

Robust Multi-Task Learning with t -Processes

Shipeng Yu, Volker Tresp, and Kai Yu

in Proc. ICML, Corvallis, OR, June 2007

Presenter: Ivo D. Shterev

Overview

- Motivation
- Properties of t -processes (TP)
- Multi-task learning (MTL) with TP
- MTL without missing labels
- MTL with missing labels
- Discussion
- Empirical study

Motivation

- MTL - tasks *sharing* common parameters.
- in most applications tasks are equally weighted.
- in some applications, like rating systems, it is important to distinguish between "good" and "bad" tasks, i.e. provide *robustness* to MTL systems.
- TP is a generalization of Gaussian Processes (GP) and therefore allows for greater flexibility in distinguishing between "good" and "bad" tasks

Properties of TP

- probability density function $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$P(\mathbf{x}) = \pi^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \nu^{\frac{\nu}{2}} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})} \left(\nu + (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)^{-\frac{\nu+d}{2}}$$

- where $\Gamma(\cdot)$ is the Gamma function, $\nu > 0$ is the degree of freedom, and d is the dimension
- generation

$$\tau \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \text{ and } \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}$$

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\tau} \boldsymbol{\Sigma}\right)$$

$$\frac{1}{\nu + 1} \sum_{i=1}^{\nu+1} \mathbf{x}_i \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Properties of TP

- properties

$$\lim_{\nu \rightarrow \infty} t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- if $\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and \boldsymbol{W} is a real matrix, then

$$\boldsymbol{W}\boldsymbol{x} \sim t_\nu(\boldsymbol{W}\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)$$

- let $\boldsymbol{x} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$,

$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ be the $[d_1, d - d_1]$ partition, then

$$\boldsymbol{x}_1 \sim t_\nu(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

$$\boldsymbol{x}_2 | \boldsymbol{x}_1 \sim t_{\nu+d_1}(\boldsymbol{\mu}_{\boldsymbol{x}_2 | \boldsymbol{x}_1}, \boldsymbol{\Sigma}_{\boldsymbol{x}_2 | \boldsymbol{x}_1})$$

Properties of TP

• where

$$\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1} = \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1} = \frac{\nu + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)}{\nu + d_1} (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})$$

• Student's distribution ($d = 1, \boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = 1$)

$$P(x) = \pi^{-\frac{1}{2}} \nu^{\frac{\nu}{2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} (\nu + x^2)^{-\frac{\nu+1}{2}}$$

Properties of TP

- Student's distribution has heavier tails

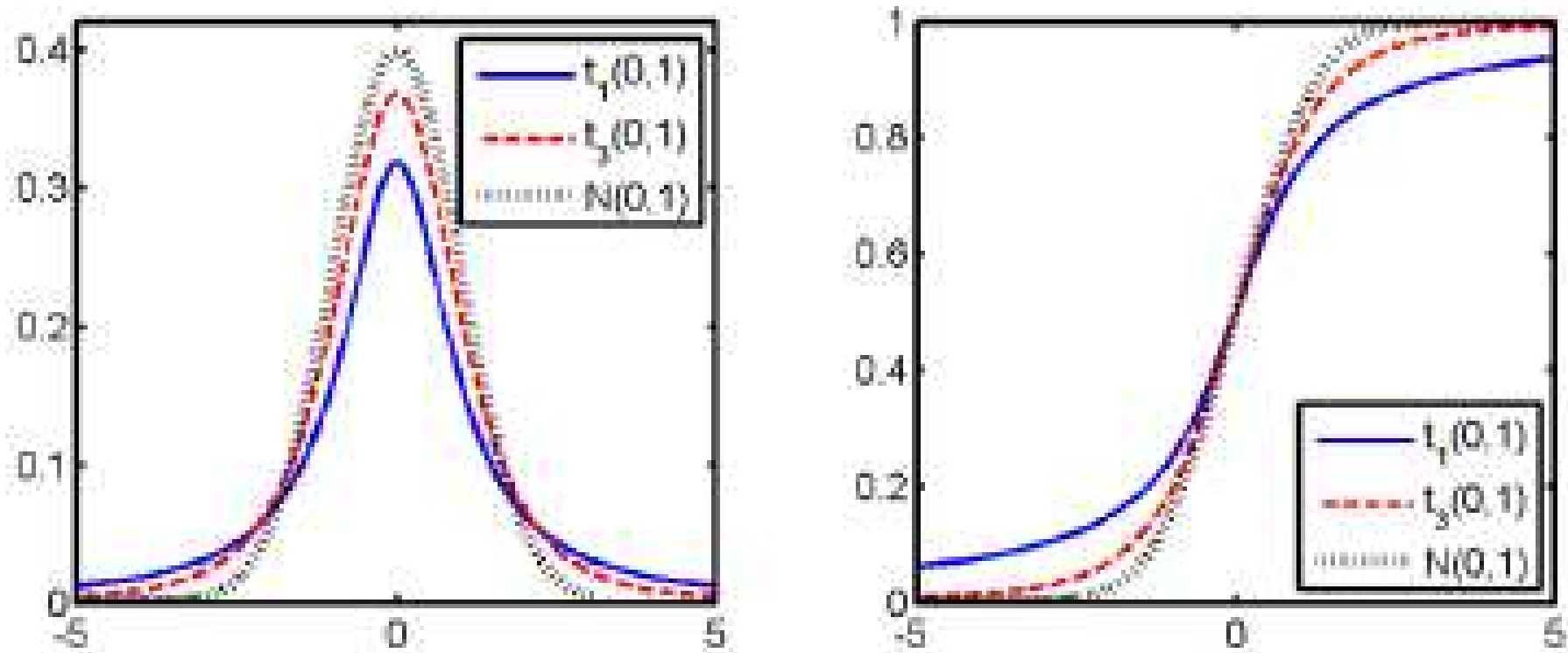


Figure 1. P.d.f. (left) and c.d.f. (right) of one dimensional t distribution $t_1(0, 1)$, $t_3(0, 1)$ and $\mathcal{N}(0, 1) = t_{+\infty}(0, 1)$.

