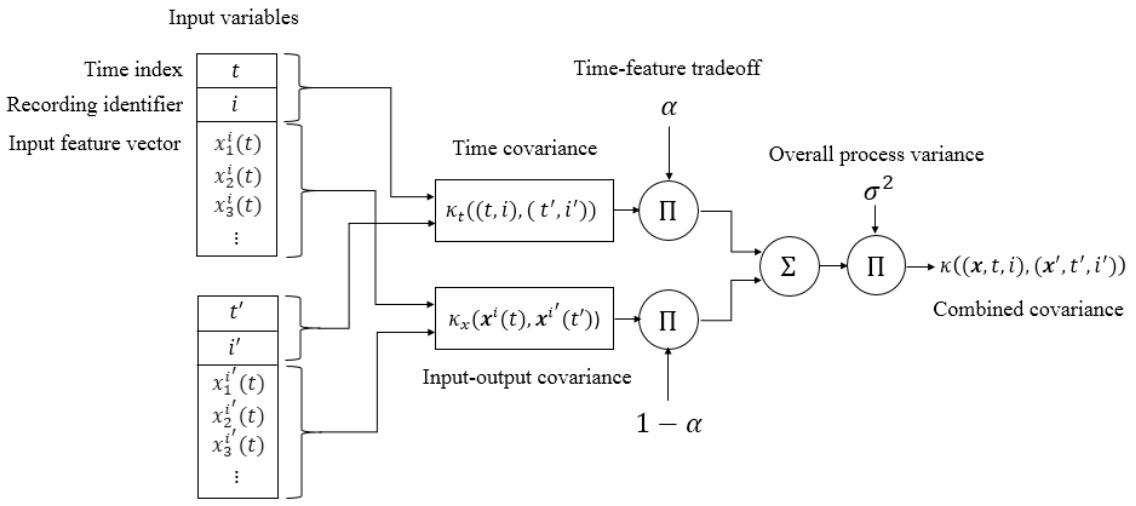

Figure 2: Diagrammatic representation of the covariance structure specified in Equation 1. Two input observations are shown on the left; time indices t and t' and time-series identifiers i and i' are processed by the time kernel κ_t, and the observation vectors $x^i(t)$ and $x^{i'}(t')$ for each feature modality are processed by the input-output kernel κ_x. These components are then combined according to the tradeoff parameter α. The result is then scaled by σ^2 to produce the final covariance value.

which is a novel approach to the combination of temporal and feature-space information for GPR:

$$\begin{aligned}&\kappa((\boldsymbol{x}, t, i), (\boldsymbol{x}, t', i')) \\ &\quad = \sigma^2 \left(\alpha \kappa_t((t, i), (t', i')) + (1 - \alpha)\kappa_x(\boldsymbol{x}, \boldsymbol{x}')\right)\end{aligned} \tag{1}$$

where each input point has the form $(\boldsymbol{x}, t, i)$, where t is the time index corresponding to observation $\boldsymbol{x}$ and i is a unique identifier representing the particular series of observations (individual recording) of which $(\boldsymbol{x}, t)$ is a part. In Equation 1, α controls the tradeoff between the relative importance placed on temporal vs input-output relationships, and σ^2 controls the overall variance of the process. Figure 2 shows a diagrammatic representation of this covariance structure. This additive model was chosen as it is able to model relationships between points based on either their relative temporal locations or their positions in the input space, unlike multiplicative models such as proposed in Lian (2010), where temporally unrelated points will have a low covariance even if they are strongly related under the input-output kernel, and vice-versa. To ensure global temporal invariance, we restrict κ_t as follows:

$$\kappa_t((t, i), (t', i')) = \begin{cases} \kappa_{t*}(t - t') & \text{where } i = i' \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

that is, the relationships between points can depend only on relative temporal distances within a single recording, not on the absolute time index of any point. Furthermore, as points in separate recordings have no meaningful temporal relations, in these cases we simply set κ_t to zero.

