Multiple Gaussian Process Models

Cedric Archambeau
Xerox Research Centre Europe
6, Ch. de Maupertuis, 38240 Meylan, France
cedric.archambeau@xerox.com

Francis Bach
INRIA-WILLOW Project-Team
23, Av. d'Italie, 75214 Paris, France
francis.bach@inria.fr

Abstract

We consider a Gaussian process formulation of the multiple kernel learning problem. The goal is to select the convex combination of kernel matrices that best explains the data and by doing so improve the generalisation on unseen data. Sparsity in the kernel weights is obtained by adopting a hierarchical Bayesian approach: Gaussian process priors are imposed over the latent functions and generalised inverse Gaussians on their associated weights. This construction is equivalent to imposing a product of heavy-tailed process priors over function space. A variational inference algorithm is derived for regression and binary classification.

1 Introduction

Kernel-based methods are well-established tools for supervised learning, allowing to perform various tasks, such as regression or binary classification, with linear and non-linear predictors. Like most statistical models, kernel-based methods can be considered within two frameworks: in the frequentist approach, estimators are obtained by minimizing a regularized empirical risk, leading e.g. to kernel ridge regression or the support vector machine [STC04, SBS00]; in the Bayesian approach, Gaussian processes (GPs) provide a Bayesian interpretation to kernel-based methods [RW06], with the potential to learn the kernel parameters from the data without having to use cross-validation.

Crucial to the predictive performance of kernel methods is the choice of the kernel function. In the Bayesian setting, the kernel function (often called covariance function) determines the correlations between the predictions we make. Assuming that the predictor's smoothness is fully specified by these correlations can be formalised by a Gaussian process imposed over function space. Techniques based on automatic relevance determination have been successful at learning the parameters of kernel functions such as the individual length scales of the squarred exponential kernel [RW06]. In the frequentist setting, a specific parameterization of kernel functions has led to a significant amount of work, namely positive linear combination of pre-defined kernel functions (or kernel matrices), leading to the multiple kernel learning (MKL) framework [LCB$^+$04, BLJ04]. The first contribution of this paper is to propose a Gaussian process (GP) formulation of the multiple kernel learning framework, which we refer to as *multiple Gaussian process* (MGP) models. Its second contribution is to provide a framework to consider all ℓ_p-norms at once and to determine *from data* whether we should use a sparsity-inducing prior or not. Currently, there is no consensus in the frequentist community on how to choose the type of regularization. In practice, however, the choice of the regulariser leads to solutions of very different kinds. For example, when considering an ℓ_1-norm a sparse solution will be obtained whether or not it is supported by the data. Obviously, if all kernels are important for prediction, this will be detrimental [GN09].

2 Multiple Gaussian process model for regression

Let $\{y_n\}_{n=1}^N$ be the set of noisy targets and $\{\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nP}\}_{n=1}^N$ the set of features, which are assumed to be non-random column vectors. We consider a weighted linear model of the P feature

vectors with i.i.d. Gaussian noise:

$$\mathbf{y}|\mathbf{X}_1,\ldots,\mathbf{X}_P,\mathbf{w}_1,\ldots,\mathbf{w}_P \sim \mathcal{N}(\textstyle\sum_p \mathbf{X}_p\mathbf{w}_p, \tau^{-1}\mathbf{I}_N), \tag{1}$$

where $\mathbf{y} = (y_1,\ldots,y_N)^\top$ and $\mathbf{I}_N$ is the identity matrix of dimension N. The weights associated to the feature matrix $\mathbf{X}_p \in \mathbb{R}^{N\times D_p}$ are denoted by $\mathbf{w}_p \in \mathbb{R}^{D_p}$ and the residual precision by τ.

The case of interest is the one where the weight vectors are sparse, i.e., many of their elements are (nearly) zero. However, we do not know a priori the degree of sparsity. From a Bayesian perspective, the spike and slab prior is the golden standard for inducing sparsity. Here, follow a different approach and choose the prior $p(\mathbf{w}_p)$ to be a Gaussian scale mixture [AM74] centred at zero. In effect we approximate the spike and slab prior by a continuous prior which favours sparse solutions. Although zero probability mass is put on exact zero values, the use of heavy-tailed priors allows us to infer large, as well as quasi zero values for the kernel weights.