Properties of TP

- data points

$$\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$$

- random function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ follows a $\mathcal{GP}(h, k)$

- if $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n \sim \mathcal{N}(\mathbf{h}, \mathbf{K})$

- where covariance and mean functions

$$\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n, k : \mathbb{R}^d \rightarrow \mathbb{R}$$

$$\mathbf{h} = \{h(\mathbf{x}_i)\}_{i=1}^n, h : \mathbb{R}^d \rightarrow \mathbb{R}$$

- marginals are preserved

$$\int \mathcal{GP}(\mathbf{f}; \mathbf{h}, \mathbf{K}) df_1 \dots df_{i-1} df_{i+1} \dots df_n \sim \mathcal{GP}(h_i, K_i)$$

Properties of TP

- data points

$$\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$$

- random function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ follows a $\mathcal{TP}_\nu(h, k)$
- follows a \mathcal{TP} if $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n \sim \mathcal{TP}_\nu(\mathbf{h}, \mathbf{K})$
- using the properties of TP, it can be shown that marginals are preserved, i.e.

$$\int \mathcal{TP}_\nu(\mathbf{f}; \mathbf{h}, \mathbf{K}) df_1 \dots df_{i-1} df_{i+1} \dots df_n \sim \mathcal{TP}_\nu(h_i, K_i)$$

Properties of TP

- sampling $f \sim \mathcal{TP}_\nu(h, k)$ is equivalent to

$$\tau \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), f \sim \mathcal{GP}\left(h, \frac{1}{\tau}k\right)$$

- if prior $f \sim \mathcal{TP}_\nu(h, k)$ and $\mathbf{f}_n = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^T$, then

$$f | \mathbf{f}_n \sim \mathcal{TP}_{\nu+n}(h^*, k^*)$$

- where

$$h^*(\mathbf{x}) = \mathbf{k}_x^T \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{h}_n) + h(\mathbf{x})$$

$$k^*(\mathbf{x}_i, \mathbf{x}_j) = \frac{\nu + (\mathbf{f}_n - \mathbf{h}_n)^T \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{h}_n)}{\nu + n} (k_{ij} - \mathbf{k}_{\mathbf{x}_i}^T \mathbf{K}_n^{-1} \mathbf{k}_{\mathbf{x}_j})$$

$$\mathbf{h}_n = [h(\mathbf{x}_1), \dots, h(\mathbf{x}_n)]^T, \mathbf{k}_x = [k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n)]^T$$

$$\mathbf{K}_n = \{k(\mathbf{x}_s, \mathbf{x}_t)\}_{s,t=1}^n, k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Properties of TP

- TP defines a mixture of GPs
- *robustness* of $\mathcal{TP}_\nu(h, k)$ is inversely proportional to ν
- $\lim_{\nu \rightarrow \infty} \mathcal{TP}(h, k) = \mathcal{GP}(h, k)$, i.e. no *robustness*, the same holds for the posterior process
- TP and GP have the same posterior mean, but for $\nu < \infty$ posterior covariances differ in scaling $\frac{\nu + (\mathbf{f}_n - \mathbf{h}_n)^T \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{h}_n)}{\nu + n}$
- mean $\mu_\tau = 1$, variance $\sigma_\tau^2 = \frac{2}{\nu}$
- if τ happens to be small ($\sigma_\tau^2 \nearrow$, i.e. $\nu \searrow$), then the sampling process looks noisy
- the posterior process has less *robustness* than the prior
- in practice we observe noisy versions of $f(\mathbf{x}_i)$, i.e.
 $P(y_i | f(\mathbf{x}_i)) \sim \mathcal{N}(f(\mathbf{x}_i), \sigma^2)$

