Sparse Variational Inference for Generalized Gaussian Process Models

Abstract

Gaussian processes (GP) provide an attractive machine learning model due to their non-parametric form, their flexibility to capture many types of observation data, and their generic inference procedures. Sparse GP inference algorithms address the cubic complexity of GPs by focusing on a small set of pseudo-examples. To date, such approaches have focused on the simple case of Gaussian likelihoods for the observations, or have couched sparse solutions within a latent variable GP model. This paper develops a direct sparse solution for GPs under general likelihoods by providing a new characterization of the gradients required for inference in terms of individual observation likelihood terms. In addition, we propose a simple new approach for optimizing the sparse variational approximation using a fixed point computation. We demonstrate experimentally that the fixed point operator acts as a contraction in many cases and therefore leads to fast convergence. An experimental evaluation for count regression, classification, and ordinal regression illustrates the generality and advantages of the new approach.

1. Introduction

Gaussian process (GP) models are a flexible class of non-parametric Bayesian methods that have been used in a variety of supervised machine learning tasks. A GP induces a normally distributed set of latent values which in turn generate observation data. GPs have been successfully applied to many observation types including regression with Gaussian likelihood, binary classification (Rasmussen & Williams, 2006), robust regression (Vanhatalo et al., 2009), ordinal regression (Chu et al., 2006), quantile regression (Boukouvalas et al., 2012), and relational learning (Sindhwani et al., 2007). In addition, a generalized GPs formulation (Shang & Chan, 2013), using observation data from a generic exponential family distribution, enables a non-parametric extension of generalized linear models. The main difficulty in applying GP models is the complexity which is cubic in the number of observations N. In addition, non-Gaussian likelihoods require some approximation of the posterior as the GP prior is non-conjugate.

In recent years, a number of approaches have been developed to address this issue. The variational Gaussian approximation has received renewed attention (Opper & Archambeau, 2009; Lázaro-gredilla & Titsias, 2011; Khan et al., 2012; Challis & Barber, 2013; Khan et al., 2013) with reformulations and algorithms that reduce the number of estimated parameters, and improve convergence of the estimates. This provides significant improvements but retains the overall $\mathrm{O}(N^3)$ complexity of inference. Several recent papers develop novel algorithmic frameworks, including online stochastic solutions for variational inference via data sub-sampling (Hoffman et al., 2013) and distribution sampling (Titsias & Lázaro-gredilla, 2014), and parallel approaches that distribute the inference across multiple cores (Gal et al., 2014).

An alternative known as sparse solutions (see e.g. (Seeger et al., 2003; Keerthi & Chu, 2006; Quiñonero Candela et al., 2005; Snelson & Ghahramani, 2006; Titsias, 2009)) uses an additional approximation to reduce complexity. In particular, Titsias (2009) formulated this approximation as an optimization of a variational bound on the marginal likelihood. In these methods, an active set of M real or "pseudo" samples, where $M \ll N$, is used as an approximate sufficient statistic for inference and prediction, reducing training complexity to $\mathrm{O}(M^2 N)$. Despite significant interest, and some work on specific models (Naish-Guzman & Holden, 2007; Vanhatalo & Vehtari, 2007), there is no general direct formulation of the sparse GPs model for general likelihoods. Although some of the work on latent variable GP models (Titsias & Lawrence, 2010; Khan et al., 2013; Challis & Barber, 2013; Gal et al., 2014) can capture the sparse formulation as a special case, the algorithms have not been applied in this way and it is not clear that they are best suited for this purpose.

In this paper, we extend the formulation of Titsias (2009) to handle arbitrary likelihoods. Our formulation and solution are generic in that they depend directly on properties of the observation likelihood function of individual observations. In particular we show that the gradients needed to

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

optimize the sparse solution can be calculated from derivative information of individual observation likelihoods. This allows for a generic solution that also applies in the generalized GP framework. The sparse model can be optimized by adapting the ideas of Challis & Barber (2013) to the sparse formulation but this approach can be slow in some cases. We propose a new method for solving the optimization problem using fixed point updates on the variational covariance. Although we are not able to analyze it theoretically, we demonstrate experimentally that the fixed point operator acts as a contraction in many cases and therefore leads to fast convergence. Finally, an experimental evaluation on count regression, classification, and ordinal regression compares these algorithms to several baselines as well as to the application of a dual optimization algorithm for latent variable models (Khan et al., 2013), and illustrates the generality and advantages of the new approach.

2. Preliminaries: Sparse Variational Approximation

We briefly review Gaussian process (GP) models and describe our notation. A more thorough introduction can be found in (Rasmussen & Williams, 2006). A GP is specified by a mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$ and is used to provide a prior distribution over functions. For any finite set of N inputs $\mathcal{X}$, the function values at $\mathcal{X}$, denoted $\boldsymbol{f}_{\mathcal{X}}$, are distributed as multivariate Gaussian with mean $\boldsymbol{m}_{\mathcal{X}}$ and covariance $K_N = k(\mathcal{X},\mathcal{X})$. The function values are typically assumed to be latent and the observations are distributed as $\prod_{i=1}^N p(y_i|f(\boldsymbol{x}_i))$, where $p(\cdot|\cdot)$ is the likelihood of the i'th observation y_i given the latent function evaluated at input $\boldsymbol{x}_i$. We let $\boldsymbol{y}$ stand for the vector of observations at inputs $\mathcal{X}$. In our notation, subscripts M or $\mathcal{U}$ refer to evaluation on the inducing set (also referred to as active or pseudo set) while N or $\mathcal{X}$ refer to evaluation on the training set, $K_{\cdot M} \equiv k(\cdot,\mathcal{U})$ and $K_{M\cdot} = K_{\cdot M}^T$. $\mathbb{S}_M^{++}$ refers to the space of symmetric positive definite matrices. For vectors, $\preceq$ denotes element-wise inequality and for matrices A, B denote $A \preceq B$ to mean that for all vectors c, we have $c^T A c \leq c^T B c$.