Formally, we impose the following product of zero-mean Gaussian scale mixtures on the weights:

$$\mathbf{w}|\boldsymbol{\gamma} \sim \textstyle\prod_p \mathcal{N}(\mathbf{0}, \gamma_p^{-1}\mathbf{I}_{D_p}), \qquad\qquad \boldsymbol{\gamma} \sim \prod_p \mathcal{N}^{-1}(\omega,\chi,\phi), \tag{2}$$

where $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_P)^\top$ is the vector of unobserved scale variables on which independent generalised inverse Gaussian densities (see Appendix A) are imposed. The marginal $p(\mathbf{w})$ is then a symmetric generalised hyperbolic density [Hu05], which has fat tails compared to the Gaussian. This family contains the Student-t, the Laplace, the Gamma-variance and Jeffrey's as special cases.

Given this probabilistic model, one can integrate out $\mathbf{w}$, leading to a closed form expression for the marginal density of the observations:

$$\mathbf{y}|\boldsymbol{\gamma} \sim \mathcal{N}\big(\mathbf{0}, \textstyle\sum_p \gamma_p^{-1}\mathbf{K}_p + \tau^{-1}\mathbf{I}_N\big), \tag{3}$$

where $\mathbf{K}_p = \mathbf{X}_p\mathbf{X}_p^\top \in \mathbb{R}^{N\times N}$ is the kernel matrix associated with the p-th feature matrix. Clearly, the marginal density is well-defined for any set of valid kernel matrices since $\gamma_p > 0$ for all p.

The linear additive model defined in (1) corresponds to the *weight-space* view representation of the multiple Gaussian process model (MGP). From (3), however, we see that the marginal density only depends on the linear kernel matrices $\{\mathbf{K}_p\}_{p=1}^P$. Hence, the model can be generalised to a non-linear additive model by replacing these linear kernel matrices by non-linear ones. This new representation corresponds to the multiple *function-space* view representation.

Let $\{f_p(\cdot)\}_{p=1}^P$ be a set of P latent functions on which we impose scaled Gaussian process priors:

$$f_p(\cdot)|\gamma_p \sim \mathcal{GP}(0, \gamma_p^{-1}k_p(\cdot,\cdot)), \tag{4}$$

for all p. The functions $\{k_p(\cdot,\cdot)\}_{p=1}^P$ are covariance functions, which are also valid kernel functions [RW06]. Again we consider i.i.d. Gaussian noise, but assume $\{y_n\}_{n=1}^N$ are noisy observations of a sum of P latent *function* values $\mathbf{f}_p \in \mathbb{R}^N$. The likelihood function and the MGP prior are given by

$$\mathbf{y}|\mathbf{f},\tau \sim \mathcal{N}(\textstyle\sum_p \mathbf{f}_p, \tau^{-1}\mathbf{I}_N) = \mathcal{N}(\mathbf{M}\mathbf{f}, \tau^{-1}\mathbf{I}_N), \tag{5}$$

$$\mathbf{f}|\boldsymbol{\gamma} \sim \textstyle\prod_p \mathcal{N}(\mathbf{0}, \gamma_p^{-1}\mathbf{K}_p) = \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{K}}), \tag{6}$$

where $\mathbf{f}^\top = (\mathbf{f}_1^\top,\ldots,\mathbf{f}_P^\top)$, $\mathbf{M} = (\mathbf{1}_P^\top \otimes \mathbf{I}_N)$ and $\widetilde{\mathbf{K}} = \mathrm{diag}\{\gamma_1^{-1}\mathbf{K}_1,\ldots,\gamma_P^{-1}\mathbf{K}_P\}$. Vector $\mathbf{1}_P$ is the unit vector of dimension P and the operator $\otimes$ denotes the Kronecker product. The prior on $\boldsymbol{\gamma}$ is still given by (2) and the marginal $p(\mathbf{y}|\boldsymbol{\gamma})$ has the same form as in the weight-space view representation.