Properties of TP

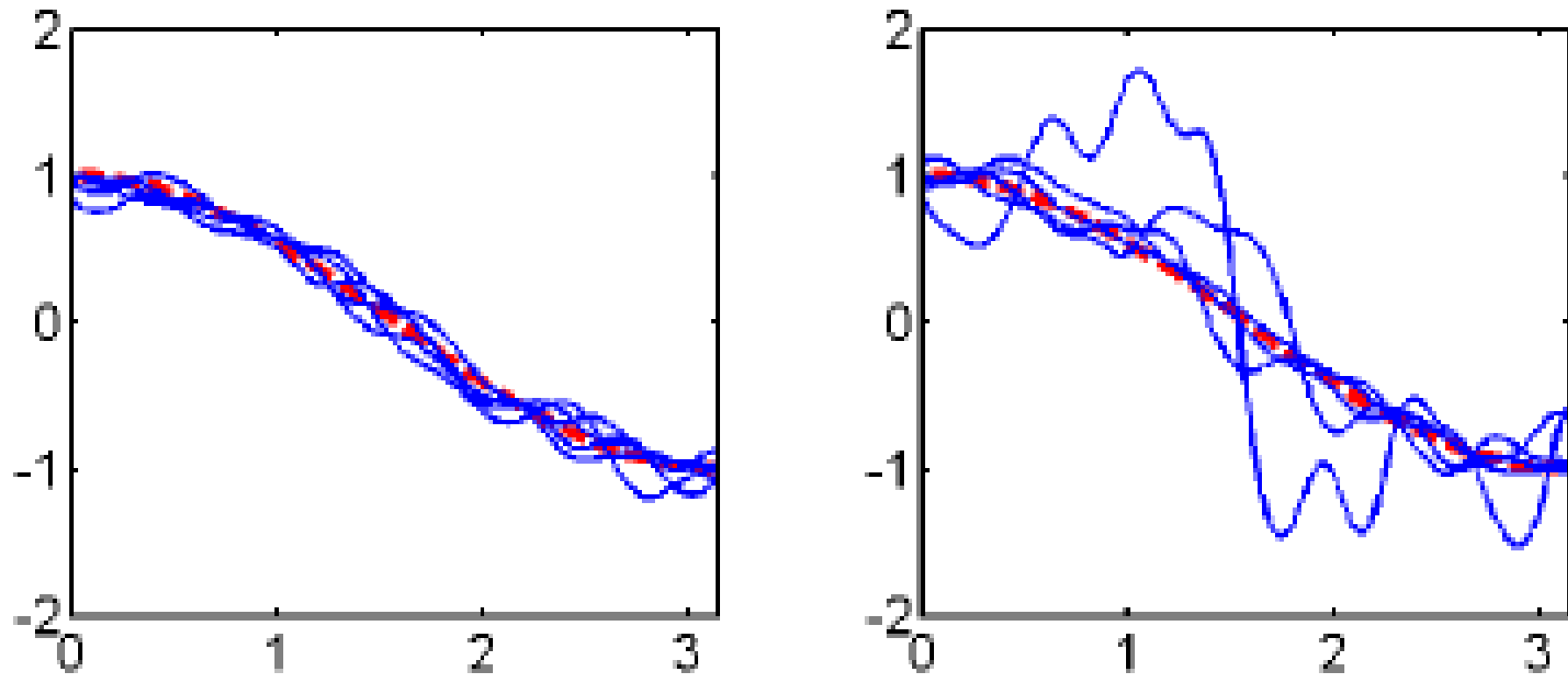


Figure 2. Five samples (blue solid) from $\mathcal{GP}(h, \kappa)$ (left) and $\mathcal{TP}_\nu(h, \kappa)$ (right), with $h(x) = \cos(x)$ (red dashed), $\kappa(x_i, x_j) = 0.01 \exp(-20(x_i - x_j)^2)$ and $\nu = 5$.

MTL with TP

- underlying data

$$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \supset \bigcup_{l=1}^m \mathbf{X}_l$$

- where m is the number of tasks
- noisy observations (labels)

$$\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \supset \bigcup_{l=1}^m \mathbf{Y}_l$$

- assume that all tasks share the same prior $\mathcal{TP}_\nu(h, k)$

MTL with TP

- sampling process

$$\begin{aligned} \mathbf{y}_l | \mathbf{f}_l, \sigma^2 &\sim \mathcal{N}(\mathbf{f}_l, \sigma^2 \mathbf{I}) \\ \mathbf{f}_l | \nu, h, k &\sim t_\nu(\mathbf{h}_l, \mathbf{K}_{l,l}) \end{aligned}$$

- or in terms of GP

$$\begin{aligned} \mathbf{y}_l | \mathbf{f}_l, \sigma^2 &\sim \mathcal{N}(\mathbf{f}_l, \sigma^2 \mathbf{I}) \\ \mathbf{f}_l | \tau_l, h, k &\sim \mathcal{N}(\mathbf{h}_l, \frac{1}{\tau_l} \mathbf{K}_{l,l}) \\ \tau_l | \nu &\sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}) \end{aligned}$$

- where \mathbf{I} is the identity matrix

MTL with TP

- conditional log-likelihood

$$\log P(\mathbf{Y}|\mathbf{X}, \sigma^2) = \sum_l \log \int P_N(\mathbf{y}_l|\mathbf{f}_l, \sigma^2) P_t(\mathbf{f}_l|\nu, \mathbf{h}, \mathbf{K}) d\mathbf{f}_l$$

- where

$$P_t(\mathbf{f}_l|\nu, \mathbf{h}, \mathbf{K}) = \int P_N(\mathbf{f}_l|\mathbf{h}, \frac{1}{\tau_l} \mathbf{K}) P_G(\tau_l|\frac{\nu}{2}, \frac{\nu}{2}) d\tau_l$$

- and P_N , P_t , P_G denote Gaussian, multivariate t , and Gamma
- initialization

$$P(\mathbf{h}, \mathbf{K}) = \mathcal{N}(\mathbf{h}, \mathbf{h}_0, \frac{1}{\pi} \mathbf{K}) \mathcal{IW}(\mathbf{K}; \mathbf{K}_0, \eta)$$

- where $\mathcal{IW}(\mathbf{K}, \mathbf{K}_0, \eta)$ denotes inverse Wishart with base covariance matrix \mathbf{K}_0 and degrees of freedom η

