Gaussian Process Regression with Student-t Likelihood

Jarno Vanhatalo, Pasi Jylanki and Aki Vehtari
NIPS-2009
Presented by Minhua Chen

July 30, 2010
Main Idea
- Motivation
- Model Inference

GP Regression Model with Student-t Likelihood
- The Model
- The Inference Problem

Bayesian Inference by Scale Mixture Representation
- Scale Mixture Representation for Student-t Distribution
- Bayesian Inference for the Augmented Model

Approximate Inference by Laplacian Approximation
- Introduction to Laplacian Approximation
- Laplacian Approximation for the GP Regression Model
- MAP Estimation for the Hyperparameters

Experimental Results

Conclusion
Figure 1: An example of regression with outliers by Neal [6]. On the left Gaussian and on the right the Student-t observation model. The real function is plotted with black line.
All the following inference methods are covered in this paper:

- **Approximate inference:** Variational Bayes (VB), Expectation Maximization (EM), Laplacian Approximation.
- **Exact inference:** Markov Chain Monte Carlo (MCMC).

The author derived the Laplacian Approximation method for the GP regression Student-t Likelihood Model and compared with other inference methods.
The Model

Given data \( X = \{x_i\}_{i=1}^n \) and the latent function \( f = \{f(x_i)\}_{i=1}^n \), the GP regression model can be expressed as \( y = f + \epsilon \) with GP prior

\[
p(f|\theta) = \mathcal{N}(f; 0, K_{ff})
\]

where \((K_{ff})_{ij} = K(x_i, x_j) = \sigma_{se}^2 \exp(-\sum_{d=1}^D (x_{id} - x_{jd})^2/l_d^2)\) is the kernel function.

The Student-t likelihood (i.e., the noise model) is

\[
p(y|f, \theta) = \prod_{i=1}^n t(y_i; f_i, \sigma, \nu)
\]

\[
= \prod_{i=1}^n \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma}} \left(1 + \frac{(y_i - f_i)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}
\]

where \( \theta = \{\sigma_{se}^2, l_{1:D}, \sigma, \nu\} \) is the set of hyperparameters.
The Inference Problem

The full likelihood of the model can be expressed as

\[ p(f, y|\theta) = p(f|\theta)p(y|f, \theta) = \mathcal{N}(f; 0, K_{ff}) \prod_{i=1}^{n} t(y_i; f_i, \sigma, \nu) \]

Then the posterior inference for \( f \) can be expressed as

\[ p(f|y, \theta) = \frac{p(f|\theta)p(y|f, \theta)}{p(y|\theta)} \]

This inference is not straightforward since Gaussian prior and Student-t likelihood are not conjugate.

- Use scale mixture representation of Student-t to make the model conjugate, then apply VB/EM/MCMC.
- Use Laplacian Approximation directly.

Once the posterior \( p(f|y, \theta) \) is inferred, we can do prediction for new data:

\[ p(y_\star|x_\star, y, \theta) = \int \int p(f_\star|f, x_\star, \theta)p(f|y, \theta)p(y_\star|f_\star, \theta)dfdf_\star \]
Scale Mixture Representation for Student-t Distribution

Scale mixture representation for Student-t distribution can be applied as follows:

\[ p(y|f, \theta) = \prod_{i=1}^{n} t(y_i; f_i, \sigma, \nu) \]
\[ = \prod_{i=1}^{n} \int \mathcal{N}(y_i; f_i, \tau_i^{-1}) \mathcal{Ga}(\tau_i; \nu/2, \nu\sigma^2/2) d\tau_i \]

Then the augmented model is

\[ p(f, y, \tau|\theta) = p(f|\theta)p(y|f, \tau, \theta)p(\tau|\theta) \]
\[ = \mathcal{N}(f; 0, K_{ff})\mathcal{N}(y; f, \text{diag}^{-1}(\tau)) \prod_{i=1}^{n} \mathcal{Ga}(\tau_i; \nu/2, \nu\sigma^2/2) \]

By integrating out \( \tau \), the above augmented model reduces to the original model. The advantage we gain is that now the likelihood is conjugate with the prior of \( f \).
VB assumes that the joint posterior can be factorized as products of marginal posteriors:

\[
p(f, \tau|y, \theta) = \frac{p(f|\theta)p(y|f, \tau, \theta)p(\tau|\theta)}{p(y|\theta)} = q(f)q(\tau)
\]

1) \( q(f) = \mathcal{N}(f; m, A) \) with \( A = (K_{ff} + \text{diag}(\langle \tau \rangle))^{-1} \),
\( m = (K_{ff} + \text{diag}(\langle \tau \rangle))^{-1}(\text{diag}(\langle \tau \rangle)y) \).

2) \( q(\tau) = \prod_{i=1}^{n} \text{Ga}(\tau_i; (\nu + 1)/2, (\nu \sigma^2 + \langle (y_i - f_i)^2 \rangle)/2) \)
with \( \langle \tau_i \rangle = \frac{\nu+1}{\nu \sigma^2 + \langle (y_i - f_i)^2 \rangle} \).

Notice that \( \langle (y_i - f_i)^2 \rangle = (y_i - m_i)^2 + A_{ii} \).
EM is a special case of VB, which gives a point estimate on $f$:

$$p(f, \tau | y, \theta) = \frac{p(f | \theta)p(y | f, \tau, \theta)p(\tau | \theta)}{p(y | \theta)} = \delta(f - \hat{f})\tilde{q}(\tau)$$

1) $\hat{f} = \arg \max_f \int \log p(p(f, \tau, y | \theta)) \cdot \tilde{q}(\tau) d\tau$

$$= (K_{ff} + \text{diag}(<\tilde{\tau}>))^{-1}(\text{diag}(<\tilde{\tau}>)y).$$

2) $\tilde{q}(\tau) = \prod_{i=1}^n \text{Ga}(\tau_i; (\nu + 1)/2, (\nu \sigma^2 + (y_i - \hat{f}_i)^2)/2)$ with $<\tilde{\tau}_i> = \frac{\nu + 1}{\nu \sigma^2 + (y_i - \hat{f}_i)^2}$. 

