Heavy-Tailed Process Priors for Selective Shrinkage

Fabian L. Wauthier
University of California, Berkeley
flw@cs.berkeley.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Abstract

Heavy-tailed distributions are often used to enhance the robustness of regression and classification methods to outliers in output space. Often, however, we are confronted with "outliers" in input space, which are isolated observations in sparsely populated regions. We show that heavy-tailed stochastic processes (which we construct from Gaussian processes via a copula), can be used to improve robustness of regression and classification estimators to such outliers by selectively shrinking them more strongly in sparse regions than in dense regions. We carry out a theoretical analysis to show that selective shrinkage occurs when the marginals of the heavy-tailed process have sufficiently heavy tails. The analysis is complemented by experiments on biological data which indicate significant improvements of estimates in sparse regions while producing competitive results in dense regions.

1 Introduction

Gaussian process classifiers (GPCs) [12] provide a Bayesian approach to nonparametric classification with the key advantage of producing predictive class probabilities. Unfortunately, when training data are unevenly sampled in input space, GPCs tend to overfit in the sparsely populated regions. Our work is motivated by an application to protein folding where this presents a major difficulty. In particular, while Nature provides samples of protein configurations near the global minima of free energy functions, protein-folding algorithms, which imitate Nature by minimizing an estimated energy function, necessarily explore regions far from the minimum. If the estimate of free energy is poor in those sparsely-sampled regions then the algorithm has a poor guide towards the minimum. More generally this problem can be viewed as one of "covariate shift," where the sampling pattern differs in the training and testing phase.

In this paper we investigate a GPC-based approach that addresses overfitting by shrinking predictive class probabilities towards conservative values. For an unevenly sampled input space it is natural to consider a *selective shrinkage* strategy: we wish to shrink probability estimates more strongly in sparse regions than in dense regions. To this end several approaches could be considered. If sparse regions can be readily identified, selective shrinkage could be induced by tailoring the Gaussian process (GP) kernel to reflect that information. In the absence of such knowledge, Goldberg and Williams [5] showed that Gaussian process regression (GPR) can be augmented with a GP on the log noise level. More recent work has focused on partitioning input space into discrete regions and defining different kernel functions on each. Treed Gaussian process regression [6] and Treed Gaussian process classification [1] represent advanced variations of this theme that define a prior distribution over partitions and their respective kernel hyperparameters. Another line of research which could be adapted to this problem posits that the covariate space is a nonlinear deformation of another space on which a Gaussian process prior is placed [3, 13]. Instead of directly modifying the kernel matrix, the observed non-uniformity of measurements is interpreted as being caused by the spatial deformation. A difficulty with all these approaches is that posterior inference is based on MCMC, which can be overly slow for the large-scale problems that we aim to address.

1

This paper shows that selective shrinkage can be more elegantly introduced by replacing the Gaussian process underlying GPC with a stochastic process that has heavy-tailed marginals (e.g., Laplace, hyperbolic secant, or Student-t). While heavy-tailed marginals are generally viewed as providing robustness to outliers in the *output space* (i.e., the response space), selective shrinkage can be viewed as a form of robustness to outliers in the *input space* (i.e., the covariate space). Indeed, selective shrinkage means the data points that are far from other data points in the input space are regularized more strongly. We provide a theoretical analysis and empirical results to show that inference based on stochastic processes with heavy-tailed marginals yields precisely this kind of shrinkage.

The paper is structured as follows: Section 2 provides background on GPCs and highlights how selective shrinkage can arise. We present a construction of heavy-tailed processes in Section 3 and show that inference reduces to standard computations in a Gaussian process. An analysis of our approach is presented in Section 4 and details on inference algorithms are presented in Section 5. Experiments on biological data in Section 6 demonstrate that heavy-tailed process classification substantially outperforms GPC in sparse regions while performing competitively in dense regions. The paper concludes with an overview of related research and final remarks in Sections 7 and 8.

2 Gaussian process classification and shrinkage

A Gaussian process (GP) [12] is a prior on functions $z : \mathcal{X} \to \mathbb{R}$ defined through a mean function (usually identically zero) and a symmetric positive semidefinite kernel $k(\cdot, \cdot)$. For a finite set of locations $X = (x_1, \ldots, x_n)$ we write $z(X) \sim p(z(X)) = \mathcal{N}(0, K(X, X))$ as a random variable distributed according to the GP with finite-dimensional kernel matrix $[K(X, X)]_{i,j} = k(x_i, x_j)$. Let y denote an n-vector of binary class labels associated with measurement locations X[1]. For Gaussian process classification (GPC) [12] the probability that a test point x_* is labeled as class $y_* = 1$, given training data (X, y), is computed as