MTL with TP

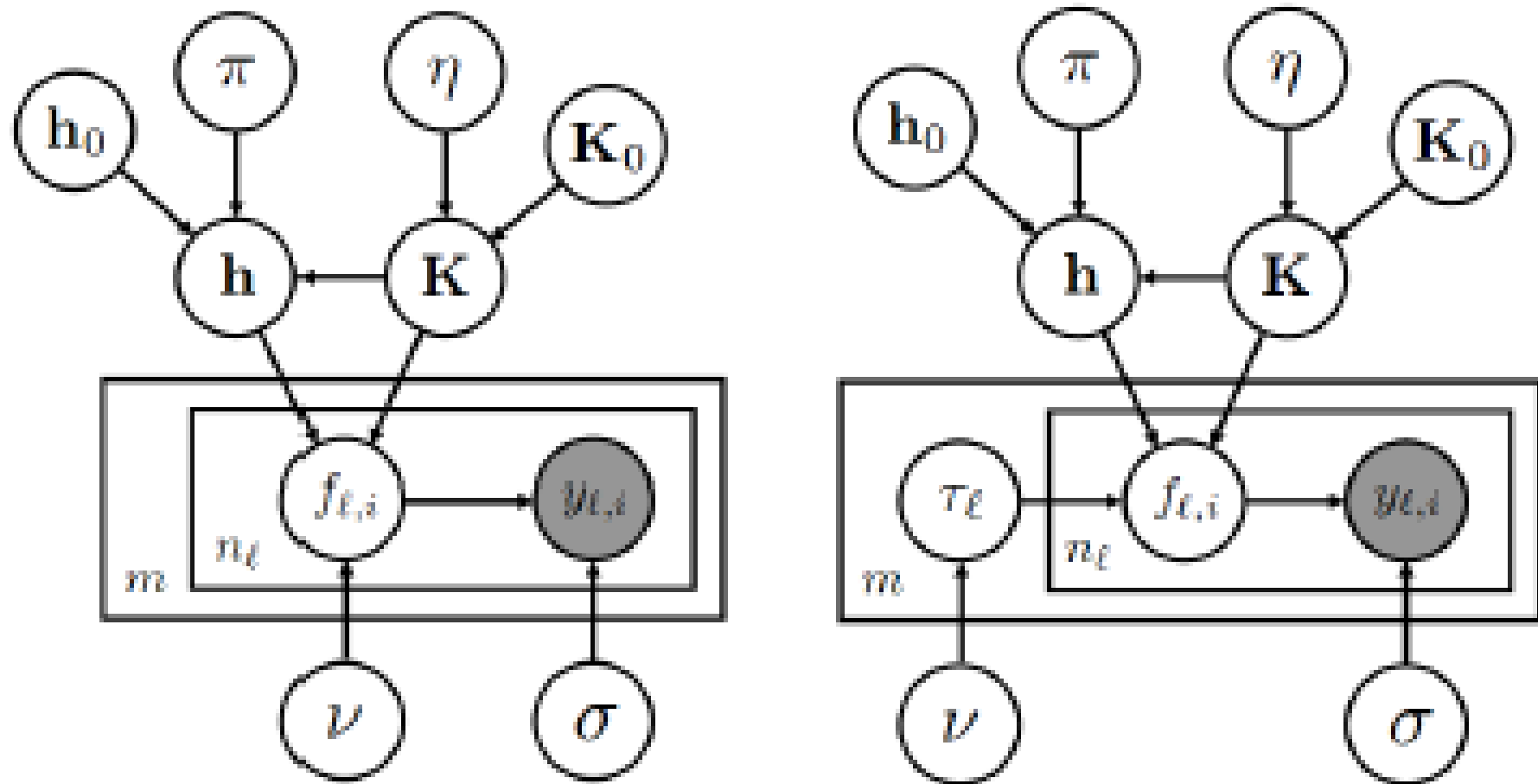


Figure 3. Graphical models for TP multi-task learning (left) and the infinite mixture interpretation (right).

MTL without missing labels

- for input data \mathbf{X} and fully observed labels \mathbf{Y} , the joint posterior is

$$P(\{\mathbf{f}_l, \tau_l\}) = \frac{1}{Z} \prod_l P_N(\mathbf{y}_l | \mathbf{f}_l, \sigma^2) P_N(\mathbf{f}_l | \mathbf{h}, \frac{1}{\tau_l} \mathbf{K}) P_G(\tau_l | \frac{\nu}{2}, \frac{\nu}{2})$$

- with approximation for VB

$$Q(\{\mathbf{f}_l, \tau_l\}) = \prod_l P_N(\mathbf{f}_l | \boldsymbol{\mu}_l, \mathbf{C}_l) P_G(\tau_l | \alpha_l, \beta_l)$$

- where $\boldsymbol{\mu}_l \in \mathbb{R}^{n_l}$, $\mathbf{C}_l \in \mathbb{R}^{n_l \times n_l}$, $\alpha_l > 0$, $\beta_l > 0$ are variational parameters
- and Q is found by minimizing $\int Q \log \frac{Q}{P} d\mathbf{f}_l d\tau_l$

MTL without missing labels

- giving rise to the following update equations

$$\alpha_l = \frac{\nu + n}{2}$$

$$\beta_l = \frac{\nu + (\boldsymbol{\mu}_l - \mathbf{h})^T \mathbf{K}^{-1} (\boldsymbol{\mu}_l - \mathbf{h}) + \text{tr}(\mathbf{K}^{-1} \mathbf{C}_l)}{2}$$

$$\mathbf{C}_l = \left(\frac{1}{\sigma^2} \mathbf{I} + \frac{\alpha_l}{\beta_l} \mathbf{K}^{-1} \right)^{-1}$$

$$\boldsymbol{\mu}_l = \mathbf{C}_l \left(\frac{1}{\sigma^2} \mathbf{y}_l + \frac{\alpha_l}{\beta_l} \mathbf{K}^{-1} \mathbf{h} \right)$$

- where $\text{tr}(\cdot)$ denotes matrix trace

MTL without missing labels

• we use

$$\sigma_{ML}^2 = \arg \max_{\sigma^2} P(\mathbf{y}|\sigma^2)$$

$$\mathbf{h}_{MAP} = \arg \max_{\mathbf{h}} P(\mathbf{h})P(\mathbf{y}|\mathbf{h})$$

$$\mathbf{K}_{MAP} = \arg \max_{\mathbf{K}} P(\mathbf{K})P(\mathbf{y}|\mathbf{K})$$

• giving rise to the following update equations

$$\mathbf{h} = \frac{1}{\pi + \sum_l \frac{\alpha_l}{\beta_l}} (\pi \mathbf{h}_0 + \sum_l \frac{\alpha_l}{\beta_l} \boldsymbol{\mu}_l)$$

$$\mathbf{K} = \frac{1}{\eta + m} \left(\pi (\mathbf{h} - \mathbf{h}_0)(\mathbf{h} - \mathbf{h}_0)^T + \eta \mathbf{K}_0 + \sum_l \frac{\alpha_l}{\beta_l} (\mathbf{C}_l + (\boldsymbol{\mu}_l - \mathbf{h}) \times (\boldsymbol{\mu}_l - \mathbf{h})^T) \right)$$

MTL without missing labels

• and

$$\sigma^2 = \frac{1}{mn} \sum_l \|\mathbf{y}_l - \boldsymbol{\mu}_l\|^2 + \text{tr}(\mathbf{C}_l)$$

