Contribution
We extend the multi-task learning model with two novel aspects. First, we allow the fixed effect to be multi-modal so that each task may draw its fixed effect from a different cluster. Second, we extend the model so that each task may be an arbitrary phase-shifted image of the original time series. This yields our model the Shift-invariant Grouped Multi-task Learning for Gaussian Processes (GMT).

Outline of the inference algorithm
Our main technical contribution is the inference algorithm for the proposed model. We develop details for the EM algorithm optimizing the Maximum A Posteriori (MAP) estimates for the parameters of the model. Technically, the two main insights are in estimating the expectation for the coupled hidden variables (the cluster identities and the task specific portion of the time series) and in solving the regularized least squares problem for a set of phase-shifted observations. As a special case our algorithm yields the Gaussian mixture model (GMM) for phase shifted time series.

Model
The generative model, characterized by parameter set \( \{K_0, \alpha, \{t_j\}, \sigma^2\} \) is summarized as follows:

\[
\begin{align*}
\tilde{f}_s \mid K_0 & \sim \exp \left\{ -\frac{1}{2} \| \tilde{f}_s \|_{K_0}^2 \right\}, z_j \mid \alpha \sim \text{Discrete}(\alpha) \\
\tilde{f}^j \mid K & \sim \exp \left\{ -\frac{1}{2} \| \tilde{f}^j \|_{K}^2 \right\}, f^j = \tilde{f}_s * \delta_{t_j} + \tilde{f}^j, y^j \sim N \left( f^j(x^j), \sigma^2 \right),
\end{align*}
\]

where \( k \) is the number of centers which is determined \( \text{a priori}, M \) is number of instances and \( s = 1, \ldots, k, j = 1, \ldots, M \). \( f + \delta_t \) denotes the circular right shift of \( f \) by time \( t \).

Inference

In the Expectation step, the posterior distribution of the hidden variable are

\[
Pr(x, f | Y, M; \mathcal{M}^o) = \prod_j Pr(z_j, f_j | Y, M; \mathcal{M}^o) = \prod_j Pr(z_j | f_j, Y, M; \mathcal{M}^o) \times Pr(f_j | z_j, Y, M; \mathcal{M}^o)
\]

which is equal to a multinominal distribution times \( \mathcal{N}(\mu_\ell, C^\ell) \). By a version of Fubini’s theorem we have \( \mathbb{E}_{(x,f,Y;M^o)} \log \mathcal{L} = \mathbb{E}_{(x,f,Y;M^o)} \mathbb{E}_{(f,z_j,Y;\mathcal{M}^o)} \log \mathcal{L} \), therefore

\[
Q = \frac{1}{2} \sum_s \| \tilde{f}_s \|_{K_0}^2 + \sum_s Pr(x_s | f_s, Y, \mathcal{M}^o) \left\{ \sum_j \sum_s z_{js} \times \int dPr(f_j | z_j = s) \log \alpha_s Pr(y^j | f_j, z_j = s; \mathcal{M}) Pr(f_j; \mathcal{M}) \right\}.
\]

In the Maximization step, \( M^o = \arg \max_{M} Q(f, \{t_j\}, \sigma^2) = \frac{1}{2} \sum_s \| \tilde{f}_s \|_{K_0}^2 + \sum_s n_j \log \sigma + \frac{1}{2\sigma^2} \sum_j \text{Tr}(C^j_f) + \frac{1}{2\sigma^2} \sum_j \sum_s \gamma_{js} \left( \| y^j - [f_s * \delta_{t_j}(x^j) - \mu_{\ell_j}] \|_{C^j_f}^2 \right) \]

We solve \( Q_1 \) using a coordinate descent algorithm while solving a weighted regularized least square problem inside each iteration. \( Q_2 \) can be solved directly to yield the kernel matrix in the nonparametric case. For parametric kernels we use Conjugate Gradients to optimize \( Q_2 \).

Regression on Synthetic data
We generated non-phase-shifted data with \( k = 2 \). The right figure shows a comparison of the learned functions among single, multi-task, and grouped multi-task learning where the red dotted line is the reference true function. The left two figures show a comparison of the RMSE. The left one gives 3 pairwise comparisons when sample size is 5 and the right one shows RMSE as a function of sample size.

Astrophysics time series
The concrete application motivating this research is the classification of stars into several meaningful categories from the astronomy literature. We evaluate our method on the OGLEII dataset, which consists of 14087 periodic variable stars with (3425,3390,7272) in the categories (CEPH, RRL, EB); 2 examples per class are shown in corresponding columns in the figure below. We run our model on densely and synchronously sampled data as well as sparsely and non-synchronously sampled time series. GMT is significantly better with sparse data.

Acknowledgement
This research was partly supported by NSF grant IIS-0803409.