$$p(y_* = 1 | X, y, x_*) = \mathbb{E}_{p(z(x_*)|X,y,x_*)} \left(\frac{1}{1 + \exp\{-z(x_*)\}} \right) \tag{1}$$

$$p(z(x_*)|X, y, x_*) = \int p(z(x_*)|X, z(X), x_*) p(z(X)|X, y) dz(X).$$

The predictive distribution $p(z(x_*)|X, y, x_*)$ represents a regression on $z(x_*)$ with a complicated observation model $y|z$. The central observation from Eq. (1) is that we could selectively shrink the prediction $p(y_* = 1|X, y, x_*)$ towards a conservative value $1/2$ by selectively shrinking $p(z(x_*)|X, y, x_*)$ closer to a point mass at zero.

3 Heavy-tailed process priors via the Gaussian copula

In this section we construct the heavy-tailed stochastic process by transforming a GP. As with the GP, we will treat the new process as a prior on functions. Suppose that $\text{diag}\,(K(X, X)) = \sigma^2 \mathbf{1}$. We define the heavy-tailed process $f(X)$ with marginal c.d.f. G_b as

$$z(X) \sim \mathcal{N}(0, K(X, X)) \tag{2}$$
$$u(X) = \Phi_{0,\sigma^2}(z(X)) \tag{3}$$
$$f(X) = G_b^{-1}(u(X)) = G_b^{-1}(\Phi_{0,\sigma^2}(z(X))).$$

Here the function $\Phi_{0,\sigma^2}(\cdot)$ is the c.d.f. of a centered Gaussian with variance σ^2. Presently, we only consider the case when G_b is the (continuous) c.d.f. of a heavy-tailed density g_b with scale parameter b that is symmetric about the origin. Examples include the Laplace, hyperbolic secant and Student-t distribution. We note that other authors have considered asymmetric or even discrete distributions [2, 11, 16] while Snelson et al. [15] use arbitrary monotonic transformations in place of $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$. The process $u(X)$ has the density of a Gaussian copula [10, 16] and is critical in transferring the correlation structure encoded by $K(X, X)$ from $z(X)$ to $f(X)$. If we define

---

[1]To improve the clarity of exposition, we only deal with binary classification for now. A full multiclass classification model is used in our experiments.

2

$z(f(X)) = \Phi_{0,\sigma^2}^{-1}(G_b(f(X)))$, it is well known [7, 9, 11, 15, 16] that the density of $f(X)$ satisfies

$$p(f(X)) = \frac{\prod_{i=1} g_b(f(x_i))}{|K(X,X)/\sigma^2|^{1/2}} \exp\left\{-\frac{1}{2} z(f(X))^\top \left[K(X,X)^{-1} - \frac{I}{\sigma^2}\right] z(f(X))\right\}. \quad (4)$$

Observe that if $K(X,X) = \sigma^2 I$ then $p(f(X)) = \prod_{i=1} g_b(f(x_i))$. Also note that if G_b were chosen to be Gaussian, we would recover the Gaussian process. The predictive distribution $p(f(x_*)|X, f(X), x_*)$ can be interpreted as a Heavy-tailed process regression (HPR). It is easy to see that its computation can be reduced to standard computations in a Gaussian model by nonlinearly transforming observations $f(X)$ into z-space. The predictive distribution in z-space satisfies

$$p(z(x_*)|X, f(X), x_*) = \mathcal{N}(\mu_*, \Sigma_*) \quad (5)$$

$$\mu_* = K(x_*, X)K(X,X)^{-1}z(f(X)) \quad (6)$$

$$\Sigma_* = K(x_*, x_*) - K(x_*, X)K(X,X)^{-1}K(X, x_*). \quad (7)$$

The corresponding distribution in f-space follows by another change of variables. Having defined the heavy-tailed stochastic process in general we now turn to an analysis of its shrinkage properties.

4 Selective shrinkage

By "selective shrinkage" we mean that the degree of shrinkage applied to a collection of estimators varies across estimators. As motivated in Section 2, we are specifically interested in selectively shrinking posterior distributions near isolated observations more strongly than in dense regions. This section shows that we can achieve this by changing the *form* of prior marginals (heavy-tailed instead of Gaussian) and that this induces stronger selective shrinkage than any GPR could induce. Since HPR uses a GP in its construction, which can induce some selective shrinkage on its own, care must be taken to investigate only the additional benefits the transformation $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$ has on shrinkage. For this reason we assume a particular GP prior which leads to a special type of shrinkage in GPR and then check how an HPR model built on top of that GP changes the observed behavior.