• where $\|\cdot\|$ denotes L_2 norm

MTL with missing labels

- update equations, n_l observed labels

$$\boldsymbol{\mu}_l = \mathbf{K}_{n,l} \mathbf{R}_l (\mathbf{y}_l - \mathbf{h}_l) + \mathbf{h}$$

$$\mathbf{C}_l = \frac{\beta_l}{\alpha_l} (\mathbf{K} - \mathbf{K}_{n,l} \mathbf{R}_l \mathbf{K}_{n,l}^T)$$

$$\alpha_l = \frac{\nu + n_l}{2}$$

$$\beta_l = \frac{\nu + (\mathbf{y}_l - \boldsymbol{\mu}_l)^T \mathbf{R}_l \mathbf{K}_{l,l} \mathbf{R}_l (\mathbf{y}_l - \boldsymbol{\mu}_l) + \sigma^2 \text{tr}(\mathbf{R}_l)}{2}$$

$$\mathbf{R}_l = (\mathbf{K}_{l,l} + \sigma^2 \frac{\alpha_l}{\beta_l} \mathbf{I})^{-1}$$

$$\sigma^2 = \frac{1}{m \sum_l n_l} \sum_l (\|\mathbf{y}_l - \boldsymbol{\mu}_l\|^2 + \text{tr}(\mathbf{C}_l))$$

- and $\mathbf{K}_{n,l}$ is a $n \times n_l$ sub-matrix of \mathbf{K} , $n_l < n$

MTL with missing labels

Algorithm 1 Robust Multi-Task Learning

Require: A size- n item set with input features $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Require: m tasks of partial labels $\mathbf{Y} = \{y_1, \dots, y_m\}$, in which task ℓ labels a subset of $n_\ell \leq n$ items.

- 1: Choose prior mean \mathbf{h}_0 (e.g., zero function), base kernel \mathbf{K}_0 (e.g., a Gaussian kernel), degrees of freedom $\nu > 0$, noise level $\sigma^2 > 0$, and hyperparameter $\pi > 0$, $\eta > 0$.
 - 2: Initialize $\mathbf{h} = \mathbf{h}_0$ and $\mathbf{K} = \mathbf{K}_0$.
 - 3: repeat
 - 4: for $\ell = 1, \dots, m$ do
 - 5: Iterate (5) to obtain $\mu_\ell, \mathbf{C}_\ell, \alpha_\ell, \beta_\ell$ for ℓ -th task.
 - 6: end for
 - 7: Update shared parameter $\mathbf{h}, \mathbf{K}, \sigma^2$ via (2), (3), (6).
 - 8: until the improvement is smaller than a threshold.
-

MTL with missing labels

- computational complexity is $\mathcal{O}(m(n\hat{n}^2 + \hat{n}^3))$, similar to that of a GP model
- where $\hat{n} = \max n_l$
- label prediction - given a test point \mathbf{x}^* , what is the probability of its label y_l^*

$$P(y_l^* | \mathcal{D}, \Theta) = \int P(y_l^* | f_l^*, \Theta) P(f_l^* | \mathcal{D}, \Theta) df_l^*$$

- where $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, $f_l^* = f_l(\mathbf{x}^*)$, $P(y_l^* | f_l^*, \Theta) \sim \mathcal{N}(f_l^*, \sigma^2)$, model parameters $\Theta = \{\sigma^2, \nu, \mathbf{h}, \mathbf{K}\}$
- we can compute

$$P(f_l^* | \mathcal{D}, \Theta) = \int P(f_l^* | \mathbf{f}_l, \Theta) P(\mathbf{f}_l | \mathcal{D}, \Theta) d\mathbf{f}_l$$

MTL with missing labels

• from the properties of TP, $P(f_l^* | \mathbf{f}_l, \Theta) \sim t_{\nu+n_l}(\mu_l^*, \sigma_l^{*2})$

• where

$$\begin{aligned}\mu_l^* &= \mathbf{k}^T \mathbf{K}_{l,l}^{-1} (\mathbf{f}_l - \mathbf{h}_l) + h(\mathbf{x}^*) \\ \sigma_l^{*2} &= \frac{\nu + (\mathbf{f}_l - \mathbf{h}_l)^T \mathbf{K}_{l,l}^{-1} (\mathbf{f}_l - \mathbf{h}_l)}{\nu + n_l} (k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T \mathbf{K}_{l,l}^{-1} \mathbf{k})\end{aligned}$$

• $P(f_l^* | \mathcal{D}, \Theta)$ is difficult to compute

• we can write

$$P(y_l^* | \mathcal{D}, \Theta) = \int P(\tau_l | \mathcal{D}, \Theta) P(y_l^* | \tau_l, \mathcal{D}, \Theta) d\tau_l$$

• where $P(y_l^* | \tau_l, \mathcal{D}, \Theta) \sim \mathcal{GP}(\mathbf{h}, \frac{1}{\tau_l} \mathbf{K}) \equiv \mathcal{N}(\hat{\mu}_l^*, \hat{\sigma}_l^{*2})$

MTL with missing labels

• and

$$\hat{\mu}_l^* = \mathbf{k}^T (\mathbf{K}_{l,l} + \sigma^2 \tau_l \mathbf{I})^{-1} (\mathbf{y}_l - \mathbf{h}_l) + h(\mathbf{x}^*)$$

$$\hat{\sigma}_l^{*2} = \frac{1}{\tau_l} (k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T (\mathbf{K}_{l,l} + \sigma^2 \tau_l \mathbf{I})^{-1} \mathbf{k})$$

• and $P(\tau_l | \mathcal{D}, \Theta)$ is the posterior of $\tau_l \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$