This covariance structure generalizes two standard approaches to the practical application of GPR: the use of a one-dimensional temporal covariance function to perform pointwise interpolation in time, and the use of a typically multidimensional covariance function to model a mapping between individual points in an input and output space. This approach combines the two, with the influence of each model controlled by a single tradeoff parameter α. This allows the system to predict the values of a time-varying function within an output space based on a corresponding time-varying function within an input space, while simultaneously imposing temporal relationships (generally, smoothness or periodicity conditions) on the output function and carrying out interpolation from known temporally nearby values within the output space if they are available.

4. Experiments

4.1. Simulated data

To demonstrate the advantages of the proposed approach over the straightforward application of GPR to the input-output relationships and over a multiplicative model similar to that proposed in Lian (2010), we constructed a simple data set with the following components:

- A time index set T, of which input and output data are functions.
- An output function $f_y(t) : T \mapsto Y = \mathbb{R}$, which we expect to vary smoothly with time.
- An input function $f_x(t) : T \mapsto X = \mathbb{R}^2$, for which if $f_x(t)$ and $f_x(t')$ are similar, then $f_y(t)$ and $f_y(t')$ will tend to be similar.

To model these relationships with GPR, we can construct a composite covariance function κ by first specifying functions κ_x, and κ_t to model the non-temporal relationships between points in X and Y, and the relative temporal relationships between points within T respectively. As component kernels, we use squared-exponential kernels which decrease isotropically with distance and have characteristic lengthscale l, evaluated over two dimensions for κ_x and one for κ_t:

$$\kappa_x(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{x}'|^2}{2l_x^2}\right) \tag{3}$$

$$\kappa_t(t, t) = \exp\left(-\frac{|t - t'|^2}{2l_t^2}\right) \tag{4}$$

$$\kappa((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = \sigma^2\left(\alpha\kappa_x(\boldsymbol{x}, \boldsymbol{x}') + (1 - \alpha)\kappa_t(t, t)\right) \tag{5}$$

For the simulation experiment, the time series $f_x(T)$ was generated randomly by taking 40 normally-distributed points at evenly spaced intervals along 2000 time indices, which were then interpolated with cubic curves to produce smoothly-varying functions of time. $f_y(T)$ was then generated by sampling from a Gaussian distribution with a covariance matrix derived from κ with preset parameters $l_t = 10, l_x = 1.5, \alpha = 0.5, \sigma^2 = 2$, which were not used during training. The central 500 points of $f_y(T)$ were set aside as test data, and the others were used to train the system (that is, three times as much training data as test data). In the training process, the kernel parameters were estimated by maximizing the likelihood of the training data. Then, the trained kernel was used to predict the 500 test points, and the concordance correlation coefficient between the mean values of the predictive distributions and the known test outputs was evaluated. The concordance correlation coefficient ρ_c

is a modification of the standard Pearson correlation coefficient which penalizes differences in both shape and magnitude:

$$\rho_c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{6}$$

where ρ is the Pearson correlation coefficient between the sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, μ_x and μ_y are the corresponding means, and σ_x^2 σ_y are the variances. The component kernels κ_x and κ_t were also used to make predictions which were evaluated against the test data. We also tested a multiplicative kernel $\kappa_\times$:

$$\kappa_\times((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = \sigma^2 \left(\kappa_x(\boldsymbol{x}, \boldsymbol{x}')\kappa_t(t, t)\right) \tag{7}$$

