Twin Gaussian Processes for Structured Prediction

Liefeng Bo and Cristian Sminchisescu
TTI-C, *{blf0218,crismin}@tti-c.org*

We describe twin Gaussian processes (TGP), a structured prediction method that uses Gaussian processes (GP) priors for both covariates and responses, and estimates outputs by minimizing the Kullback-Leibler divergence between two GP modeled as normal distributions over finite index sets. TGP captures not only the interdependencies between covariates (as in a typical GP) but also those between responses, so correlations among both inputs and outputs are accounted for. We demonstrate the method for the reconstruction of 3d human poses in monocular video in the standard HumanEva test bench (see table 1), but the idea applies to other structured learning tasks.

Consider a visual task where we aim to recover, *e.g.*, the 3d pose of a human from an image. This is difficult primarily because human poses are high dimensional—anthropometric models have in the order of 30-60 joint angles or joint positions depending on the desired accuracy. But human poses and motions are also structured due to typical human constraints like balance and synchronicity of movement, as apparent in walkers, runners, or dancers. Recently, there has been an increasing interest in approaches to 3d human pose reconstruction [2, 4] based on learning predictive models trained using sets of image features and corresponding 3d poses. A shortcoming of existing methods is that they do not sufficiently exploit interdependencies between outputs. Our work aims to improve the modeling of correlations by using GP and a new prediction criterion.

While Gaussian processes [1] are powerful tools for modeling non-linear input-output dependencies, being potentially interesting for 3d human pose reconstruction, most GP models focus on the prediction of a single output. Although generalizations to multiple outputs can be derived by training independent models for each output, this fails to leverage information about correlations between variables into the predictor [5]. This motivates our TGP model, that encodes the relations between both inputs and outputs using GP priors. Since samples from two Gaussian processes reflect marginal similarities among inputs and outputs, the idea is to match them as well as possible. This emphasizes the objective that, *e.g.*, for perceptual inference, similar images should lead to similar 3d percepts and vice-versa. This goal is achieved by minimizing the Kullback-Leibler divergence between the two marginal GP, modeled as normal distributions over sets of finite indices.

Given a training set of inputs $R = \{\mathbf{r}_i\}_{i=1}^n$ and outputs $X = \{\mathbf{x}_i\}_{i=1}^n$, let $C_R(\mathbf{r},\mathbf{r})$ the covariance function defined over inputs and $C_X(\mathbf{x},\mathbf{x})$ the covariance function defined over outputs. We specify a joint Gaussian distribution over the training inputs and the test input $\mathbf{r}$: $\begin{bmatrix}\mathbf{f}_R \\ f_r\end{bmatrix} \sim \mathcal{N}_R\left(\mathbf{0}, \begin{bmatrix}\mathbf{C}_R & \mathbf{C}_R^{\mathbf{r}} \\ (\mathbf{C}_R^{\mathbf{r}})^\top & C_R(\mathbf{r},\mathbf{r})\end{bmatrix}\right)$, and a joint Gaussian distribution over training outputs and the test output $\mathbf{x}$ (we aim to estimate $\mathbf{x}$): $\begin{bmatrix}\mathbf{f}_X \\ f_x\end{bmatrix} \sim \mathcal{N}_X\left(\mathbf{0}, \begin{bmatrix}\mathbf{C}_X & \mathbf{C}_X^{\mathbf{x}} \\ (\mathbf{C}_X^{\mathbf{x}})^\top & C_X(\mathbf{x},\mathbf{x})\end{bmatrix}\right)$, where $\mathbf{C}_X$ is a $n \times n$ matrix with $(\mathbf{C}_X)_{ij} = C_X(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{C}_X^{\mathbf{x}}$ is a $n \times 1$ column vector with $(\mathbf{C}_X^{\mathbf{x}})_i = C_X(\mathbf{x}_i, \mathbf{x})$. TGP makes predictions by minimizing the Kullback-Leibler divergence between $\mathcal{N}_X$ and $\mathcal{N}_R$ with respect to the output vector $\mathbf{x}$:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{X}}(D_{KL}(\mathcal{N}_X \parallel \mathcal{N}_R)) \tag{1}$$