Discussion

- in this paper ν is fixed but it can be learned by maximizing the log-likelihood in the M-step

$$\nu = \arg \max_{\nu} \log P(\mathbf{Y}|\mathbf{X})$$

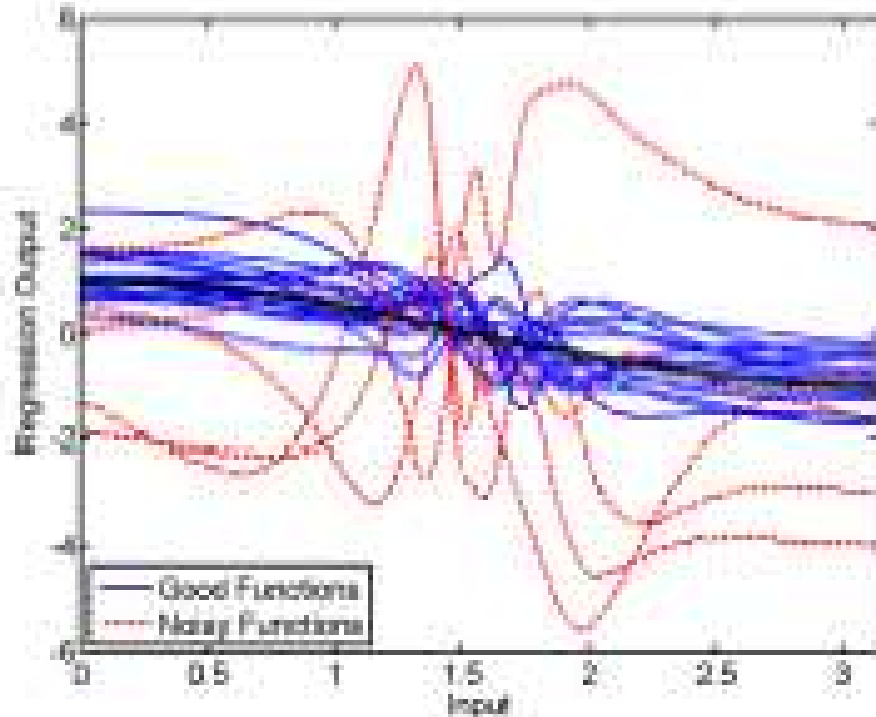
- incorporating linearly transformed TP, i.e.

$$f_l(\mathbf{x}) = \mathbf{w}_l^T \mathbf{x}, \text{ for } \mathbf{w}_l \in \mathbb{R}^d$$

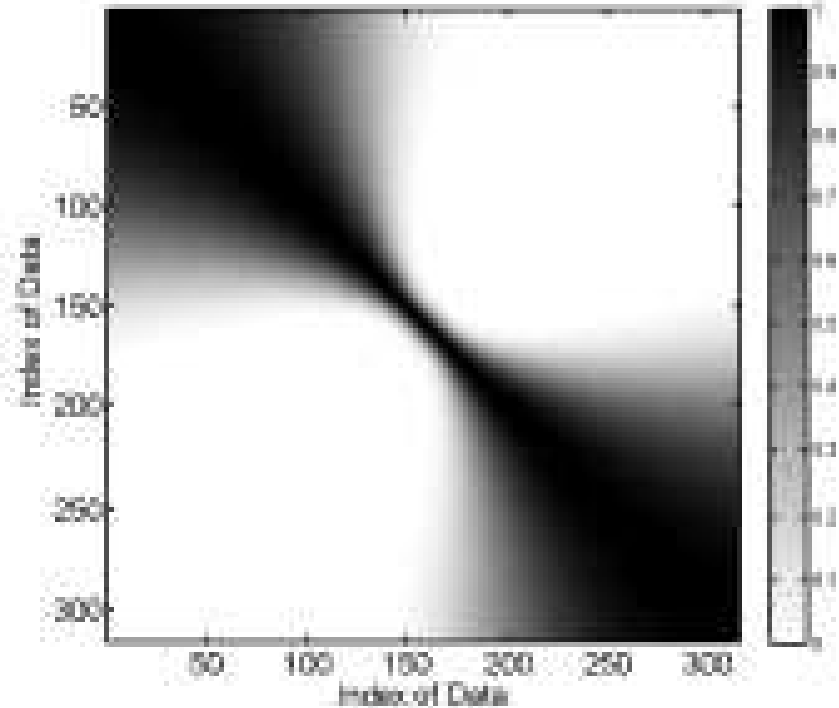
- instead of $P(\mathbf{y}_l|\mathbf{f}_l) \sim \mathcal{N}(\mathbf{f}_l, \mathbf{K})$, use $P(\mathbf{y}_l|\mathbf{f}_l) \sim t_{\nu}(\mathbf{f}_l, \mathbf{K})$, therefore improve robustness for both hidden variables and their labels