4.2. Results

This process was performed 50 times with newly-generated data each time. The average concordance values obtained were 0.73 for κ, 0.00 for κ_t, 0.02 for κ_x and 0.39 for $\kappa_\times$, where a higher value corresponds to a more accurate prediction of the test data. While neither the input-output model κ_x or the temporal model κ_t were able to produce reliably accurate predictions on this data, the models κ and $\kappa_\times$, by combining these two kernels, were able to produce significantly more robust predictions. Figure 3 shows the results obtained for a single simulated dataset. Note that while the predictions produced by κ_x, which do not use any temporal information, vary more rapidly than the true values, the addition of the time component κ_t to produce the composite kernel κ imposes a smoothness condition on the output which reduces the rate at which the prediction varies, and under this condition, the system is able to use both the input values $\boldsymbol{x}$ and the corresponding time indices t to produce a prediction which more closely resembles the true output function. Furthermore, in the regions near the edge of the 500 point test region, where temporally nearby training outputs exist, the time kernel κ_t is also able to provide an extrapolative capability. Both κ and $\kappa_\times$ benefit from this effect, and as such they both produce accurate predictions in these edge regions. However, within the central region, where there are no temporally nearby training points from which to extrapolate, the temporal covariance component κ_t between points in this region and training points is close to zero. In the multiplicative model $\kappa_\times$, this means that the complete covariance is forced to be close to zero, limiting the ability of this model to make meaningful predictions from the training data in this region. However, in the proposed additive model κ, a zero temporal covariance merely causes the predictions to rely only on the input-output covariance model κ_x.

4.3. Continuous Emotion recognition

To demonstrate the applicability of the proposed approach to CER, we applied it to the AVEC 2016 affect recognition sub-challenge Valstar et al. (2016), which requires continuous arousal and valence values to be predicted for spontaneous, task-driven dyadic interactions using the recorded outputs of audio, video and electrophysiological sensors, along with annotations of perceived arousal and valence values produced by a panel of humans to obtain an approximation to the ground-truth emotional content of the training recordings. However, in this case, we restricted our attention to audio data only, and we attempt to predict arousal, as it is considered to have a stronger relationship with audio-derived acoustic features than valence Gunes and Schuller (2013). As per the challenge specifications, the accuracy of predictions was evaluated using the concordance correlation coefficient between predicted time series (which in this case will be taken as the mean of the predicted distribution) and the average of the human annotations.

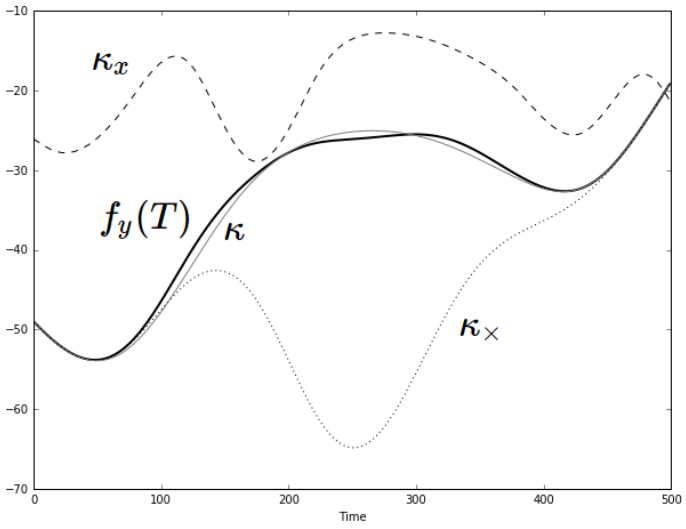

Figure 3: Predicted values from randomly-generated simulated data. The bold solid line represents the true test output generated to have covariances specified by a model covariance function with both temporal and input-output relationships, the thin solid line the mean value of the predictive distribution produced by the model covariance function κ, the dashed line the mean value of the predictive distribution produced by the non-temporal component kernel κ_x, and the dotted line he mean value of the predictive distribution produced by the multiplicative kernel $\kappa_\times$. The predictions produced by the feature-only kernel κ_x do not benefit from extrapolation in the edge regions or regularization in the central region, and so this model is not able to produce an accurate prediction relative to κ, which additionally accounts for temporal relationships. Further, while both the additive model κ and the multiplicative model $\kappa_\times$ incorporate both temporal and input-output relationships, the multiplicative model is unable to produce accurate predictions in the central regions where no temporally nearby training points are present.