The MGP model corresponds to imposing P independent *non-Gaussian* process priors over function space. If we condition on the corresponding scale variable, any finite subset of latent function values is distributed according to a multivariate Gaussian marginal. For any of these marginals one can integrate out the scale variable, such that any finite set of latent function values is distributed according to a product of P independent multivariate Gaussian scale mixture densities:

$$\mathbf{f} \sim \prod_p \int \mathcal{N}(\mathbf{0}, \gamma_p^{-1}\mathbf{K}_p)\; p(\gamma_p)\; d\gamma_p \propto \prod_p \frac{K_{\omega+\frac{N}{2}}\Big(\sqrt{\chi(\phi+\mathbf{f}_p^\top\mathbf{K}_p^{-1}\mathbf{f}_p)}\Big)}{\Big(\sqrt{(\phi+\mathbf{f}_p^\top\mathbf{K}_p^{-1}\mathbf{f}_p)/\chi}\Big)^{\omega+\frac{N}{2}}}. \tag{7}$$

where $K_\omega(\cdot)$ is the modified Bessel function of the second kind. Hence, the prior measure imposed over function space is a heavy-tailed one, known as the generalised hyperbolic measure [BN77,

Hu05]. The Gaussian process is recovered for $\omega \to \infty$ and the symmetric multivariate zero-mean hyperbolic process is obtained for $\omega = -1$. Other special cases include the multivariate Gamma-variance process ($\omega < 0$ and $\phi = 0$), the multivariate Laplace process ($\omega = -1$ and $\phi = 0$), the multivariate Student-t process ($\omega > 0$ and $\chi = 0$) and the multivariate Cauchy process ($\omega = 1/2$ and $\chi = 0$).

The most straightforward approach for the estimation of $\boldsymbol{\gamma}$ is to use type II maximum a postriori (or type II maximum likelihood in absence of prior on $\boldsymbol{\gamma}$ as adopted in [KGUD10]). The optimisation can be performed using standard nonlinear optimisation tools, but the regulariser needs to be chosen in advance. Instead, we turn our attention to the inference problem of these parameters from data. We view $\boldsymbol{\gamma}$ as a latent variable and the desired level of sparsity is learnt from the data by optimising the hyperparameters by type II maximum likelihood (ML).

3 Variational inference with type II maximum likelihood

We follow a mean field approach [Bea03, Bis06]. In order to find an analytically tractable solution, the posterior over the latent function values $\mathbf{f}$ and the scale vector $\boldsymbol{\gamma}$ is assumed to factorise given the data, that is $q(\mathbf{f}, \boldsymbol{\gamma}) = q(\mathbf{f}) \prod_p q(\gamma_p)$. It can be shown that the variational posteriors maximising the negative variational free energy (a lower bound to the log-marginal likelihood) are given by $q(\mathbf{f}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{\gamma}) = \prod_p \mathcal{N}^-(\omega_p, \chi_p, \phi_p)$. The parameters are defined as

$$\boldsymbol{\mu} = \tau \boldsymbol{\Sigma} \mathbf{M}^\top \mathbf{y}, \quad \boldsymbol{\Sigma} = \left(\bar{\mathbf{K}}^{-1} + \tau \mathbf{M}^\top \mathbf{M}\right)^{-1}, \quad \omega_p = \omega + N/2, \quad \chi_p = \chi, \quad \phi_p = \phi + \langle \mathbf{f}_p^\top \mathbf{K}_p^{-1} \mathbf{f}_p \rangle,$$

where $\bar{\mathbf{K}} = \mathrm{diag}\{\langle \gamma_1 \rangle^{-1} \mathbf{K}_1, \ldots, \langle \gamma_P \rangle^{-1} \mathbf{K}_P\}$.

The predictive MGP is obtained by assuming that $q(\boldsymbol{\gamma})$ is peaked around its mean such that $q(f(\mathbf{x})) \approx q(f(\mathbf{x})|\langle \boldsymbol{\gamma} \rangle)$. The true predictive density can then be approximated by the following analytically tractable integral:

$$y(\mathbf{x})|\mathbf{y} \sim \int p(y(\mathbf{x})|f(\mathbf{x}))\, q(f(\mathbf{x}))\, df(\mathbf{x}) = \mathcal{GP}\big(m(\mathbf{x}), v(\mathbf{x}) + \tau^{-1}\big). \tag{8}$$

where

$$m(\mathbf{x}) = \textstyle\sum_p \langle \gamma_p \rangle^{-1} \mathbf{k}_p(\mathbf{x}_p)^\top \mathbf{B}^{-1} \mathbf{y}, \tag{9}$$

$$v(\mathbf{x}) = \textstyle\sum_p \langle \gamma_p \rangle^{-1} k_p(\mathbf{x}_p, \mathbf{x}_p) - \sum_p \sum_q \langle \gamma_p \rangle^{-1} \langle \gamma_q \rangle^{-1} \mathbf{k}_p(\mathbf{x}_p)^\top \mathbf{B}^{-1} \mathbf{k}_q(\mathbf{x}_q)\big\}, \tag{10}$$

with $\mathbf{B} = \sum_r \langle \gamma_r \rangle^{-1} \mathbf{K}_r + \tau^{-1} \mathbf{I}_N$. From these expression we see that the posterior mean and variance have the same form as in standard GP regression; the kernel is simply replaced by a convex combination of kernels. Note, moreover, that the expression $m(\mathbf{x})$ has the same form as the one we would obtain with a frequentist method such as kernel ridge regression.