Given the observations $\boldsymbol{y}$ our goal is to calculate the posterior distribution over $\boldsymbol{f}_{\mathcal{X}}$ (i.e., inference) as well as make predictions $p(y^*|\boldsymbol{x}^*,\boldsymbol{y})$ at a new input $\boldsymbol{x}^*$. Calculating the posterior requires cubic run time in the number of data points and is not feasible for large datasets. Sparse GP methods approximate this by reducing the number of "relevant points" to $M \ll N$. The standard approach first augments the data with M pseudo inputs $\mathcal{U} = \{u_l \in \mathbb{R}^D | 1 \leq l \leq M \ll N\}$ and assumes for prediction that $p(y^*|\boldsymbol{x}^*,\mathbf{y}) = p(y^*|\boldsymbol{x}^*,\boldsymbol{f}_{\mathcal{U}})$. Titsias (2009) formulated this task as an optimization where the set $\mathcal{U}$ and its values $\boldsymbol{f}_{\mathcal{U}}$ are chosen to maximize a variational lower bound on the marginal likelihood of the data. In this paper we extend this formulation to handle general likelihood functions.

2.1. Variational Lower Bound

Following the model of (Titsias, 2009), the posterior $p(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}}|\boldsymbol{y})$ is approximated by the variational distribution

$$q(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}}) = p(\boldsymbol{f}_{\mathcal{X}}|\boldsymbol{f}_{\mathcal{U}})\phi(\boldsymbol{f}_{\mathcal{U}}) \tag{1}$$

where ϕ is a multivariate Gaussian distribution with (unknown) mean $\boldsymbol{m}$ and covariance V.

The approximate posterior $q(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}})$ is found by minimizing the Kullback-Liebler (KL) divergence between $q(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}})$ and the full posterior $p(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}}|\boldsymbol{y})$ which is equivalent to maximizing the following lower bound on the log marginal likelihood (Titsias, 2009; Khan et al., 2012; Gal et al., 2014):

$$\begin{aligned}\log p(\boldsymbol{y}) \geq \sum_{i=1}^N \mathbb{E}_{q(f(\boldsymbol{x}_i))}\left[\log p(y_i|f(\boldsymbol{x}_i))\right] \\ -\mathrm{KL}(\phi(\boldsymbol{f}_{\mathcal{U}})||p(\boldsymbol{f}_{\mathcal{U}}))\end{aligned} \tag{2}$$

where $q(f(\boldsymbol{x}_i))$ denotes the marginal distribution of $f(\boldsymbol{x}_i)$ with respect to the approximate posterior $q(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}})$. We refer to the RHS of Equation (2) as the variational lower bound (VLB).

Since $q(\boldsymbol{f}_{\mathcal{X}},\boldsymbol{f}_{\mathcal{U}})$ is jointly Gaussian, the marginal distribution is given by a univariate Gaussian with mean $m_q(\boldsymbol{x}_i)$ and variance $v_q(\boldsymbol{x}_i)$ where

$$m_q(\boldsymbol{x}) = m(\boldsymbol{x}) + K_{xM}K_M^{-1}(\boldsymbol{m} - \boldsymbol{m}_{\mathcal{U}}) \tag{3a}$$

$$v_q(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{x}) + K_{xM}K_M^{-1}(V - K_M)K_M^{-1}K_{Mx} \tag{3b}$$

3. Inference for the Sparse Model

By first-order optimality, $(\boldsymbol{m}^\star, V^\star)$ is found via the conditions $\frac{\partial \mathrm{VLB}}{\partial \boldsymbol{m}}|_{\boldsymbol{m}=\boldsymbol{m}^\star} = 0$ and $\frac{\partial \mathrm{VLB}}{\partial V}|_{V=V^\star} = 0$. We start by showing how the derivatives can be calculated.

3.1. Characterization of the Variational Solution

The first term of the VLB represents the goodness of fit for the model. As in (Challis & Barber, 2013) we use a change of variables to simplify the analysis. In particular, by making the change of variables $f_i = z_i\sqrt{v_{q_i}} + m_{q_i}$, we can express the expectation with respect to the approximate marginal q_i as (we drop the argument $\boldsymbol{x}_i$ for notational convenience): $\mathbb{E}_{q_i(f_i)}\left[\log p(y_i|f_i)\right] =$

$$\frac{1}{\sqrt{2\pi}}\int \log p(y_i|z_i\sqrt{v_{q_i}} + m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \tag{4}$$

We can now develop the derivatives of the observation term with respect to the variational parameters by taking derivatives of (4). Starting with m_{q_i} we get:

$$\begin{aligned}
&\frac{\partial}{\partial m_{q_i}}\left[\frac{1}{\sqrt{2\pi}}\int \log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i\right] \\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\partial}{\partial m_{q_i}}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}{\partial m_{q_i}}\ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i \\
&= \frac{1}{\sqrt{2\pi}}\int \ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i \\
&= \mathbb{E}_{\mathcal{N}(z_i|0,1)}\left[\ell_i(z_i)\right] \qquad (5)
\end{aligned}$$

where $\ell_i(z_i) \equiv \frac{\partial}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})$. Since $\frac{\partial m_{q_i}}{\partial \boldsymbol{m}} = K_M^{-1}K_{Mi}$, we get that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{m}}\mathbb{E}_{q_i}\left[\log p(y_i|f_i)\right] &= K_M^{-1}K_{Mi} \qquad (6)\\
\mathbb{E}_{\mathcal{N}(z_i|0,1)}&\left[\frac{\partial}{\partial f_i}\log p(y_i|f_i)\Big|_{f_i=z_i\sqrt{v_{q_i}}+m_{q_i}}\right]
\end{aligned}$$