Gibbs sampling:

1) Sample $f$ from $p(f|\tau, y, \theta)$
2) Sample $\tau$ from $p(\tau|f, y, \theta)$

Iteration between the above two processes.
Introduction to Laplacian Approximation

Given an arbitrary density function \( p(x) \), the Taylor expansion of \( \log p(x) \) is

\[
\log p(x) = \log p(x_0) + (x - x_0)^\top \nabla \log p(x) \big|_{x=x_0} \\
+ \frac{1}{2} (x - x_0)^\top (\nabla \nabla \log p(x) \big|_{x=x_0})(x - x_0) + o(\|x - x_0\|^2)
\]

Specifically, choose \( x_0 = x_* \) as a stationary point, that is, \( \nabla \log p(x) \big|_{x=x_*} = 0 \), then

\[
p(x) \hat{=} c \exp \left( -\frac{1}{2} (x - x_*)^\top \Sigma^{-1} (x - x_*) \right) = \mathcal{N}(x; x_*, \Sigma)
\]

where \( c \) is a normalizing constant and

\[
\Sigma^{-1} = -\nabla \nabla \log p(x) \big|_{x=x_*}.
\]
Laplacian Approximation for the GP Regression Model

Specifically for this GP regression model, the posterior \( p(\mathbf{f}|\mathbf{y}, \theta) \) is approximated by a normal distribution as

\[
p(\mathbf{f}|\mathbf{y}, \theta) = \mathcal{N}(\mathbf{f}; \hat{\mathbf{f}}, \Sigma)
\]

1) \( \hat{f} = \arg \max_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{y}, \theta) = \arg \max_{\mathbf{f}} (\log p(\mathbf{y}|\mathbf{f}, \theta) + \log p(\mathbf{f}|\theta)) \).

2) \( \Sigma^{-1} = -\nabla \nabla \log p(\mathbf{f}|\mathbf{y}, \theta) \bigg|_{\mathbf{f}=\hat{\mathbf{f}}} = K_{ff}^{-1} + \mathbf{W} \)

with \( W_{ii} = -(\nu + 1) \frac{(y_i - \hat{f}_i)^2 - \nu \sigma^2}{((y_i - \hat{f}_i)^2 + \nu \sigma^2)^2} \) and \( W_{ij} = 0 \) if \( i \neq j \).

\( \hat{f} \) is usually done by numerical optimization methods, but in this paper it is done by using EM (iterative) estimation.
MAP Estimation for the Hyperparameters in the Laplacian Approximation

The MAP estimation for \( \theta \) is done via

\[
\theta^*_\theta = \arg \max_{\theta} \log p(\theta, y) = \arg \max_{\theta} (\log p(y|\theta) + \log p(\theta))
\]

The marginal term can be approximated as

\[
p(y|\theta) = \frac{p(f|\theta)p(y|f, \theta)}{p(f|y, \theta)} = \frac{\mathcal{N}(f; 0, K_{ff})p(y|f, \theta)}{\mathcal{N}(f; \hat{f}, \Sigma)} \quad (\forall f)
\]

Specifically set \( f = \hat{f} \), then

\[
\log p(y|\theta) = \log p(y|\hat{f}, \theta) - \frac{1}{2} \hat{f}^T K_{ff}^{-1} \hat{f} - \frac{1}{2} \log |K_{ff}| - \frac{1}{2} \log |K_{ff}^{-1} + W|
\]

Cholesky decomposition and rank one update are used to speed up the above computation.
**Results**

<table>
<thead>
<tr>
<th></th>
<th>Neal</th>
<th>Friedman</th>
<th>Housing</th>
<th>Concrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.393</td>
<td>0.324</td>
<td>0.324</td>
<td>0.230</td>
</tr>
<tr>
<td>T-lapl</td>
<td>0.028</td>
<td>0.220</td>
<td>0.289</td>
<td>0.231</td>
</tr>
<tr>
<td>T-vb</td>
<td>0.029</td>
<td>0.220</td>
<td>0.294</td>
<td>0.212</td>
</tr>
<tr>
<td>T-mcmc</td>
<td>0.055</td>
<td>0.253</td>
<td>0.287</td>
<td>0.197</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Neal</th>
<th>Friedman</th>
<th>Housing</th>
<th>Concrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.254</td>
<td>0.227</td>
<td>1.249</td>
<td>-0.0642</td>
</tr>
<tr>
<td>T-lapl</td>
<td>-2.181</td>
<td>-0.16</td>
<td>0.080</td>
<td>-0.116</td>
</tr>
<tr>
<td>T-vb</td>
<td>-2.228</td>
<td>-0.049</td>
<td>0.091</td>
<td>-0.132</td>
</tr>
<tr>
<td>T-mcmc</td>
<td>-1.907</td>
<td>-0.106</td>
<td>0.029</td>
<td>-0.241</td>
</tr>
</tbody>
</table>
• Laplacian Approximation method is derived for the GP regression Student-t likelihood model. Performance is competitive to other inference methods (e.g., VB).

• Student-t likelihood model performs significantly better than Gaussian likelihood model on four real data sets.