In this section we provide an idealized analysis that allows us to compare the selective shrinkage obtained by GPR and HPR. Note that we focus on regression in this section so that we can obtain analytical results. We work with n measurement locations, $X = (x_1, \ldots, x_n)$, whose index set $\{1, \ldots, n\}$ can be partitioned into a "dense" set D with $|D| = n-1$ and a single "sparse" index $s \notin D$. Assume that $x_d = x_{d'}, \forall d, d' \in D$, so that we may let (without loss of generality) $\tilde{K}(x_d, x_{d'}) = 1, \forall d \neq d' \in D$. We also assert that $x_d \neq x_s \ \forall d \in D$ and let $\tilde{K}(x_d, x_s) = \tilde{K}(x_s, x_d) = 0 \ \forall d \in D$. Assuming that $n > 2$ we fix the remaining entry $\tilde{K}(x_s, x_s) = \epsilon/(\epsilon + n - 2)$, for some $\epsilon > 0$. We interpret ϵ as a noise variance and let $K = \tilde{K} + \epsilon I$.

Denote any distributions computed under the GPR model by $p_{\mathrm{gp}}(\cdot)$ and those computed in HPR by $p_{\mathrm{hp}}(\cdot)$. Using $K(X, X) = K$, define $z(X)$ as in Eq. (2). Let y denote a vector of real-valued measurements for a regression task. The posterior distribution of $z(x_i)$ given y, with $x_i \in X$, is derived by standard Gaussian computations as

$$p_{\mathrm{gp}}(z(x_i)|X, y) = \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

$$\mu_i = \tilde{K}(x_i, X)K(X, X)^{-1}y$$

$$\sigma_i^2 = K(x_i, x_i) - \tilde{K}(x_i, X)K(X, X)^{-1}\tilde{K}(X, x_i).$$

For our choice of $K(X, X)$ one can show that $\sigma_d^2 = \sigma_s^2$ for $d \in D$. To ensure that the posterior distributions agree at the two locations we require $\mu_d = \mu_s$, which holds if measurements y satisfy

$$y \in \mathcal{Y}_{\mathrm{gp}} \triangleq \left\{y| \left(\tilde{K}(x_d, X) - \tilde{K}(x_s, X)\right)K(X,X)^{-1}y = 0\right\} = \left\{y \left| \sum_{d \in D} y_d = y_s\right.\right\}.$$

A similar analysis can be carried out for the induced HPR model. By Eqs. (5)–(7) HPR inference leads to identical distributions $p_{\mathrm{hp}}(z(x_d)|X, y') = p_{\mathrm{hp}}(z(x_s)|X, y')$ with $d \in D$ if measurements y' in f-space satisfy

$$y' \in \mathcal{Y}_{\mathrm{hp}} \triangleq \left\{y'| \left(\tilde{K}(x_d, X) - \tilde{K}(x_s, X)\right)K(X,X)^{-1}\Phi_{0,\sigma^2}^{-1}(G_b(y')) = 0\right\}$$

$$= \left\{y' = G_b^{-1}(\Phi_{0,\sigma^2}(y))|y \in \mathcal{Y}_{\mathrm{gp}}\right\}.$$

3

$$\text{(a) } g_b(x) = \frac{1}{2b}\exp\left\{-\frac{|x|}{b}\right\} \qquad \text{(b) } g_b(x) = \frac{1}{2b}\operatorname{sech}\left(\frac{\pi x}{2b}\right) \qquad \text{(c) } g_b(x) = \frac{1}{b\left(2+(x/b)^2\right)^{3/2}}$$
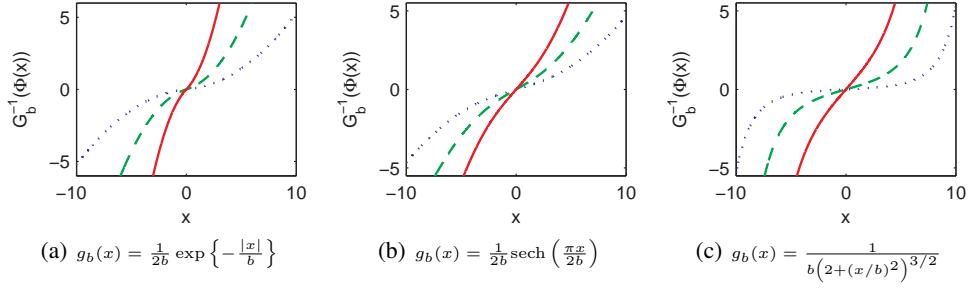
Figure 1: Illustration of $G_b^{-1}(\Phi_{0,\sigma^2}(x))$, for $\sigma^2 = 1.0$ with G_b the c.d.f. of (a) the Laplace distribution (b) the hyperbolic secant distribution (c) a Student-t inspired distribution, all with scale parameter b. Each plot shows three samples—dotted, dashed, solid—for growing b. As b increases the distributions become heavy-tailed and the gradient of $G_b^{-1}(\Phi_{0,\sigma^2}(x))$ increases.

