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 $t_\nu(\mu, \Sigma)$

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

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

- generation

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

$$x \sim N(\mu, \frac{1}{\tau} \Sigma)$$

$$\frac{1}{\nu + 1} \sum_{i=1}^{\nu+1} x_i \sim t_\nu(\mu, \Sigma)$$
Properties of TP

- properties

\[
\lim_{\nu \to \infty} t_\nu(\mu, \Sigma) \equiv \mathcal{N}(\mu, \Sigma)
\]

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

\[
W x \sim t_\nu(W \mu, W \Sigma W^T)
\]

- let \( x \sim t_\nu(\mu, \Sigma) \), and \( x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \), \( \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \),

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}
\]
be the \([d_1, d - d_1]\) partition, then

\[
x_1 \sim t_\nu(\mu_1, \Sigma_{11})
\]

\[
x_2|x_1 \sim t_{\nu+d_1}(\mu_{x_2|x_1}, \Sigma_{x_2|x_1})
\]
Properties of TP

where

\[ \mu_{x_2|x_1} = \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1) + \mu_2 \]

\[ \Sigma_{x_2|x_1} = \frac{\nu + (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)}{\nu + d_1} (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \]

Student’s distribution \((d = 1, \mu = 0, \Sigma = 1)\)

\[ P(x) = \pi^{-\frac{1}{2}} \nu^\nu \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} (\nu + x^2)^{-\frac{\nu+1}{2}} \]
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 $N(0,1) = t_{+\infty}(0,1)$. 
Properties of TP

- data points
  \[ x_1, \ldots, x_n \in \mathbb{R}^d \]

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

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

- where covariance and mean functions
  \[
  K = \{ k(x_i, x_j) \}_{i,j=1}^n, \quad k : \mathbb{R}^d \to \mathbb{R} \\
  h = \{ h(x_i) \}_{i=1}^n, \quad h : \mathbb{R}^d \to \mathbb{R}
  \]

- marginals are preserved
  \[
  \int \mathcal{GP}(f; h, K) df_1 \ldots df_{i-1} df_{i+1} \ldots df_n \sim \mathcal{GP}(h_i, K_i)
  \]
Properties of TP

- data points
  \[ x_1, \ldots, x_n \in \mathbb{R}^d \]
- random function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) follows a \( \mathcal{T}\mathcal{P}_\nu(h, k) \)
- follows a \( \mathcal{T}\mathcal{P} \) if \( f = \{ f(x_i) \}_{i=1}^n \sim \mathcal{T}\mathcal{P}_\nu(h, K) \)
- using the properties of TP, it can be shown that marginals are preserved, i.e.
  \[
  \int \mathcal{T}\mathcal{P}_\nu(f; h, K) df_1 \ldots df_{i-1} df_{i+1} \ldots df_n \sim \mathcal{T}\mathcal{P}_\nu(h_i, K_i)
  \]
Properties of TP

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

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

- if prior $f \sim TP_{\nu}(h, k)$ and $f_n = [f(x_1), \ldots, f(x_n)]^T$, then

$$f|f_n \sim TP_{\nu+n}(h^*, k^*)$$

where

$$h^*(x) = k_x^T K_n^{-1}(f_n - h_n) + h(x)$$

$$k^*(x_i, x_j) = \frac{\nu + (f_n - h_n)^T K_n^{-1}(f_n - h_n)}{\nu + n} (k_{ij} - k_{xi}^T K_n^{-1} k_{xj})$$

$$h_n = [h(x_1), \ldots, h(x_n)]^T, k_x = [k(x, x_1), \ldots, k(x, x_n)]^T$$

$$K_n = \{k(x_s, x_t)\}_{s,t=1}^n, k_{ij} = k(x_i, x_j)$$
Properties of TP

- TP defines a mixture of GPs
- *robustness* of $\mathcal{T}\mathcal{P}_\nu(h, k)$ is inversely proportional to $\nu$
- $\lim_{\nu \to \infty} \mathcal{T}\mathcal{P}(h, k) = \mathcal{GP}(h, k)$, 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 $\frac{\nu^2(f_n - h_n)^T K_n^{-1} (f_n - h_n)}{\nu + n}$
- mean $\mu_{\tau} = 1$, variance $\sigma_{\tau}^2 = \frac{2}{\nu}$
- if $\tau$ happens to be small ($\sigma_{\tau}^2 \uparrow$, 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 $f(x_i)$, i.e. $P(y_i | f(x_i)) \sim \mathcal{N}(f(x_i), \sigma^2)$
Properties of TP

Figure 2. Five samples (blue solid) from $\mathcal{GP}(h, \kappa)$ (left) and $T \mathcal{P}_\nu(h, \kappa)$ (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

\[ X = \{ x_1, ..., x_n \} \supset \bigcup_{l=1}^{m} X_l \]