The ML II updates for the hyperparameters are obtained by solving the following expressions (which are simple root finding equations, with unique solutions, hence easily solved by binary search):

$$\omega : P \ln \sqrt{\tfrac{\phi}{\chi}} - P \tfrac{d \ln K_\omega(\sqrt{\chi\phi})}{d\omega} + \textstyle\sum_p \langle \ln \gamma_p \rangle = 0, \tag{11}$$

$$\chi : \tfrac{P\omega}{\chi} - \tfrac{P}{2} \sqrt{\tfrac{\phi}{\chi}} R_\omega(\sqrt{\chi\phi}) + \tfrac{1}{2} \textstyle\sum_p \langle \gamma_p^{-1} \rangle = 0, \tag{12}$$

$$\phi : -\tfrac{P}{2} \sqrt{\tfrac{\chi}{\phi}} R_\omega(\sqrt{\chi\phi}) + \tfrac{1}{2} \textstyle\sum_p \langle \gamma_p \rangle = 0, \tag{13}$$

where $R_\omega(\cdot) = K_{\omega+1}(\cdot)/K_\omega(\cdot)$. These updates are obtained by direct maximising of the variational bound. The update for τ is obtained in the same manner.

4 MGP for classification

We restrict ourselves to binary classification and consider a scaled probit model in which the likelihood is derived from the Gaussian cumulative density. A probit model is equivalent to a Gaussian noise and a step function likelihood [AC93, OW00].

Table 1: Average root mean square error for toy regression data (lower is better). The multiple Laplace process performs worse than the Student-t and the Gamma-variance process when the generating process is sparse. ARD performs poorly when the generating process is not sparse. In the case of MKL the prior choice of the regulariser leads to more sensitivity to model misspecifications.

Number of active kernels	1 out of 10	3 out of 10	10 out of 10
Multiple Student-t	.033 ($\pm$.027)	.067 ($\pm$.032)	.719 ($\pm$.221)
Multiple Laplace	.034 ($\pm$.028)	.076 ($\pm$.035)	.704 ($\pm$.204)
Multple Gamma-variance	.033 ($\pm$.027)	.067 ($\pm$.032)	.719 ($\pm$.223)
ARD	.033 ($\pm$.027)	.066 ($\pm$.031)	.746 ($\pm$.223)
MKL ℓ_1	.037 ($\pm$.032)	.066 ($\pm$.030)	.720 ($\pm$.203)
MKL ℓ_2	.830 ($\pm$.655)	.831 ($\pm$.399)	.762 ($\pm$.251)
MKL $\ell_{4/3}$	.097 ($\pm$.062)	.233 ($\pm$.098)	.719 ($\pm$.238)

Let $\{t_n\}_{n=1}^N$ be the class labels, with $t_n \in \{-1, +1\}$ for all n. The likelihood (5) is replaced by

$$\mathbf{t}|\mathbf{y} \sim \textstyle\prod_n I(t_n y_n), \qquad \mathbf{y}|\mathbf{f} \sim \mathcal{N}(\textstyle\sum_p \mathbf{f}_p, \tau^{-1}\mathbf{I}_N), \tag{14}$$

where $I(z) = 1$ for $z \geqslant 0$ and 0 otherwise.