Similarly for v_{q_i}:

$$\begin{aligned}
&\frac{\partial}{\partial(\sqrt{v_{q_i}})}\left[\frac{1}{\sqrt{2\pi}}\int \log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i\right] \\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}{\partial(\sqrt{v_{q_i}})}\ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i \\
&= \frac{1}{\sqrt{2\pi}}\int z_i\ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i \\
&= -\frac{1}{\sqrt{2\pi}}\ell_i(z_i)e^{-\frac{1}{2}z_i^2}]_{-\infty}^{+\infty} + \int \frac{\partial}{\partial z_i}\ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i
\end{aligned} \qquad (7)$$

where in the last step we have used integration in parts. If $\ell_i(z_i) = \mathrm{o}(e^{\frac{1}{2}z_i^2})$ as $z_i \to \pm\infty$, then (7) reduces to

$$\begin{aligned}
&\int \frac{\partial}{\partial z_i}\ell_i(z_i)e^{-\frac{1}{2}z_i^2}dz_i \\
&= \int \frac{\partial}{\partial z_i}\frac{\partial}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \\
&= \int \frac{\partial}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}\frac{\partial}{\partial z_i}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \\
&= \int \frac{\partial}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}\frac{\partial}{\partial z_i}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \\
&= \int \frac{\partial}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})}\frac{\sqrt{v_{q_i}}}{p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})}e^{-\frac{1}{2}z_i^2}dz_i \\
&= \sqrt{v_{q_i}}\int \frac{\partial^2}{\partial(z_i\sqrt{v_{q_i}}+m_{q_i})^2}\log p(y_i|z_i\sqrt{v_{q_i}}+m_{q_i})e^{-\frac{1}{2}z_i^2}dz_i \\
&= \sqrt{v_{q_i}}\,\mathbb{E}_{\mathcal{N}(z|0,1)}\left[\frac{\partial^2}{\partial f_i^2}\log p(y_i|f_i)\Big|_{f_i=z_i\sqrt{v_{q_i}}+m_{q_i}}\right]
\end{aligned} \qquad (8)$$

Since $\frac{\partial(\sqrt{v_{q_i}})}{\partial V} = \frac{1}{2\sqrt{v_{q_i}}}K_M^{-1}K_{Mi}K_{iM}K_M^{-1}$, we get

$$\begin{aligned}
\frac{\partial}{\partial V}\mathbb{E}_{q_i}\left[\log p(y_i|f_i)\right] &= \frac{1}{2}K_M^{-1}K_{Mi}K_{iM}K_M^{-1} \\
\mathbb{E}_{\mathcal{N}(z_i|0,1)}&\left[\frac{\partial^2}{\partial f_i^2}\log p(y_i|f_i)\Big|_{f_i=z_i\sqrt{v_{q_i}}+m_{q_i}}\right] \qquad (9)
\end{aligned}$$

Finally, defining

$$\rho_i = \mathbb{E}_{\mathcal{N}(f_i|m_{q_i},v_{q_i})}\left[\frac{\partial}{\partial f_i}\log p(y_i|f_i)\right] \qquad (10a)$$

$$\lambda_i = \mathbb{E}_{\mathcal{N}(f_i|m_{q_i},v_{q_i})}\left[\frac{\partial^2}{\partial f_i^2}\log p(y_i|f_i)\right] \qquad (10b)$$

and putting together the derivatives above with the standard derivatives of the KL divergence we get a simple characterization of the derivatives of the VLB:

$$\frac{\partial \mathrm{VLB}}{\partial \boldsymbol{m}} = \sum_i\left(\rho_i K_M^{-1}K_{Mi}\right) - K_M^{-1}(\boldsymbol{m}-\boldsymbol{m}_{\mathcal{U}}) \qquad (11a)$$

$$\frac{\partial \mathrm{VLB}}{\partial V} = \frac{1}{2}\sum_i\left(\lambda_i K_M^{-1}K_{Mi}K_{iM}K_M^{-1}\right) + \frac{1}{2}(V^{-1}-K_M^{-1}) \qquad (11b)$$

This formulation is generic in the sense that it has the same form for any likelihood function and is simply determined by ρ_i and λ_i. These quantities can be evaluated independently of the sparse model, and rely on derivatives of the observation distribution and their expectations under a Gaussian distribution. It can be seen from Section 3.3 that the regularity condition, $\ell(z) = \mathrm{o}(e^{\frac{1}{2}z^2})$ as $z \to \pm\infty$, is met by many likelihoods of interest.

3.2. Full Variational Gaussian Approximation

The full (non-sparse) variational approximation can be found as a "limiting case" of the sparse variational formulation when $\mathcal{X} = \mathcal{U}$. One can show that in this case, the projection term $K_M^{-1}K_{Mi}$ becomes $\boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the Euclidean unit coordinate vector and the VLB derivatives (11a) and (11b) become

$$\frac{\partial \mathrm{VLB}}{\partial \boldsymbol{m}} = \sum_i\left(\rho_i \boldsymbol{e}_i\right) - K_N^{-1}(\boldsymbol{m}-\boldsymbol{m}_N)$$

$$\frac{\partial \mathrm{VLB}}{\partial V} = \frac{1}{2}\sum_i\left(\lambda_i \boldsymbol{e}_i\boldsymbol{e}_i^T\right) + \frac{1}{2}(V^{-1}-K_N^{-1})$$

By setting the derivative with respect to V to zero, it is obvious that the optimal variational precision matches the prior precision everywhere except on the diagonal. These are exactly the forms of the derivatives that appear in (Opper & Archambeau, 2009) and (Khan et al., 2012).

Table 1. List of likelihood functions, their derivatives, and expectations of the derivatives with respect to $\mathcal{N}(f|m,v)$ as given by (10a) and (10b) where available in closed form (NA denotes not available). For the ordinal likelihood, L denotes the number of ordered categories, k_o is a shape parameter, and the bin edges $\{\phi_l\}_{l=1}^{L}$ obey $-\infty = \phi_0 < \phi_1 < \ldots < \phi_{L-1} < \phi_L = \infty$.

y	$p(y\|f)$	$\frac{\partial}{\partial f}\log p(y\|f)$	$\frac{\partial}{\partial f^2}\log p(y\|f)$	ρ	λ
$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-f)^2}{2\sigma^2}}$	$\frac{1}{\sigma^2}(y-f)$	$-\frac{1}{\sigma^2}$	$\frac{1}{\sigma^2}(y-m)$	$-\frac{1}{\sigma^2}$
$\{0,1,2,\ldots\}$	$\frac{1}{y!}e^{-e^f}e^{fy}$	$-e^f+y$	$-e^f$	$-e^{m+\frac{1}{2}v}+y$	$-e^{m+\frac{1}{2}v}$
$\{-1,+1\}$	$\sigma(yf)$	$y(1-\sigma(yf))$	$-\sigma(yf)\,\sigma(-yf)$	NA	NA
$\{1,2,\ldots,L\}$	$\sigma(k_o(\phi_y-f))$ $-\sigma(k_o(\phi_{y-1}-f))$	$k_o(1-\sigma(k_o(f-\phi_y))$ $-\sigma(k_o(f-\phi_{y-1})))$	$-k_o^2(\sigma(k_o(\phi_y-f))\sigma(k_o(f-\phi_y))$ $+\sigma(k_o(\phi_{y_i}-f))\sigma(k_o(f-\phi_y))$	NA	NA