- where \( m \) is the number of tasks

- noisy observations (labels)

\[ Y = \{ y_1, ..., y_n \} \supset \bigcup_{l=1}^{m} Y_l \]

- assume that all tasks share the same prior \( TP_{\nu}(h, k) \)
**MTL with TP**

- sampling process
  \[
  \begin{align*}
  y_l | f_l, \sigma^2 & \sim \mathcal{N}(f_l, \sigma^2 I) \\
  f_l | \nu, h, k & \sim \mathcal{t}_\nu(h_l, K_{l,l})
  \end{align*}
  \]

- or in terms of GP
  \[
  \begin{align*}
  y_l | f_l, \sigma^2 & \sim \mathcal{N}(f_l, \sigma^2 I) \\
  f_l | \tau_l, h, k & \sim \mathcal{N}(h_l, \frac{1}{\tau_l} K_{l,l}) \\
  \tau_l | \nu & \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)
  \end{align*}
  \]

- where \( I \) is the identity matrix
**MTL with TP**

- conditional log-likelihood

\[
\log P(Y|X, \sigma^2) = \sum_l \log \int P_N(y_l|f_l, \sigma^2) P_t(f_l|\nu, h, K) d f_l
\]

- where

\[
P_t(f_l|\nu, h, K) = \int P_N(f_l|h, \frac{1}{\tau_l} K) P_G(\tau_l|\frac{\nu}{2}, \frac{\nu}{2}) d \tau_l
\]

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

\[
P(h, K) = \mathcal{N}(h, h_0, \frac{1}{\pi} K) \mathcal{IW}(K; K_0, \eta))
\]

- where \(\mathcal{IW}(K, K_0, \eta)\) denotes inverse Wishart with base covariance matrix \(K_0\) and degrees of freedom \(\eta\)
Figure 3. Graphical models for TP multi-task learning (left) and the infinite mixture interpretation (right).
**MTL without missing labels**

- For input data $X$ and fully observed labels $Y$, the joint posterior is

$$P(\{f_l, \tau_l\}) = \frac{1}{Z} \prod_l P_N(y_l | f_l, \sigma^2) P_N(f_l | h, \frac{1}{\tau_l} K) P_G(\tau_l | \nu, \nu)$$

- With approximation for VB

$$Q(\{f_l, \tau_l\}) = \prod_l P_N(f_l | \mu_l, C_l) P_G(\tau_l | \alpha_l, \beta_l)$$

- Where $\mu_l \in \mathbb{R}^{n_l}$, $C_l \in \mathbb{R}^{n_l \times n_l}$, $\alpha_l > 0$, $\beta_l > 0$ are variational parameters

- And $Q$ is found by minimizing $\int Q \log \frac{Q}{P} d\textbf{f}_l d\tau_l$
MTL without missing labels

giving rise to the following update equations

\[
\alpha_l = \frac{\nu + n}{2}
\]

\[
\beta_l = \frac{\nu + (\mu_l - h)^T K^{-1}(\mu_l - h) + tr(K^{-1}C_l)}{2}
\]

\[
C_l = \left( \frac{1}{\sigma^2} I + \frac{\alpha_l}{\beta_l} K^{-1} \right)^{-1}
\]

\[
\mu_l = C_l \left( \frac{1}{\sigma^2} y_l + \frac{\alpha_l}{\beta_l} K^{-1} h \right)
\]

where \( tr(\cdot) \) denotes matrix trace
MTL without missing labels

we use

\[ \sigma_{ML}^2 = \arg \max_{\sigma^2} P(y|\sigma^2) \]
\[ h_{MAP} = \arg \max_h P(h)P(y|h) \]
\[ K_{MAP} = \arg \max_K P(K)P(y|K) \]

giving rise to the following update equations

\[ h = \frac{1}{\pi + \sum_l \alpha_l \beta_l} (\pi h_0 + \sum_l \frac{\alpha_l}{\beta_l} \mu_l) \]
\[ K = \frac{1}{\eta + m} \left( \pi (h - h_0)(h - h_0)^T + \eta K_0 + \sum_l \frac{\alpha_l}{\beta_l} (C_l + (\mu_l - h) \times (\mu_l - h)^T) \right) \]
MTL without missing labels

and

\[ \sigma^2 = \frac{1}{mn} \sum_l \|y_l - \mu_l\|^2 + tr(C_l) \]

where \( \| \cdot \| \) denotes \( L_2 \) norm
MTL with missing labels

- update equations, $n_l$ observed labels

\[
\begin{align*}
\mu_l &= K_{n,l}R_l(y_l - h_l) + h \\
C_l &= \frac{\beta_l}{\alpha_l} (K - K_{n,l}R_lK_{n,l}^T) \\
\alpha_l &= \frac{\nu + n_l}{2} \\
\beta_l &= \frac{\nu + (y_l - \mu_l)^T R_l K_{l,l} R_l (y_l - \mu_l) + \sigma^2 tr(R_l)}{2} \\
R_l &= (K_{l,l} + \sigma^2 \frac{\alpha_l}{\beta_l} I)^{-1} \\
\sigma^2 &= \frac{1}{m \sum_l n_l} \sum_l (\|y_l - \mu_l\|^2 + tr(C_l))
\end{align*}
\]