Invoking related matrix identities and dropping constant terms, we simplify (1) to:

$$\begin{aligned}\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{X}}(D_{KL}(\mathcal{N}_X \parallel \mathcal{N}_R) &= C_X(\mathbf{x},\mathbf{x}) - 2(\mathbf{C}_X^{\mathbf{x}})^\top \mathbf{C}_R^{-1}\mathbf{C}_R^{\mathbf{r}} \\ -(C_R(\mathbf{r},\mathbf{r}) - (\mathbf{C}_R^{\mathbf{r}})^\top \mathbf{C}_R^{-1}\mathbf{C}_R^{\mathbf{r}}) &\ln(C_X(\mathbf{x},\mathbf{x}) - (\mathbf{C}_X^{\mathbf{x}})^\top \mathbf{C}_X^{-1}\mathbf{C}_X^{\mathbf{x}}))\end{aligned} \tag{2}$$

TGP requires the inversion of $n \times n$ covariance matrices and are impractical for large problems. This can be overcome by combining the model with K nearest neighbors (TGP-KNN): find the K nearest neighbors of a test input and perform TGP only on this reduced dataset. To estimate outputs we use a quasi-Newton optimizer. In our experiments, this converged in less than 10 iterations on average.

For illustration, we analyze the HumanEva dataset [3], which contains a number of sequences of walking, jogging, throw-catch, gestures, and boxing, to a total of 5942 training and 5832 test samples (the backgrounds are known and fairly uniform, hence silhouettes can be computed). The training set consists of pairs of human (image) silhouette-based descriptors and human poses represented as 45d vectors of three-dimensional body joint positions. All poses are preprocessed by subtracting the root joint location from all the joint centers for every frame. We use datasets corresponding to the same set of human motions, captured by different video cameras: C1, C2, C3. To compute shape context descriptors (HistoSC), edges are extracted from the silhouette image and 400 points are sampled on edges, both internal and external. The shape context descriptor at each of the points is computed based on 15 angular bins and 8 radial bins. The SC at each of the 400 points per image, accumulated over images subsampled from the training set (typically every 15) is used to generate a codebook using vector quantization. The codebook has 300 clusters obtained using k-means (hence the descriptor size is 300). The models we use, trained separately for each viewpoint, are K-Nearest Neighbor (KNN), Ridge Regression (RR), Gaussian Processes (GP) and our Twin GP, with and without KNN preprocessing. The free parameters of each model were optimized using 10-fold cross validation over the training set. Results (mean absolute errors on test set) are shown in table 1. Output computation for TGP requires non-linear optimization. We use a quasi-Newton method intialized with RR—in our experiments this improved the estimate significantly compared with random initialization.

Table 1: Comparative evaluation of different predictors on HumanEva-1 [3] (prediction error per human body joint position, in mm). Shape context features are used as inputs (HistoSC), and we work with separate training and test sets for each viewpoint C1, C2, C3. KNN and TGP-KNN use 25 and 500 nearest neighbors, respectively. TGP outperforms KNN, RR and GP methods.

Features/View	RR	KNN	GP	TGP	TGP-KNN
HistoSC/C1	49.3	39.9	36.4	28.5	28.9
HistoSC/C2	45.3	39.3	33.6	27.3	28.0
HistoSC/C3	44.9	38.0	31.9	25.4	25.9

Topic: visual processing and pattern recognition **Preference: poster**

References

[1] K. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[2] R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *NIPS*, 2002.

[3] L. Sigal and M. Black. HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. Technical Report CS-06-08, Brown University, 2006.

[4] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, volume 1, pages 390–397, 2005.

[5] J. Weston, O. Chapelle, A. Elisseeff, B. Scholkopf, and V. Vapnik. Kernel dependency estimation. In *NIPS*, 2002.