3.3. Some Observation Models

In this section we illustrate the generality of the model by providing details of several specific observation likelihood functions. Table 1 provides a list of likelihood functions, derivatives, and evaluations of (10a) and (10b) for standard GP regression with Gaussian likelihood, count regression with Poisson likelihood, binary classification with Bernoulli-logit likelihood, and ordinal regression with a cumulative-logit likelihood. Closed form expressions for ρ and λ are not available for all likelihood functions. In these cases, Gaussian-Hermite quadrature is used to calculate the expectations. We remark here that all likelihoods are log concave in f which is useful for empirical analysis of our proposed fixed point operator in the next section.

In addition, our formulation applies directly (but is not limited) to the framework of generalized GP models (Shang & Chan, 2013) in which $p(y_i|\theta_i)$ is given by an exponential family distribution where θ_i is related to f_i through the link function. In this case the individual derivatives defining ρ_i, λ_i are given by standard quantities as in Eq (39-41) of (Shang & Chan, 2013).

4. VLB Optimization

Parameterized in $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, the optimal variational parameters are given by

$$\boldsymbol{m}^\star = K_{MN}\boldsymbol{\rho}^\star + \boldsymbol{m}_{\mathcal{U}} \tag{12a}$$

$$V^\star = \left(K_M^{-1} - K_M^{-1}K_{MN}\,\mathrm{diag}(\boldsymbol{\lambda}^\star)\,K_{NM}K_M^{-1}\right)^{-1} \tag{12b}$$

It is only for standard GP regression with Gaussian likelihood that closed form solutions for $\boldsymbol{m}^\star$ and $V^\star$ can be obtained (matching the ones in (Titsias, 2009)). In general, (12a) and (12b) are a set of nonlinear equations coupled through their dependencies on $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$.

We explore two inference algorithms for our model. Our first algorithm optimizes $(\boldsymbol{m}^\star, V^\star)$ by coordinate ascent across the parameters. Newton's method is used to optimize the variational mean at a cost of $\mathrm{O}(M^3)$. The optimization of the covariance is more difficult due to its dimensionality and due to the requirement for it to be positive definite. The cost of Newton's method for optimizing the covariance would be prohibitive at $\mathrm{O}(M^6)$ limiting us to gradient ascent procedures. In addition, a naive gradient based optimization must project the resulting gradient steps onto the legal region. Challis & Barber (2013) proposed an optimization through the Cholesky factor of the covariance which automatically guarantees the positive-definite requirement. Although Challis & Barber (2013) proposed a joint optimization in the mean and Cholesky factor of the covariance in their inference approach for the latent Gaussian model, in practice, we have found coordinate ascent to be superior to joint optimization. In particular, our experiments below report results for coordinate ascent where $\boldsymbol{m}$ is optimized with Newton's method, and V is optimized via L-BFGS over the Cholesky factor.

4.1. Fixed Point Operator

We propose an alternative approach to optimizing the covariance through the following fixed-point operator, $T : \mathbb{S}_{++}^M \to \mathbb{S}_{++}^M$ derived from the optimality condition (12b),

$$T(V) = \left(K_M^{-1} - K_M^{-1}K_{MN}\,\mathrm{diag}(\boldsymbol{\lambda})\,K_{NM}K_M^{-1}\right)^{-1} \tag{13}$$

By inspection of 13 and 11b, it is obvious that T contains $V^\star$ in its fixed point set. To prove that the limit of the sequence defined by $V^{(k+1)} = T(V^{(k)})$ is equal to $V^\star$ for any initial $V^{(0)}$, requires showing that T is a contraction mapping, that is, there exists an $L \in [0,1)$ such that $\|T(V) - T(U)\| \leq L\|V - U\|$, for all U, V.

The presence of the nonlinear operation in $\boldsymbol{\lambda}$ that maps the covariance to a vector has rendered a general proof of contraction for arbitrary likelihoods difficult. Reparameterizing to another form such as the variational precision or Cholesky factor does not remove this nonlinearity.

We next show that although the contraction property does not always hold, it does hold in many cases of interest. In particular, we test the property experimentally by simulat-

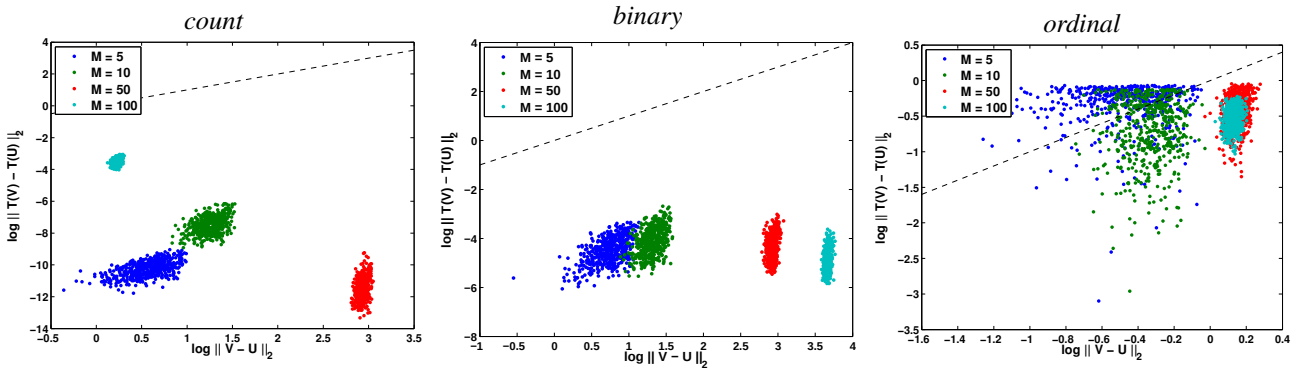

Figure 1. Results of contraction tests shown on log scale. The color coding refers to active set size. The dashed line represents the curve $\|T(U) - T(V)\|_2 = \|U - V\|_2$ in log space.

ing observations, drawing random matrices from $\mathcal{S}^M_{++}$, applying (13), and testing whether the contraction property is maintained. Now, since $\boldsymbol{\lambda}^\star \preceq 0$ for log concave likelihoods, $(V^\star)^{-1} \succeq K_M^{-1}$ is implied from (12b). This limits the pairs of covariances that require testing to those that satisfy $0 \preceq U, V \preceq K_M$.