To compare the shrinkage properties of GPR and HPR we analyze select pairs of measurements in $\mathcal{Y}_{\mathrm{gp}}$ and $\mathcal{Y}_{\mathrm{hp}}$. The derivation requires that $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$ is strongly concave on $(-\infty, 0]$, strongly convex on $[0, +\infty)$ and has gradient > 1 on $\mathbb{R}$. To see intuitively why this should hold, note that for G_b with fatter tails than a Gaussian, $|G_b^{-1}(\Phi_{0,\sigma^2}(x))|$ should eventually dominate $|\Phi_{0,b^2}^{-1}(\Phi_{0,\sigma^2}(x))| = (b/\sigma)|x|$. Figure 1 demonstrates graphically that the assumption holds for several choices of G_b, provided b is large enough, i.e., that g_b has sufficiently heavy tails. Indeed, it can be shown that for scale parameters $b > 0$, the first and second derivatives of $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$ scale linearly with b. Consider a measurement $0 \neq y \in \mathcal{Y}_{\mathrm{gp}}$ with $\operatorname{sign}(y(x_d)) = \operatorname{sign}(y(x_{d'}))$, $\forall d, d' \in D$. Analyzing such y is relevant, as we are most interested in comparing how multiple *reinforcing* observations at clustered locations and a single isolated observation are absorbed during inference. By definition of $\mathcal{Y}_{\mathrm{gp}}$, for $d^* = \operatorname{argmax}_{d \in D}|y_d|$ we have $|y_{d^*}| < |y_s|$ as long as $n > 2$. The corresponding element $y' = G_b^{-1}(\Phi_{0,\sigma^2}(y)) \in \mathcal{Y}_{\mathrm{hp}}$ then satisfies

$$|y'(x_s)| = \left|G_b^{-1}(\Phi_{0,\sigma^2}(y(x_s)))\right| > \left|\frac{G_b^{-1}(\Phi_{0,\sigma^2}(y(x_{d^*})))}{y(x_{d^*})}y(x_s)\right| = \left|\frac{y'(x_{d^*})}{y(x_{d^*})}y(x_s)\right|. \quad (8)$$

Thus HPR inference leads to identical predictive distributions in f-space at the two locations even though the isolated observation $y'(x_s)$ has disproportionately larger magnitude than $y'(x_{d^*})$, relative to the GPR measurements $y(x_s)$ and $y(x_{d^*})$. As this statement holds for any $y \in \mathcal{Y}_{\mathrm{gp}}$ satisfying our earlier sign requirement, it indicates that HPR systematically shrinks isolated observations more strongly than GPR. Since the second derivative of $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$ scales linearly with scale $b > 0$, an intuitive connection suggests itself when looking at inequality (8): the heavier the marginal tails, the stronger the inequality and thus the stronger the selective shrinkage effect.

The previous derivation exemplifies in an idealized setting that HPR leads to improved shrinkage of predictive distributions near isolated observations. More generally, because GPR transforms measurements only linearly, while HPR additionally pre-transforms measurements nonlinearly, our analysis suggests that for any GPR we can find an HPR model which leads to stronger selective shrinkage. The result has intuitive parallels to the parametric case: just as ℓ_1-regularization improves shrinkage of parametric estimators, heavy-tailed processes improve shrinkage of nonparametric estimators. We note that although our analysis kept $K(X, X)$ fixed for GPR and HPR, in practice we are free to tune the kernel to yield a desired scale of predictive distributions. The above analysis has been carried out for regression, but motivates us to now explore heavy-tailed processes in the classification case.

5 Heavy-tailed process classification

The derivation of *heavy-tailed process classification* (HPC) is similar to that of standard multiclass GPC with Laplace approximation in Rasmussen and Williams [12]. However, due to the nonlinear transformations involved, some nice properties of their derivation are lost. We revert notation and let y denote a vector of class labels. For a C-class classification problem with n training points we

4

introduce a vector of nC latent function measurements $(f_1^1, \ldots, f_n^1, f_1^2, \ldots, f_n^2, \ldots, f_1^C, \ldots, f_n^C)^\top$. For each block $c \in \{1, \ldots, C\}$ of n variables we define an independent heavy-tailed process prior using Eq. (4) with kernel matrix K_c. Equivalently, we can define the prior jointly on f by letting K be a block-diagonal kernel matrix with blocks $K_1, \ldots, K_C$. Each kernel matrix K_c is defined by a (possibly different) symmetric positive semidefinite kernel with its own set of parameters. The following construction relaxes the earlier condition that $\text{diag}\,(K) = \sigma^2 \mathbf{1}$ and instead views $\Phi_{0,\sigma^2}(\cdot)$ as some nonlinear transformation with parameter σ^2. By this relaxation we effectively adopt Liu et al.'s [9] interpretation that Eq. (4) defines the copula. The scale parameters b could in principle vary across the nC variables, but we keep them constant at least within each block of n. Labels y are represented in a 1-of-n form and generated by the following observation model

$$p(y_i^c = 1|f_i) = \pi_i^c = \frac{\exp\{f_i^c\}}{\sum_{c'} \exp\{f_i^{c'}\}}. \tag{9}$$