In order to apply the proposed approach to this task, we must specify a covariance function κ_x to represent relationships between the content of the input audio recording at a particular point in time and the corresponding output emotional parameter value, a covariance function κ_t to represent the temporal interrelationships between output points, and a method of optimization to determine the values of the system parameters. For the temporal covariance model κ_t, we use a squared-exponential function:

$$\kappa_T((t,i),(t',i')) = \begin{cases} \exp\left(-\frac{|t-t'|^2}{2l_t^2}\right) & \text{where } i = i' \\ 0 & \text{otherwise} \end{cases}$$

This covariance function corresponds to the assumption that the emotional content of a communication does not change arbitrarily with time: if you are angry at a particular moment in time, then you are more likely to remain angry for the following second than to suddenly become happy. In

the case where the points are from different recordings (that is, where $i \neq i'$, and therefore t and t' are not comparable), then we do not attempt to model a temporal relationship between them, simply setting κ_t to 0.

To model the relationships between the content of the audio recordings, we first applied feature extraction using the eGeMAPS acoustic feature set with a 4 second window for the calculation of functionals, which is specified as the standard audio feature set for arousal in the AVEC2016 affect recognition challenge Eyben et al. (2016); Valstar et al. (2016). Then, in order to reduce the computational resources required for model training, the 88 eGeMAPS features were then reduced to 40 dimensions using PCA. We then treat these 40 principal components as the inputs to κ_x. However, a simple spherical covariance function such as the squared-exponential used for κ_t may not be suitable for κ_x, as the 40-dimensional input space is large and our data is limited, so we may need to make predictions for query points which have no nearby training points in the input space. As such, we used a multidimensional additive covariance function as presented in Duvenaud et al. (2011), which allows for the modelling of interactions within any subset of the dimensions of the feature space. As component kernels, we used 40 one-dimensional squared exponential functions, each corresponding to one of the principal component axes. The hyperparameters of the additive kernel were optimized by maximizing the likelihood of the training data by gradient ascent. Then, the remaining parameters of the system were optimized using the Spearmint black-box optimizer Snoek et al. (2012) targeting the average concordance correlation coefficient produced on a two-fold cross validation over the training data.

With a system trained as described above, a concordance correlation coefficient of 0.598 was obtained on the training set when predicting the averaged arousal annotations. For comparison, the AVEC2016 baseline system was able to produce a coefficient of 0.648 Valstar et al. (2016). As such, this indicates that while the implemented system does not outperform the state-of-the-art, it is able to produce reasonable performance on a difficult real-world CER task, which suggests that the proposed approach of using GPR with additively combined temporal and input-output covariance models has some applicability to this type of task. To demonstrate the contribution of the temporal component of the model, we randomly partitioned the training data, using two thirds to train a system which was tested on the remaining third. In each experiment, all parameters of the model except α were kept constant. By varying α, we can isolate the effect of the addition of the time kernel κ_t: when $\alpha = 0$, only κ_x is used to make predictions, which corresponds to an ordinary pointwise use of GPR. As α increases, the influence of the temporal component increases and eventually comes to dominate. In this practical case, the size of the covariances produced by κ_x are of a much greater magnitude than the covariances produced by κ_t; as such, we define $\alpha = 1 - 10^{-a}$, as the values of interest for α are close to 1. Figure 4 shows a plot of the concordance correlation coefficients achieved by models with various values of a. As can be seen in the figure, while at $\alpha = 0$ ($a = 0$), where the model uses only input-output relationships, the prediction generated is meaningful, with a concordance correlation coefficient of 0.38, the introduction of the time kernel is able to increase the accuracy of the predictions up to a coefficient of 0.78 where $\alpha = 1 - 10^{-7.5}$. Past this point, the performance reduces rapidly as the time kernel comes to dominate, and the covariance structure is no longer able to properly model input-output relationships.