We report here on tests using a zero-mean GP prior with Gaussian RBF kernel ($\ell = \frac{\sqrt{10}}{3}, \sigma^2 = 1$). The inputs $\{\mathcal{U}, \mathcal{X}\}$ are 1000 i.i.d (uniform) samples from the domain $[0, 1]^{10}$. We compare the 2-norm of the distance between covariance pairs before and after the mapping. We generate such data for each of the 3 observation models defined in the previous section and repeat the process 500 times.

The results are given in Figure 1. We observe that the contraction property appears to hold under the conditions tested for the count and binary models for all active set sizes, but not for the ordinal model at small set sizes. Additional tests (not reported here) using the Matern RBF kernel ($\nu = \frac{1}{2}$) and a polynomial degree 2 kernel showed the contraction property being maintained for all models under the same conditions. A broader characterization of (13) with respect to contraction is the subject of continuing work.

5. Experiments

To evaluate the proposed method, we apply it to count regression, binary classification, and ordinal regression. The datasets used in the experiment are summarized in Table 2. The dataset *ucsdpeds1l* contains counts of pedestrians extracted from video data and was used in (Chan & Vasconcelos, 2012). The datasets *stock* and *bank* were used in previous ordinal regression experiments with GPs (Chu & Ghahramani, 2005). The remaining datasets are available from the UCI Machine Learning Repository. In all experiments, data is normalized using training data only and the same normalization is applied to the test data.

As baselines for the sparse methods we compare against

Table 2. Summary of data sets. Values in parentheses refer to number of categories

NAME	SAMPLES	NO. DIM.	MODEL TYPE
UCSDPEDS1L	4000	30	COUNT
ABALONE	4177	8	COUNT
USPS35	1540	256	BINARY
MUSK	6958	166	BINARY
STOCK (5)	950	9	ORDINAL
BANK (10)	8192	32	ORDINAL

subset of data (SoD) algorithms that reduce data size to the active set in a similar manner but unlike the sparse methods ignore the additional data. We use four different variants of SoD. The first is the Laplace approximation. The remaining are all variational Gaussian approximations but differ in the method of optimization. We test the gradient ascent method described above and our fixed point method by restricting to the active subset to perform the optimization (i.e., $N = M$ and $K_{NM} = K_{MM}$). This can be seen as if we are applying the methods to the "full data" given by the active set. Finally, we also test the inference method by Khan et al. (2013) which performs the variational optimization in the dual space. Here too the "full data" is given by the active set and we use $W = I$ in their formulation.

For the sparse model, we compare three optimization methods: the gradient ascent method, the fixed point method, and the dual method in the form capturing the sparse model (using $W = K_{NM}K_{MM}^{-1}$) and implemented with L-BFGS.

We ran all algorithms on all problems, except that we could not apply the dual method in ordinal regression since we were not aware of a closed-form for the Fenchel conjugate which is required in the dual objective function.