For inference we are ultimately interested in computing

$$p(y_*^c = 1|X, y, x_*) = \mathbb{E}_{p(f_*|X,y,x_*)} \left(\frac{\exp\{f_*^c\}}{\sum_{c'} \exp\{f_*^{c'}\}} \right), \tag{10}$$

where $f_* = (f_*^1, \ldots, f_*^C)^\top$. The previous section motivates that improved selective shrinkage will occur in $p(f_*|X, y, x_*)$, provided the prior marginals have sufficiently heavy tails.

5.1   Inference

As in GPC, most of the intractability lies in computing the predictive distribution $p(f_*|X, y, x_*)$. We use the Laplace approximation to address this issue: a Gaussian approximation to $p(z|X, y)$ is found and then combined with the Gaussian $p(z_*|X, z, x_*)$ to give us an approximation to $p(z_*|X, y, x_*)$. This is then transformed to a (typically non-Gaussian) distribution in f-space using a change of variables. Hence we first seek to find a mode and corresponding Hessian matrix of the log posterior $\log p(z|X, y)$. Recalling the relation $f = G_b^{-1}(\Phi_{0,\sigma^2}(z))$, the log posterior can be written as

$$J(z) \triangleq \log p(y|z) + \log p(z) = y^\top f - \sum_i \log \sum_c \exp\{f_i^c\}\} - \frac{1}{2} z^\top K^{-1} z - \frac{1}{2} \log |K| + \text{const.}$$

Let Π be an $nC \times n$ matrix of stacked diagonal matrices $\text{diag}\,(\pi^c)$ for n-subvectors π^c of π. With $W = \text{diag}\,(\pi) - \Pi\Pi^\top$, the gradients are

$$\nabla J(z) = \text{diag}\left(\frac{df}{dz}\right)(y - \pi) - K^{-1}z$$

$$\nabla^2 J(z) = \text{diag}\left(\frac{d^2 f}{dz^2}\right)\text{diag}\,(y - \pi) - \text{diag}\left(\frac{df}{dz}\right)W\text{diag}\left(\frac{df}{dz}\right) - K^{-1}.$$

Unlike in Rasmussen and Williams [12], $-\nabla^2 J(z)$ is not generally positive definite owing to its first term. For that reason we cannot use a Newton step to find the mode and instead resort to a simpler gradient method. Once the mode $\hat{z}$ has been found we approximate the posterior as

$$p(z|X, y) \approx q(z|X, y) = \mathcal{N}\left(\hat{z}, -\nabla^2 J(\hat{z})^{-1}\right),$$

and use this to approximate the predictive distribution by

$$q(z_*|X, y, x_*) = \int p(z_*|X, z, x_*) q(z|X, y) df.$$

Since we arranged for both distributions in the integral to be Gaussian, the resulting Gaussian can be straightforwardly evaluated. Finally, to approximate the one-dimensional integral with respect to $p(f_*|X, y, x_*)$ in Eq. (10) we could either use a quadrature method, or generate samples from $q(z_*|X, y, x_*)$, convert them to f-space using $G_b^{-1}(\Phi_{0,\sigma^2}(\cdot))$ and then approximate the expectation by an average. We have compared predictions of the latter method with those of a Gibbs sampler; the Laplace approximation matched Gibbs results well, while being much faster to compute.