As in the case of regression, we consider a mean field approximation and assumes the posterior is of the form $q(\mathbf{y})q(\mathbf{f})q(\boldsymbol{\gamma})$. We further assume the variational posterior $q(\mathbf{y})$ is a product of truncated Gaussians (see Appendix B):

$$q(\mathbf{y}) \propto \textstyle\prod_n I(t_n y_n)\mathcal{N}(\nu_n, \lambda_n) = \left(\prod_{t_n=+1} \mathcal{N}_+(\nu_n, \lambda_n)\right)\left(\prod_{t_n=-1} \mathcal{N}_-(\nu_n, \lambda_n)\right), \tag{15}$$

where $\nu_n = \sum_p \langle f_p(\mathbf{x}_{np})\rangle$ and $\lambda_n = 1/\tau$. The posterior mean and the posterior covariance of $\mathbf{f}$ are unchanged, except that $\mathbf{y}$ is replaced by $\boldsymbol{\nu}_\pm$. The elements of $\boldsymbol{\nu}_\pm$ are defined in (20). The posterior $q(\boldsymbol{\gamma})$ and the updates for the hyperparameters are identical to the ones in MGP regression.

In Bayesian classification the label with highest probability $P(t(\mathbf{x})|\mathbf{t})$ is selected. Since an exact computation is analytically intractable, we assume the posteriors $q(\mathbf{y})$ and $q(\boldsymbol{\gamma})$ are highly peaked around their mean leading to the following classification rule:

$$P(t(\mathbf{x}) = \pm 1|\mathbf{t}) \approx P(t(\mathbf{x}) = \pm 1|\mathbf{t}, \boldsymbol{\nu}_\pm, \langle\boldsymbol{\gamma}\rangle) = \Phi\big(\pm m(\mathbf{x})/\sqrt{v(\mathbf{x}) + \tau^{-1}}\big), \tag{16}$$

where $m(\mathbf{x})$ and $v(\mathbf{x})$ are as before with $\mathbf{y}$ replaced by $\boldsymbol{\nu}_\pm$. Deciding whether the label is -1 or $+1$ is equivalent to using the sign of $m(\mathbf{x})$ as the decision rule.

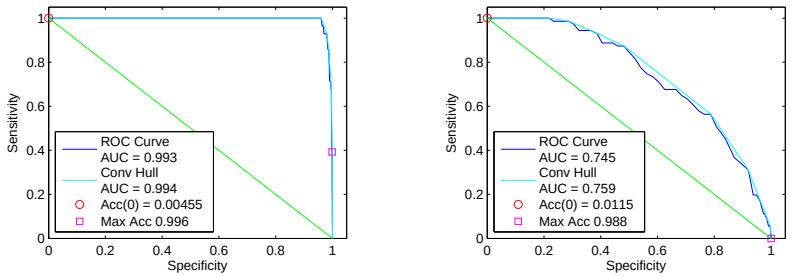
5 Discussion

We compare frequentist and Bayesian approaches to kernel combination. We demonstrate the flexibility and the performance of the MGP models on the following two detat sets:

Toy regression data. We generate random functions from the Hilbert spaces induced by 10 Laplacian kernels and add Gaussian i.i.d. noise; we show results on three different settings: a sparse problem, where only one kernel is used to generate the response, a semi-sparse problem with 3 functions are used and a non-sparse problem where all ten functions are active. Table 1 compares several MGP models and several cross-validated MKL models with fixed regularisation norms. Fig. 2 in the Appendix shows that the hierarchical Bayesian approach is able to adapt to the sparsity of the data.

Flowers data set.[1] Due to a lack a space we do not describe the data and the features, but only mention it is a standard MKL benchmark for multi-class image classification. For each of the 102 flower classes we learn a one-versus-all classifier. Fig. 1 shows the ROC curve for two classes when considering a Student-t process, for which we obtained an average $\mathrm{AUC} = .948 \pm .057$. The average AUC for Gamma-variance process and ARD are respectively given by $= .957 \pm .050$ and $.947 \pm .058$. All are better than state-of-the-art MKL results [MEZ08].

[1] www.robots.ox.ac.uk/ vgg/data/flowers/102/



(a) Typical roc curve (class 102).

(b) Worse roc curve (class 96).

Figure 1: Flowers data. ROC curves obtained for Student-t process one-versus-all classification for two flower classes. The ROC curves obtained for Gamma-variance process are slightly better.

References

[AC93] J. H. Albers and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[AM74] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36(1):99–102, 1974.

[Bea03] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, United Kingdom, 2003.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[BLJ04] F. R. Bach, G: R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In Carla E. Brodley, editor, *21st International Conference on Machine Learning (ICML)*. ACM, 2004.

[BN77] O. E. Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of the particle size. *Proceedings of the Royal Society, Series A, Mathematical and Physical Sciences*, 353:401–419, 1977.

[GN09] P. V. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *International Conference on Computer Vision 2009 (ICCV)*, pages 221–228, 2009.