All experiments use the GPML toolbox[1] for implementa-

[1] http://www.gaussianprocess.org/gpml/

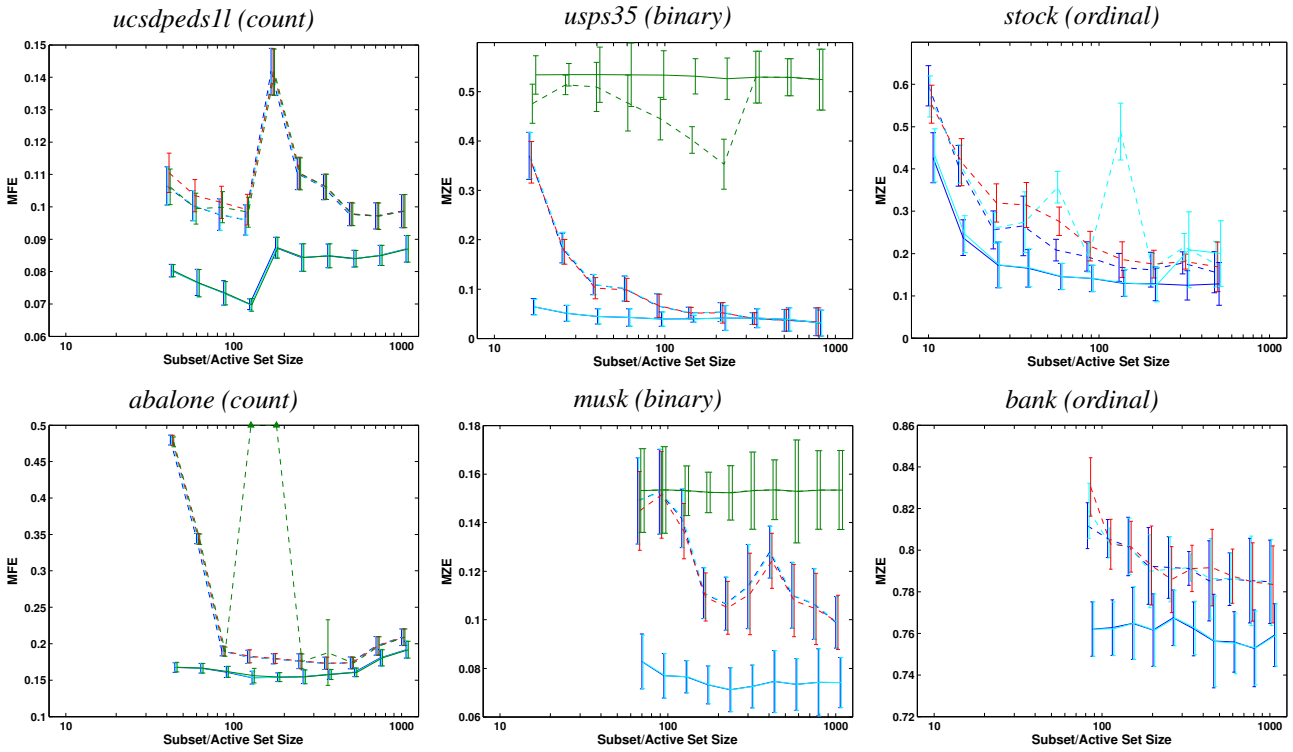

Figure 2. Learning curves with respect to subset/active set size. MFE is mean fractional error and MZE is mean zero-one error. Lower values represent better performance. Triangles on the edges of plots refer to data that exists outside the axes of the plot. Legend for plots: Laplace on SoD (- -), gradient ascent on SoD (- -), dual on SoD (- -), fixed point on SoD (- -), gradient ascent on full data (—), dual on full data (—), fixed point on full data (—).

tion of GP mean, covariance, and likelihood functions as well as for calculation of the approximate marginal likelihood via Laplace approximation and its derivatives. For consistency across methods, the minFunc software[2] is used for all gradient-based optimization.[3]

In our experiments we compare the algorithms when using the same active sets. As shown in previous work, search for useful inducing points in the sparse framework can yield a significant advantage in accuracy over subset of data, at the cost of increased run time, and this is one of the advantages of the variational framework. However, this complicates the comparison between methods. In addition, we start by comparing the methods when using the same fixed hyper parameters. This gives a direct comparison of the inference algorithms in the same context. The last comparison in this section includes learning of hyperparameters as well.

The settings for algorithms is as follows. A zero-mean GP with Gaussian RBF kernel is used in all cases. For the count likelihood, the predictions are the mean predictive estimates. For the binary classification and ordinal likelihoods, the predictions are the predictive modes. For all methods, initial variational parameters are found by running the Laplace approximation on the subset/active set. When used with SoD, the initial parameter of the dual method is obtained by solving a linear system (Eq 17 of (Khan et al., 2013)) with input parameters obtained from Laplace approximation on the subset. When used on the sparse model, the elements of the dual parameter are initialized to 1 for count regression and $\frac{1}{4}$ if $y_i = -1$ and $\frac{3}{4}$ if $y_i = +1$ for binary classification. The hyperparameters are either estimated from the subset/active set or set to default values ($\sigma^2 = 1$) prior to training using the same procedure across methods.

To investigate the performance, we generate learning curves as a function of active set size. For a given set size, the subset/active set is randomly selected from the data without replacement. After the inducing set is selected, 10-fold cross validation is performed with the remaining data. The results with respect to set size are shown in the plots in Figure 2. The curves are jittered horizontally to allow for comparison. The left column shows the count regression

code/matlab/doc/

[2] http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

[3] Stopping conditions are $\|\nabla f(x_k)\|_\infty \leq 10^{-5}$, $f(x_{k-1}) - f(x_k) \leq 10^{-9}$, or $k > 500$ where f is the objective function being optimized, k represents the iteration number, and x is the current optimization variable.

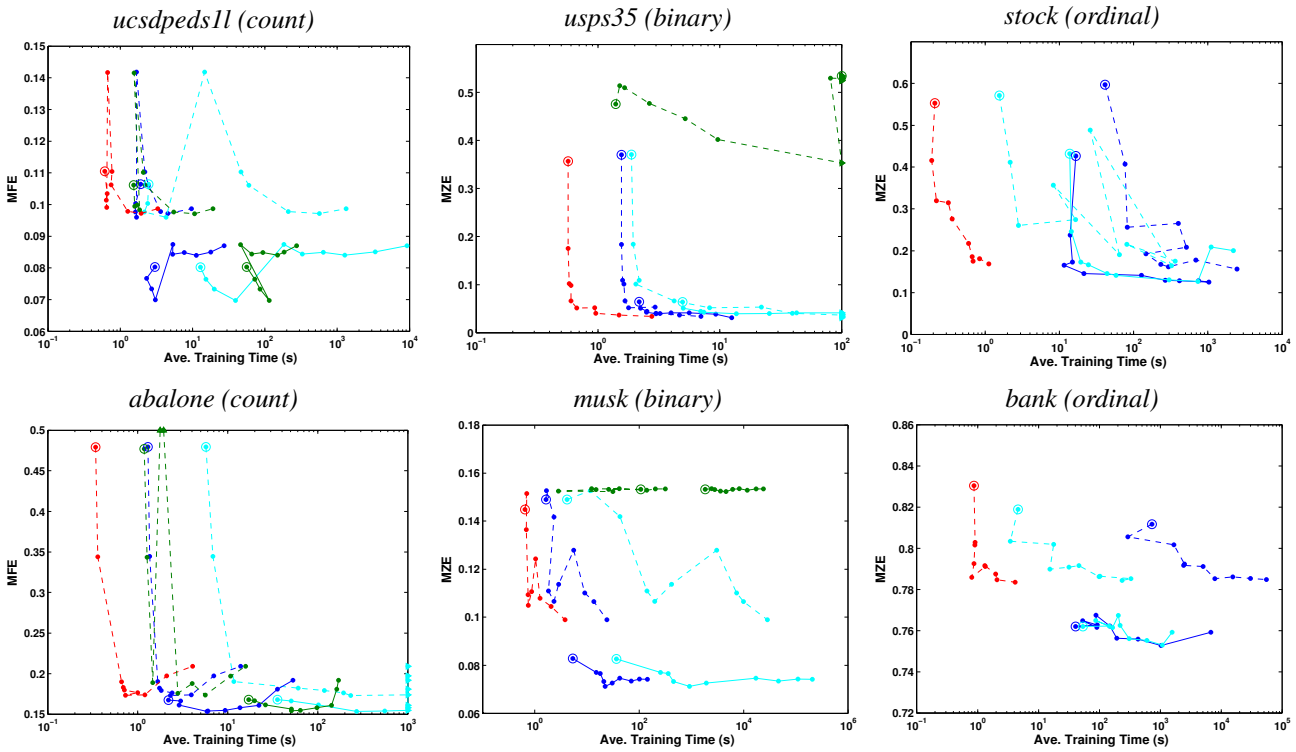

Figure 3. Training time / accuracy curves. Each dot represents a different subset/active set size. A circled dot represents the smallest subset/active set size for a method. Legend for plots: Laplace on SoD (- -), gradient ascent on SoD (- -), dual on SoD (- -), fixed point on SoD (- -), gradient ascent on full data (—), dual on full data (—), fixed point on full data (—).