5

(a)   (b)
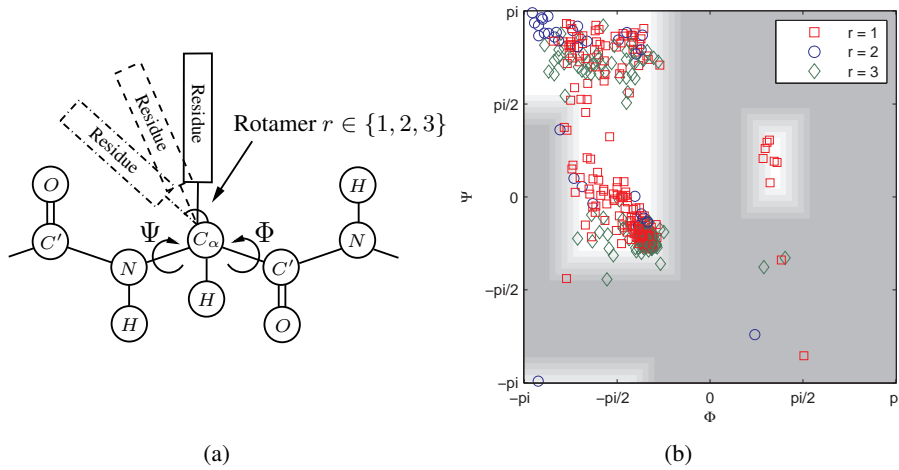
Figure 2: (a) Schematic of a protein segment. The backbone is the sequence of C', N, C_α, C', N atoms. An amino-acid-specific sidechain extends from the C_α atom at one of three discrete angles known as "rotamers." (b) Ramachandran plot of 400 (Φ, Ψ) measurements and corresponding rotamers (by shapes/colors) for amino-acid arginine (`arg`). The dark shading indicates the sparse region we considered in producing results in Figure 3. Progressively lighter shadings indicate how the sparse region was grown to produce Figure 4.

5.2 Parameter estimation

Using a derivation similar to that in [12], we have for $\hat{f} = G_b^{-1}(\Phi_{0,\sigma^2}(\hat{z}))$ that the Laplace approximation of the marginal log likelihood is

$$\log p(y|x) \approx \log q(y|x) = J(\hat{z}) - \frac{1}{2}\log|-2\pi\nabla^2 J(\hat{z})| \tag{11}$$

$$= y^\top \hat{f} - \sum_i \log \sum_c \exp\left\{\hat{f}_i^c\right\} - \frac{1}{2}\hat{z}^\top K^{-1}\hat{z} - \frac{1}{2}\log|K| - \frac{1}{2}\log|-\nabla^2 J(\hat{z})| + \text{const}.$$

We optimize kernel parameters θ by taking gradient steps on $\log q(y|x)$. The derivative needs to take into account that perturbing the parameters can also perturb the mode $\hat{z}$ found for the Laplace approximation. At an optimum $\nabla J(\hat{z})$ must be zero, so that

$$\hat{z} = K\,\text{diag}\left(\frac{d\hat{f}}{d\hat{z}}\right)(y - \hat{\pi}), \tag{12}$$

where $\hat{\pi}$ is defined as in Eq. (9) but using $\hat{f}$ rather than f. Taking derivatives of this equation allows us to compute the gradient $d\hat{z}/d\theta$. Differentiating the marginal likelihood we have

$$\frac{d\log q(y|x)}{d\theta} = (y - \hat{\pi})^\top \text{diag}\left(\frac{d\hat{f}}{d\hat{z}}\right)\frac{d\hat{z}}{d\theta} - \frac{d\hat{z}}{d\theta}K^{-1}\hat{z} + \frac{1}{2}\hat{z}^\top K^{-1}\frac{dK}{d\theta}K^{-1}\hat{z} -$$

$$\frac{1}{2}\text{tr}\left(K^{-1}\frac{dK}{d\theta}\right) - \frac{1}{2}\text{tr}\left(\nabla^2 J(\hat{z})^{-1}\frac{d\nabla^2 J(\hat{z})}{d\theta}\right).$$

The remaining gradient computations are straightforward, albeit tedious. In addition to optimizing the kernel parameters, it may also be of interest to optimize the scale parameter b of marginals G_b. Again, differentiating Eq. (12) with respect to b allows us to compute $d\hat{z}/db$. We note that when perturbing b we change $\hat{f}$ by changing the underlying mode $\hat{z}$ as well as by changing the parameter b which is used to compute $\hat{f}$ from $\hat{z}$. Suppressing the detailed computations, the derivative of the marginal log likelihood with respect to b is

$$\frac{d\log q(y|x)}{db} = (y - \hat{\pi})^\top\frac{d\hat{f}}{db} - \frac{d\hat{z}}{db}^\top K^{-1}\hat{z} - \frac{1}{2}\text{tr}\left(\nabla^2 J(\hat{z})^{-1}\frac{d\nabla^2 J(\hat{z})}{db}\right).$$

6
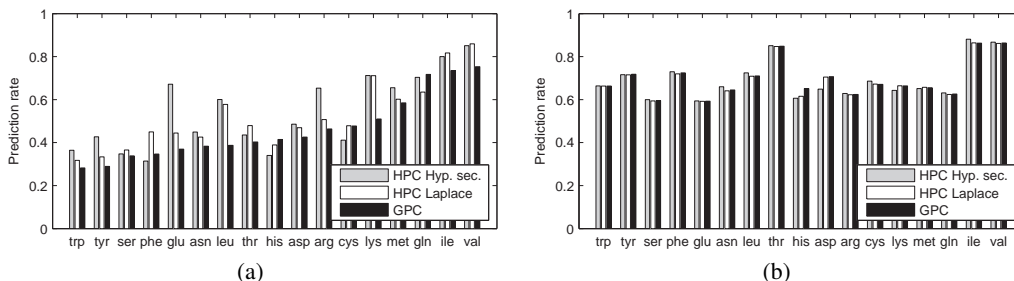
Figure 3: Rotamer prediction rates in percent in (a) sparse and (b) dense regions. Both flavors of HPC (hyperbolic secant and Laplace marginals) significantly outperform GPC in sparse regions while performing competitively in dense regions.