- and $K_{n,l}$ is a $n \times n_l$ sub-matrix of $K$, $n_l < n$
Algorithm 1 Robust Multi-Task Learning

Require: A size-$n$ item set with input features $X \in \mathbb{R}^{n \times d}$.
Require: $m$ tasks of partial labels $Y = \{y_1, \ldots, y_m\}$, in which task $\ell$ labels a subset of $n_\ell \leq n$ items.

1: Choose prior mean $h_0$ (e.g., zero function), base kernel $K_0$ (e.g., a Gaussian kernel), degrees of freedom $\nu > 0$, noise level $\sigma^2 > 0$, and hyperparameter $\pi > 0$, $\eta > 0$.
2: Initialize $h = h_0$ and $K = K_0$.
3: repeat
4: for $\ell = 1, \ldots, m$ do
5: Iterate (5) to obtain $\mu_\ell$, $C_\ell$, $\alpha_\ell$, $\beta_\ell$ for $\ell$-th task.
6: end for
7: Update shared parameter $h$, $K$, $\sigma^2$ via (2), (3), (6).
8: until the improvement is smaller than a threshold.
MTL with missing labels

• Computational complexity is $O(m(n\hat{n}^2 + \hat{n}^3))$, similar to that of a GP model.

• Where $\hat{n} = \max n_l$.

• Label prediction - given a test point $x^*$, what is the probability of its label $y_l^*$

\[
P(y_l^* | D, \Theta) = \int P(y_l^* | f_l^*, \Theta) P(f_l^* | D, \Theta) df_l^*
\]

• Where $D = \{X, Y\}$, $f_l^* = f_l(x^*)$, $P(y_l^* | f_l^*, \Theta) \sim \mathcal{N}(f_l^*, \sigma^2)$, model parameters $\Theta = \{\sigma^2, \nu, h, K\}$.

• We can compute

\[
P(f_l^* | D, \Theta) = \int P(f_l^* | f_l, \Theta) P(f_l | D, \Theta) df_l
\]
**MTL with missing labels**

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

where

\[
\begin{align*}
\mu_l^* &= k^T K_{l,l}^{-1}(f_l - h_l) + h(x^*) \\
\sigma_l^{*2} &= \nu + (f_l - h_l)^T K_{l,l}^{-1}(f_l - h_l) \frac{(k(x^*, x^*) - k^T K_{l,l}^{-1}k)}{\nu + n_l}
\end{align*}
\]

- \( P(f_l^* | \mathcal{D}, \Theta) \) 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 \( P(y_l^* | \tau_l, \mathcal{D}, \Theta) \sim GP(h, \frac{1}{\tau_l} K) \equiv \mathcal{N}(\hat{\mu}_l^*, \hat{\sigma}_l^{*2}) \)
MTL with missing labels

and

\[
\hat{\mu}_l^* = k^T(K_{l,l} + \sigma^2\tau_l I)^{-1}(y_l - h_l) + h(x^*)
\]

\[
\hat{\sigma}_l^{*2} = \frac{1}{\tau_l}(k(x^*, x^*) - k^T(K_{l,l} + \sigma^2\tau_l I)^{-1}k)
\]

and \( P(\tau_l|D, \Theta) \) is the posterior of \( \tau_l \sim Gamma(\nu_2, \nu_2) \)
Discussion

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

\[ \nu = \arg \max_{\nu} \log P(Y|X) \]

- incorporating linearly transformed TP, i.e.

\[ f_l(x) = w_l^T x, \text{ for } w_l \in \mathbb{R}^d \]

- instead of $P(y_l|f_l) \sim \mathcal{N}(f_l, K)$, use $P(y_l|f_l) \sim t_\nu(f_l, K)$, therefore improve robustness for both hidden variables and their labels
Empirical study

- toy MTL problem - learn $h$ and $K$

- data set is 350 points (from 0 to $\pi$), $\nu = 5$, $h(x) = \cos(x)$, $\sigma^2 = 0.01$

- 15 "good" and 5 "noisy" functions
Empirical study

- toy MTL problem
- base kernel is Gaussian
- learned $h$ is noisy
Empirical study

- toy MTL problem

- learned $h$ is smoother
Empirical study

- toy MTL problem

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