tasks where the performance metric is mean fractional error (MFE). For count regression, we see all the sparse variational methods achieving the same performance. We expect equivalent performance between the dual and primal methods since strong duality holds with the Poisson likelihood. Notably, this performance is better than either Laplace approximation or variational Gaussian approximation with just a subset of data. The middle column shows the binary classification tasks where the performance metric is mean zero-one error (MERR). Here, gradient ascent and fixed point methods with the sparse model achieve the best performance across the datasets. The dual method applied on the SoD and sparse variational models yielded poor performance apparently due to convergence failures. Given the loss of strong duality for this likelihood, it is not guaranteed that the optimal solution would be located even if the optimization converged. Finally, the last column shows the output of the ordinal regression problems where mean-zero one error is used as the performance metric. Again, the sparse variational model with both gradient ascent and fixed point methods results in improved performance. To summarize, looking only at subset size the sparse methods have lower error than SoD and when they converge they provide similar results. The dual method is less stable for the classification task.

Figure 3 shows the same performance metrics with respect to the average time (across folds) required for training. For the sparse approach, the fixed point method is significantly faster than the gradient ascent or dual methods in the count regression and binary classification tasks, and is very close to the gradient method in ordinal regression. Comparing the variants of SoD we see that the fixed point method also shows some advantage in binary classification problems. This suggests that it might be a good alternative for variational inference in the full data case. Focusing on ordinal regression, we see that the fixed point method is no longer faster in the sparse case and is significantly slower than the gradient method for SoD. This may be partially explained by the results of the contraction experiment in the previous section. The combination of Gaussian RBF kernel and ordinal likelihood was the only one which resulted in the contraction property not being maintained. In summary, the gradient ascent method is the most consistent across problems but the fixed point method performs better in the cases where it was shown empirically to be a contraction.

Finally, we consider the comparison to the Laplace approximation. This method is simpler and can therefore handle larger active sets for the same run time. The figures clearly show that, in general, the fixed point method for the sparse model has higher training times than the Laplace approx-

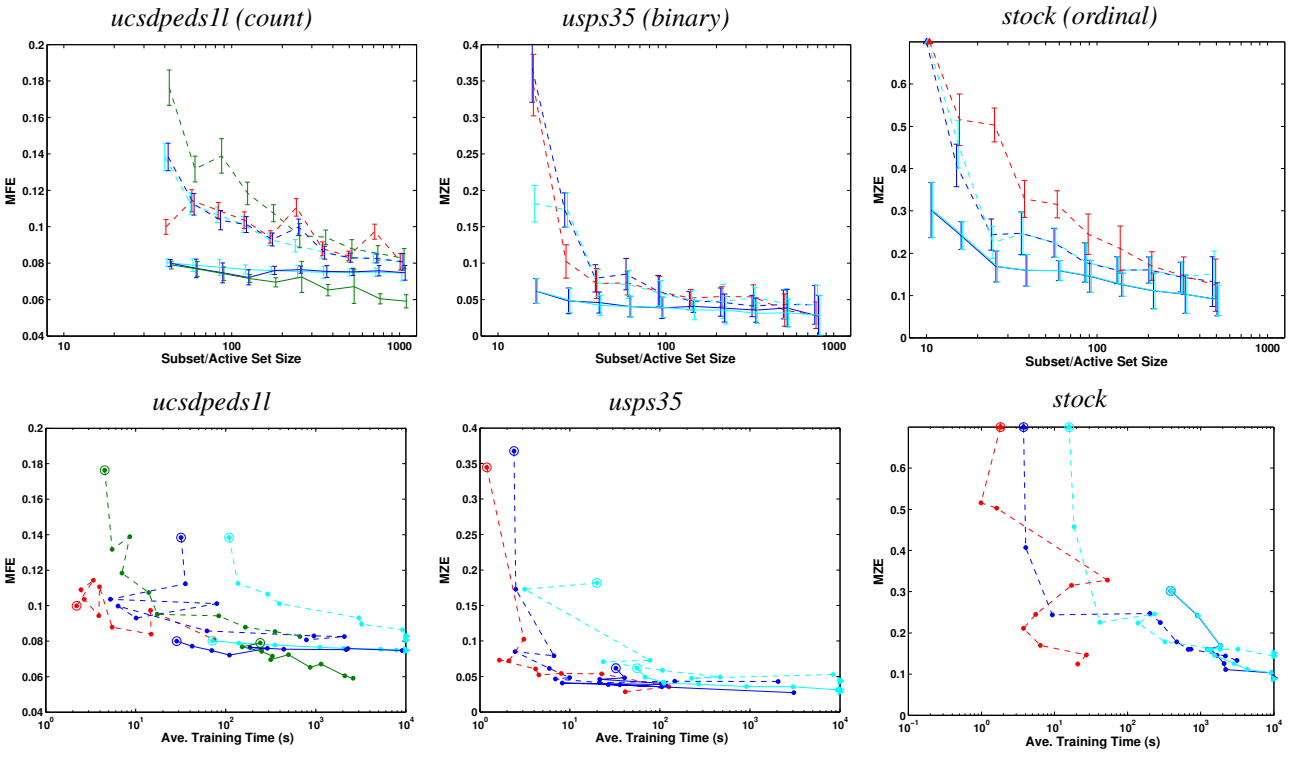

Figure 4. Results of hyperparameter optimization with respect to training time. Each dot represents a different subset/active set size. A circled dot represents the smallest subset/active set size for a method. Legend for plots: Laplace on SoD (- -), gradient ascent on SoD (- -), dual on SoD (- -), fixed point on SoD (- -), gradient ascent on full data (—), dual on full data (—), fixed point on full data (—).

imation. On the other hand, in a few cases, the sparse method with a small active set size outperforms the Laplace approximation even when a much larger dataset is used so that they have the same run time. This holds for the *ucsdpeds1l* count dataset, and the *musk* classification dataset.