6 Experiments

To a first approximation, the three-dimensional structure of a folded protein is defined by pairs of continuous backbone angles (Φ, Ψ), one pair for each amino-acid, as well as discrete angles, so-called rotamers, that define the conformations of the amino-acid sidechains that extend from the backbone. The geometry is outlined in Figure 2(a). There is a strong dependence between backbone angles (Φ, Ψ) and rotamer values; this is illustrated in the "Ramachandran plot" shown in Figure 2(b), which plots the backbone angles for each rotamer (indicated by the shapes/colors). The dependence is exploited in computational approaches to protein structure prediction, where estimates of rotamer probabilities given backbone angles are used as one term in an energy function that models native protein states as minima of the energy. Poor estimates of rotamer probabilities in sparse regions can derail the prediction procedure. Indeed, sparsity has been a serious problem in state-of-the-art rotamer models based on kernel density estimates (Roland Dunbrack, personal communication). Unfortunately, we have found that GPC is not immune to the sparsity problem. To evaluate our algorithm we consider rotamer-prediction tasks on the 17 amino-acids (out of 20) that have three rotamers at the first dihedral angle along the sidechain[2]. Our previous work thus applies with the number of classes $C = 3$ and the covariates being (Φ, Ψ) angle pairs. Since the input space is a torus we defined GPC and HPC using the following von Mises-inspired kernel for d-dimensional angular data:

$$k(x_i, x_j) = \sigma^2 \exp \left\{ \lambda \left(\left(\sum_{k=1}^{d} \cos(x_{i,k} - x_{j,k}) \right) - d \right) \right\},$$

where $x_{i,k}, x_{j,k} \in [0, 2\pi]$ and $\sigma^2, \lambda \geq 0$[3]. To find good GPC kernel parameters we optimize an ℓ_2-regularized version of the Laplace approximation to the log marginal likelihood reported in Eq. 3.44 of [12]. For HPC we let G_b be either the centered Laplace distribution or the hyperbolic secant distribution with scale parameter b. We estimate HPC kernel parameters as well as b by similarly maximizing an ℓ_2-regularized form of Eq. (11). In both cases we restricted the algorithms to training sets of only 100 datapoints. Since good regularization parameters for the objectives are not known a priori we train with and test them on a grid for each of the 17 rotameric residues in ten-fold cross-validation. To find good regularization parameters for a particular residue we look up that combination which, averaged over the ten folds of the remaining 16 residues, produced the best test results. Having chosen the regularization constants we report average test results computed in ten-fold cross validation.

We evaluate the algorithms on predefined sparse and dense regions in the Ramachandran plot, as indicated by the background shading in Figure 2(b). Across 17 residues the sparse regions usually contained more than 70 measurements (and often more than 150), each of which appears in one of the 10 cross validations. Figure 3 compares the label prediction rates on the dense and sparse

---

[2]Residues alanine and glycine are non-discrete while proline has two rotamers at the first dihedral angle.

[3]The function $\cos(x_{i,k} - x_{j,k}) = [\cos(x_{i.k}), \sin(x_{i,k})][\cos(x_{j.k}), \sin(x_{j,k})]^\top$ is a symmetric positive semi-definite kernel. By Propositions 3.22 (i) and (ii) and Proposition 3.25 in Shawe-Taylor and Cristianini [14], so is $k(x_i, x_j)$ above.

7
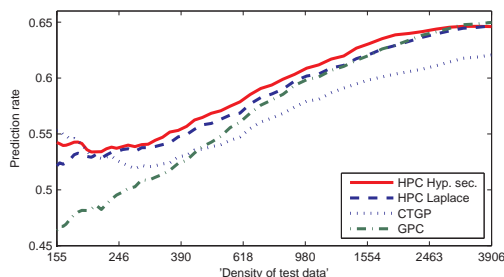
Figure 4: Average rotamer prediction rate in the sparse region for two flavors of HPC, standard GPC well as CTGP [1] as a function of the average number of points per residue in the sparse region.