Empirical study

- toy MTL problem - learn h and K



(a) 20 samples from a TP

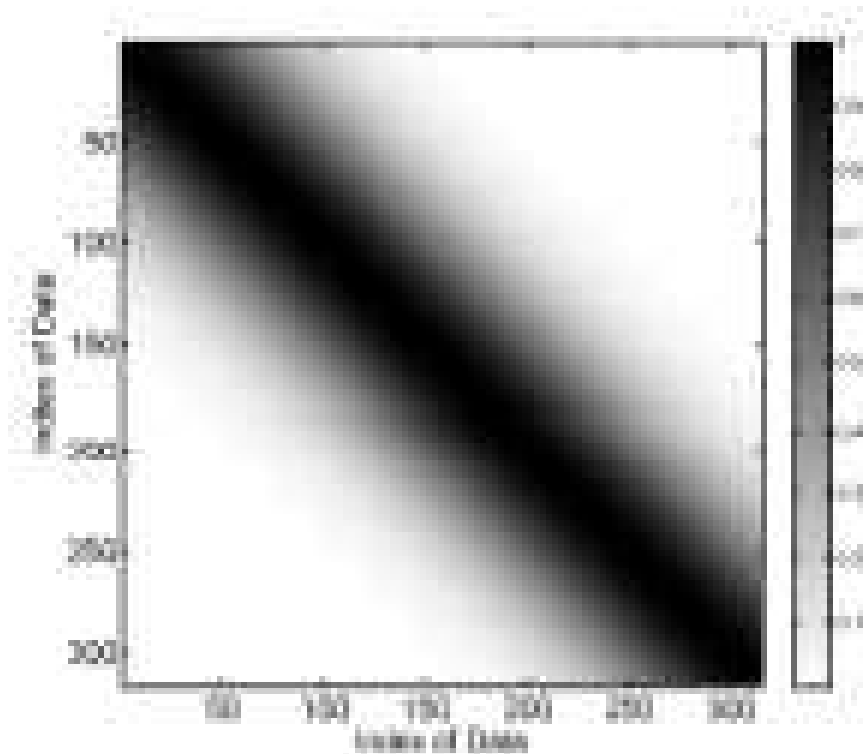


(b) Kernel of the TP

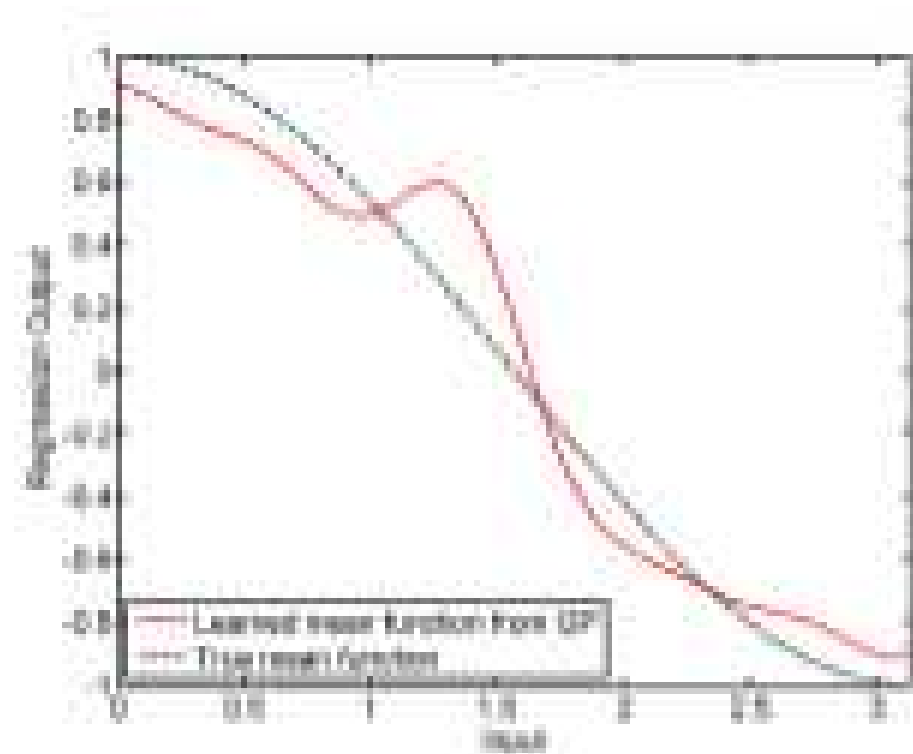
- data set is 350 points (from 0 to π), $\nu = 5$, $h(x) = \cos(x)$, $\sigma^2 = 0.01$
- 15 "good" and 5 "noisy" functions