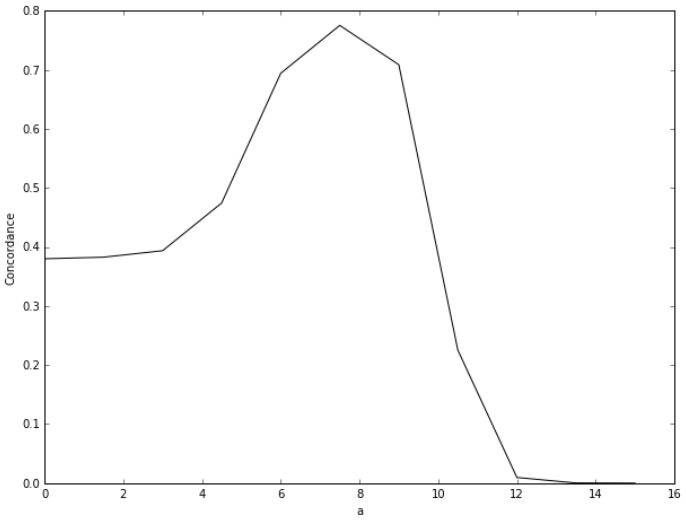

Figure 4: Concordance correlation coefficients obtained when varying α while making predictions of the arousal parameter corresponding to partition of the AVEC2016 training data. The horizontal axis represents a where $\alpha = 1 - 10^{-a}$. At $a = 0$, only the input-output covariance model is used, corresponding to an straightforward pointwise use of GPR. As a increases, the relative influence of the temporal model is increased, which improves the performance of the system up to an optimum value around $a = -7.5$, after which the accuracy drops towards zero as information about the inputs is attenuated.

5. Conclusion

In this paper, we propose a covariance model which allows Gaussian process regression to be applied to functional prediction tasks that require modelling of both input-output relationships and temporal relationships. To construct this model, we additively combine separate covariance functions to represent temporal and input-output relationships, and we restrict the temporal covariance function so as to incorporate an assumption of global temporal invariance: only relative temporal distances are considered, not the absolute time indices of each point. This model was then applied to both simulated data and a real-world CER task, demonstrating its ability to make predictions of emotional arousal from audio with accuracy approaching that of a competitive baseline system, while offering the qualitative advantages of fully Bayesian inference. Furthermore, in experiments using a partitioning of the training data, we show that the addition of the proposed temporal structure to the covariance model produces predictions with over twice the accuracy of a straightforward pointwise application of GPR in the input feature space.

References

David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen. Additive Gaussian processes. In *Advances in neural information processing systems*, pages 226–234, 2011.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10):787–800, 2007.

Hatice Gunes and BjöRn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

Heng Lian. Gaussian process models for nonparametric functional regression with functional responses, 2010.

Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, 2012.

Ercheng Pei, Le Yang, Dongmei Jiang, and Hichem Sahli. Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 208–214. IEEE, 2015.

Jian Qing Shi, Roderick Murray-Smith, and DM Titterington. Hierarchical Gaussian process mixtures for regression. *Statistics and Computing*, 15(1):31–41, 2005.

JQ Shi, B Wang, Roderick Murray-Smith, and DM Titterington. Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723, 2007.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Michel Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Guiota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016-depression, mood, and emotion recognition workshop and challenge. *arXiv preprint arXiv:1605.01600*, 2016.

Bo Wang and Jian Qing Shi. Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association*, 109(507):1123–1133, 2014.

Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, Roddy Cowie, et al. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, pages 597–600, 2008.

Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

Dongrui Wu, Thomas D Parsons, Emily Mower, and Shrikanth Narayanan. Speech emotion estimation in 3d space. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 737–742. IEEE, 2010.