[Hu05] Wenbo Hu. *Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk*. PhD thesis, Florida State University, United States of America, 2005.

[Jør82] B. Jørgensen. *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Springer-Verlag, 1982.

[KGUD10] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal in Computer Vision*, 2010.

[LCB+04] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[MEZ08] M.-E.Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[OW00] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

[RW06] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[SBS00] Bernhard Schölkopf, Chris Burges, and Alex Smola, editors. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 2000.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

A Generalised inverse Gaussian density

The generalised inverse Gaussian density is defined as follows [Jør82]:

$$x \sim \mathcal{N}^{-1}(\omega, \chi, \phi) = \frac{\chi^{-\omega}(\sqrt{\chi\phi})^{\omega}}{2K_{\omega}(\sqrt{\chi\phi})} x^{\omega-1} e^{-\frac{1}{2}(\chi x^{-1} + \phi x)}, \tag{17}$$

where $x > 0$ and $K_{\omega}(\cdot)$ is the modified Bessel function of the second kind with index $\omega \in \mathbb{R}$. Depending on the value taken by ω, we have the following constraints on χ and ϕ:

$$\begin{cases} \omega > 0: & \chi \geqslant 0,\ \phi > 0, \\ \omega = 0: & \chi > 0,\ \phi > 0, \\ \omega < 0: & \chi > 0,\ \phi \geqslant 0. \end{cases}$$

Let us define $R_{\omega}(\cdot) = K_{\omega+1}(\cdot)/K_{\omega}(\cdot)$. The following expectations are useful:

$$\langle x \rangle = \sqrt{\frac{\chi}{\phi}} R_{\omega}(\sqrt{\chi\phi}), \qquad \langle x^{-1} \rangle = \sqrt{\frac{\phi}{\chi}} R_{-\omega}(\sqrt{\chi\phi}), \qquad \langle \ln x \rangle = \ln \sqrt{\frac{\chi}{\phi}} + \frac{d \ln K_{\omega}(\sqrt{\chi\phi})}{d\omega}, \tag{18}$$

When $\chi = 0$ and $\omega > 0$, the generalised inverse Gaussian density reduces to the Gamma density. When $\phi = 0$ and $\omega < 0$, it reduces to the inverse Gamma density. The expectations simplify also.

B Truncated Gaussian density

The (positive/negative) truncated Gaussian density is defined as follows:

$$\mathcal{N}_{\pm}(\mu, \sigma^2) = \Phi(\pm\mu/\sigma)^{-1} \mathcal{N}(\mu, \sigma^2), \tag{19}$$

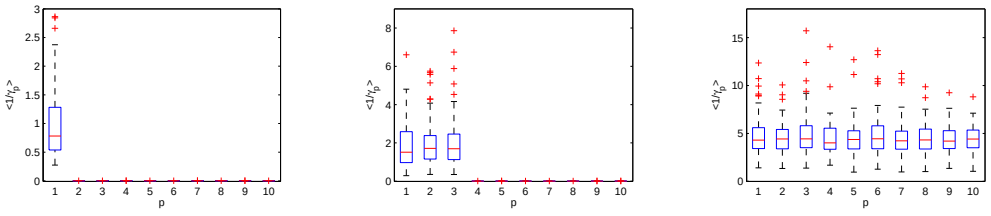
where $\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(0,1) dz$ is the cumulative density of the unit Gaussian.

Let $x_{\pm} \sim \mathcal{N}_{\pm}(\mu, \sigma^2)$. The mean and variance are given by

$$\langle x_{\pm} \rangle = \mu \pm \sigma^2 \mathcal{N}_{\pm}(0|\mu, \sigma^2), \tag{20}$$

$$\langle (x_{\pm} - \langle x_{\pm} \rangle)^2 \rangle = \sigma^2 \mp \sigma^2 \mu \mathcal{N}_{\pm}(0|\mu, \sigma^2) - \sigma^4 \mathcal{N}_{\pm}(0|\mu, \sigma^2)^2. \tag{21}$$

C Example of the inferred kernel weights



(a) 1 active kernel. (b) 3 active kernels. (c) 10 active kernels.

Figure 2: Toy regression data. Shown are the box-and-whisker plots of the expected weight for each kernel when consider a Gamma prior (Student-t process). Other generalised inverse Gaussian priors perform as well. The MGP model was run on hundred different data set realisations.