regions. Averaged over all 17 residues HPC outperforms GPC by 5.79% with Laplace and 7.89% with hyperbolic secant marginals. With Laplace marginals HPC underperforms GPC on only two residues in sparse regions: by 8.22% on glutamine (`gln`), and by 2.53% on histidine (`his`). On dense regions HPC lies within 0.5% on 16 residues and only degrades once by 3.64% on `his`. Using hyperbolic secant marginals HPC often improves GPC by more than 10% on sparse regions and degrades by more than 5% only on cysteine (`cys`) and `his`. On dense regions HPC usually performs within 1.5% of GPC. In Figure 4 we show how the average rotamer prediction rate across 17 residues changes for HPC, GPC, as well as CTGP [1] as we grow the sparse region to include more measurements from dense regions. The growth of the sparse region is indicated by progressively lighter shadings in Figure 2(b). As more points are included the significant advantage of HPC lessens. Eventually GPC does marginally better than HPC and much better than CTGP. The values reported in Figure 3 correspond to the dark shaded region, with an average of 155 measurements.

7 Related research

Copulas [10] allow convenient modelling of multivariate correlation structures as separate from marginal distributions. Early work by Song [16] used the Gaussian copula to generate complex multivariate distributions by complementing a simple copula form with marginal distributions of choice. Popularity of the Gaussian copula in the financial literature is generally credited to Li [8] who used it to model correlation structure for pairs of random variables with known marginals. More recently, the Gaussian process has been modified in a similar way to ours by Snelson et al. [15]. They demonstrate that posterior distributions can better approximate the true noise distribution if the transformation defining the warped process is learned. Jaimungal and Ng [7] have extended this work to model multiple parallel time series with marginally non-Gaussian stochastic processes. Their work uses a "binding copula" to combine several subordinate copulas into a joint model. Bayesian approaches focusing on estimation of the Gaussian copula covariance matrix for a given dataset are given in [4, 11]. Research also focused on estimation in high-dimensional settings [9].

8 Conclusions

This paper analyzed learning scenarios where outliers are observed in the input space, rather than the output space as commonly discussed in the literature. We illustrated heavy-tailed processes as a straightforward extension of GPs and an economical way to improve the robustness of estimators in sparse regions beyond those of GP-based methods. Importantly, because these processes are based on a GP, they inherit many of its favorable computational properties; predictive inference in regression, for instance, is straightforward. Moreover, because heavy-tailed processes have a parsimonious representation, they can be used as building blocks in more complicated models where currently GPs are used. In this way the benefits of heavy-tailed processes extend to any GP-based model that struggles with covariate shift.

Acknowledgements

We thank Roland Dunbrack for helpful discussions and providing access to the rotamer datasets.

8

References

[1] Tamara Broderick and Robert B. Gramacy. Classification and Categorical Inputs with Treed Gaussian Process Models. *Journal of Classification*. To appear.

[2] Wei Chu and Zoubin Ghahramani. Gaussian Processes for Ordinal Regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

[3] Doris Damian, Paul D. Sampson, and Peter Guttorp. Bayesian Estimation of Semi-Parametric Non-Stationary Spatial Covariance Structures. *Environmetrics*, 12:161–178.

[4] Adrian Dobra and Alex Lenkoski. Copula Gaussian Graphical Models. Technical report, Department of Statistics, University of Washington, 2009.

[5] Paul W. Goldberg, Christopher K. I. Williams, and Christopher M. Bishop. Regression with Input-dependent Noise: A Gaussian Process Treatment. In *Advances in Neural Information Processing Systems*, volume 10, pages 493–499. MIT Press, 1998.

[6] Robert B. Gramacy and Herbert K. H. Lee. Bayesian Treed Gaussian Process Models with an Application to Computer Modeling. *Journal of the American Statistical Association*, 2007.

[7] Sebastian Jaimungal and Eddie K. Ng. Kernel-based Copula Processes. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 628–643. Springer-Verlag, 2009.

[8] David X. Li. On Default Correlation: A Copula Function Approach. Technical Report 99-07, Riskmetrics Group, New York, April 2000.

[9] Han Liu, John Lafferty, and Larry Wasserman. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10:1–37, 2009.

[10] Roger B. Nelsen. *An Introduction to Copulas*. Springer, 1999.

[11] Michael Pitt, David Chan, and Robert J. Kohn. Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika*, 93(3):537–554, 2006.

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

[13] Alexandra M. Schmidt and Anthony O'Hagan. Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations. *Journal of the Royal Statistical Society*, 65(3):743–758, 2003. Ser. B.

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

[15] Ed Snelson, Carl E. Rasmussen, and Zoubin Ghahramani. Warped Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 16, pages 337–344, 2004.

[16] Peter Xue-Kun Song. Multivariate Dispersion Models Generated From Gaussian Copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

9