Figure 4 displays the results of performing both inference for the variational posterior and hyperparameter optimization. Three of the datasets were used to compare the algorithms. Hyperparameter optimization was implemented using L-BFGS with the VLB as the objective function for all methods except Laplace approximation where the approximate marginal likelihood was used. In the case of the count data, both the sparse fixed point and dual methods provide some improvement in performance over SoD, but the fixed point method achieves improvement faster. In binary classification, there exist a range of active set sizes for which the fixed point method provides some improvement over Laplace approximation. The dual inference method suffered convergence issues under both SoD and sparse models for the binary classification task. Finally, as in the previous experiment, the fixed point method is slower on ordinal data. In summary, as in the previous experiment, with hyperparameter optimization the fixed point method is competitive with other sparse methods and sometimes faster, and can provide performance improvements over SoD.

6. Conclusion

The paper introduced a direct formulation of sparse GP with general likelihoods. The model combines the concept of active sets with the variational Gaussian approximation in a general framework. A novel characterization of the derivatives of the variational lower bound enables a generic solution that readily includes non-conjugate likelihood functions as well as the generalized GPs. The paper also proposed and evaluated a method based on fixed point iteration for optimizing the variational covariance, and showed that this operator acts as a contraction in practice in many cases. While the sparse variational inference problem can be solved with recently developed methods for latent variables GP models, our proposed method that includes fixed point updates generally outperforms these approaches both in terms of quality and stability.

The fixed point method was shown to be useful but it is not a contraction in all cases. Characterizing the the fixed point operator and specifically under what conditions it is a contraction operator is an important direction for future work. Given that it often converges in a few iterations, we propose that it can be a useful alternative to current approaches for the full variational Gaussian approximation.

References

Boukouvalas, Alexis, Barillec, Remi, and Cornford, Dan. Gaussian Process Quantile Regression using Expectation Propagation. June 2012. URL http://arxiv.org/abs/1206.6391.

Challis, Edward and Barber, David. Gaussian Kullback-Leibler Approximate Inference. *Journal of Machine Learning Research*, 14:2239–2286, 2013. URL http://jmlr.org/papers/v14/challis13a.html.

Chan, A. B. and Vasconcelos, N. Counting People With Low-Level Features and Bayesian Regression. *Image Processing, IEEE Transactions on*, 21 (4):2160–2177, April 2012. doi: 10.1109/TIP.2011.2172800. URL http://dx.doi.org/10.1109/TIP.2011.2172800.

Chu, Wei and Ghahramani, Zoubin. Gaussian Processes for Ordinal Regression. *J. Mach. Learn. Res.*, 6: 1019–1041, December 2005. ISSN 1532-4435. URL http://portal.acm.org/citation.cfm?id=1088707.

Chu, Wei, Sindhwani, Vikas, Ghahramani, Zoubin, and Keerthi, Sathiya S. Relational learning with gaussian processes. In *NIPS*, pp. 289–296, 2006.

Gal, Yarin, van der Wilk, Mark, and Rasmussen, Carl. Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3257–3265. Curran Associates, Inc., 2014.

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347, May 2013. URL http://jmlr.org/papers/v14/hoffman13a.html.

Keerthi, Sathiya S. and Chu, Wei. A matching pursuit approach to sparse Gaussian process regression. In *In Advances in Neural Information Processing Systems 18*. MIT Press, 2006.

Khan, Mohammad E., Mohamed, Shakir, and Murphy, Kevin P. Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression. In *NIPS*, pp. 3149–3157, 2012.

Khan, Mohammad E., Aravkin, Aleksandr Y., Friedlander, Michael P., and Seeger, Matthias. Fast Dual Variational Inference for Non-Conjugate LGMs. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 951–959, 2013. URL http://arxiv.org/abs/1306.1052.

Lázaro-gredilla, Miguel and Titsias, Michalis K. Variational heteroscedastic Gaussian process regression. In *In 28th International Conference on Machine Learning (ICML-11*, pp. 841–848. ACM, 2011.

Naish-Guzman, Andrew and Holden, Sean B. The Generalized FITC Approximation. In *NIPS*, 2007.

Opper, Manfred and Archambeau, Cédric. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792, 2009. doi: 10.1162/neco.2008.08-07-592. URL http://dx.doi.org/10.1162/neco.2008.08-07-592.

Quiñonero Candela, Joaquin, Rasmussen, Carl E., and Herbrich, Ralf. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:2005, 2005.

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

Seeger, Matthias, Williams, Christopher K. I., Lawrence, Neil D., and Dp, Sheeld S. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. In *Workshop on AI and Statistics 9*, 2003.

Shang, Lifeng and Chan, Antoni B. On Approximate Inference for Generalized Gaussian Process Models, November 2013. URL http://arxiv.org/abs/1311.6371.

Sindhwani, Wei C., Ghahramani, Zoubin, and Keerthi, Sathiya S. Relational learning with gaussian processes. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, pp. 289. MIT Press, 2007.

Snelson, Edward and Ghahramani, Zoubin. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pp. 1257–1264. MIT press, 2006.

Titsias, Michalis. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5, 2009. URL http://jmlr.csail.mit.edu/proceedings/papers/v5/titsias09a/titsias09a.pdf.

Titsias, Michalis and Lázaro-gredilla, Miguel. Doubly Stochastic Variational Bayes for non-Conjugate Inference. In Jebara, Tony and Xing, Eric P. (eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979. JMLR Workshop and Conference Proceedings, 2014. URL http://jmlr.org/proceedings/papers/v32/titsias14.pdf.

Titsias, Michalis K. and Lawrence, Neil D. Bayesian Gaussian process latent variable model. In *Thirteenth International Conference on Artificial Intelligence and Statistics*, May 2010. URL http://jmlr.org/proceedings/papers/v9/titsias10a/titsias10a.pdf.

Vanhatalo, Jarno and Vehtari, Aki. Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology. *Journal of Machine Learning Research - Proceedings Track*, 1: 73–89, 2007.

Vanhatalo, Jarno, Jylänki, Pasi, and Vehtari, Aki. Gaussian process regression with Student-t likelihood. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1910–1918. 2009.