Empirical study

- toy MTL problem



(c) The base kernel

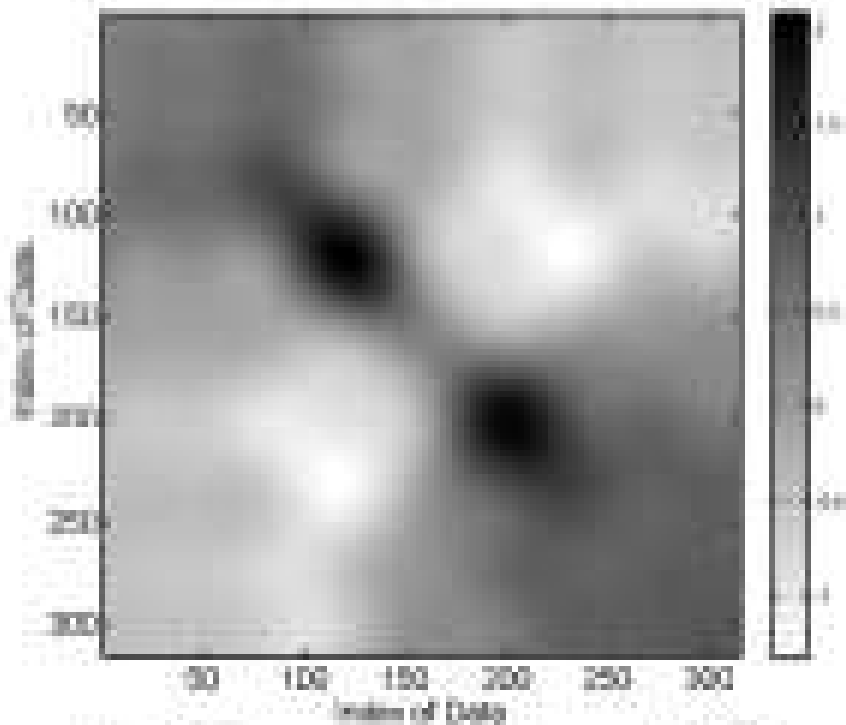


(d) Learned h from GP

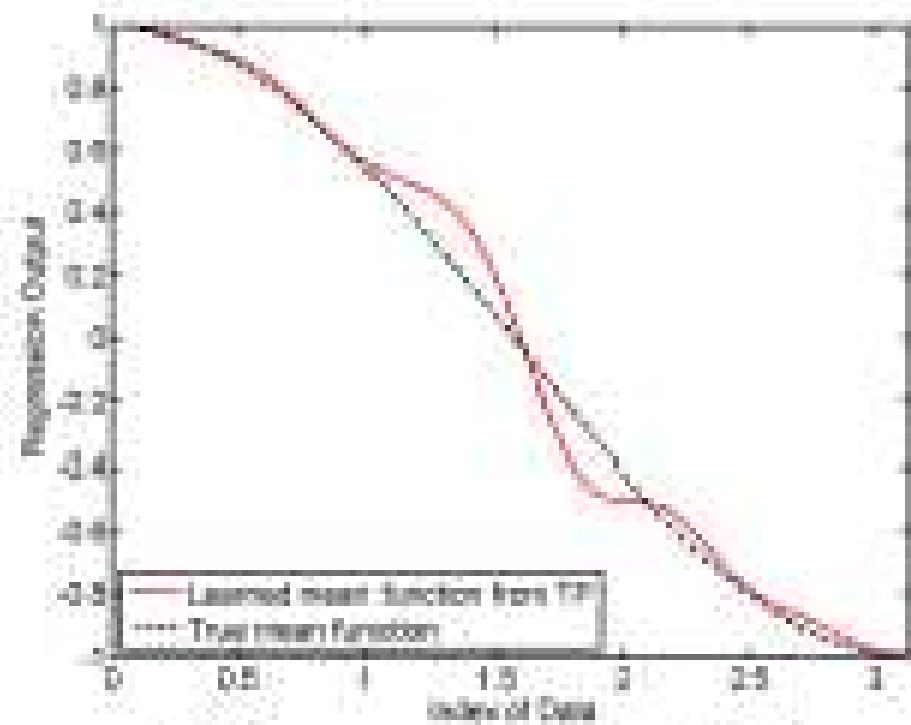
- base kernel is Gaussian
- learned h is noisy

Empirical study

- toy MTL problem



(e) Learned K from GP

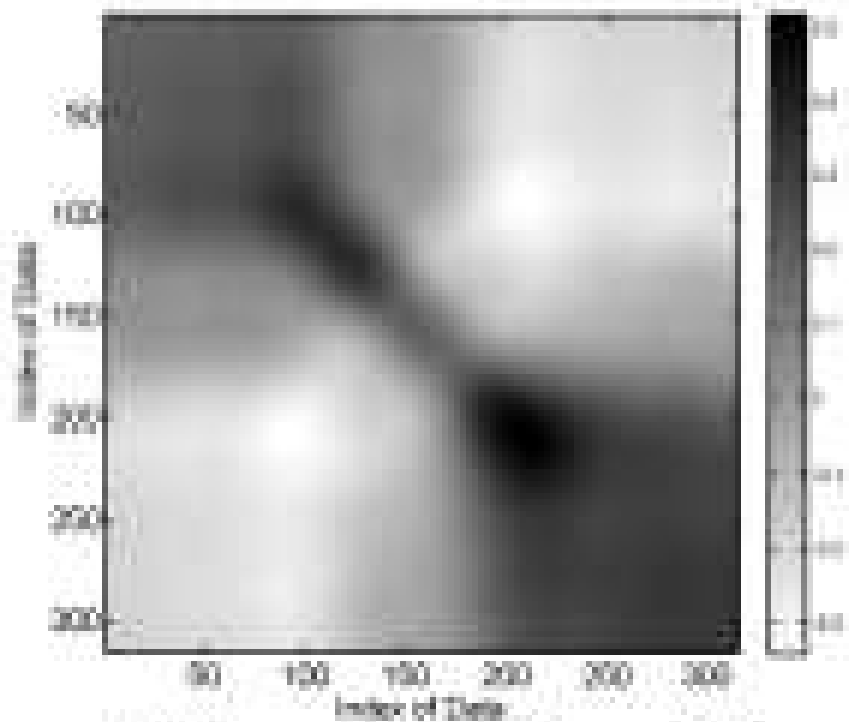


(f) Learned h from TP

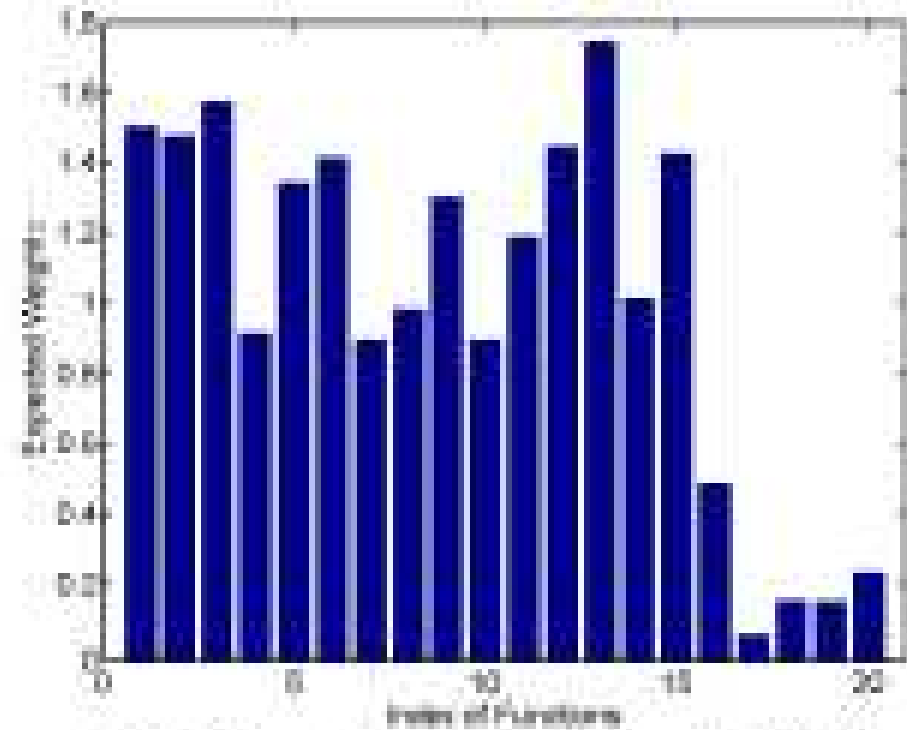
- learned h is smoother

Empirical study

- toy MTL problem



(g) Learned K from TP



(h) Function weights in TP

- last 5 weights are the smallest, corresponding to the noisy functions