Student-t Processes as Alternatives to Gaussian Processes

Amar Shah
University of Cambridge

Andrew Gordon Wilson
University of Cambridge

Zoubin Ghahramani
University of Cambridge

Abstract

We investigate the Student-t process as an alternative to the Gaussian process as a nonparametric prior over functions. We derive closed form expressions for the marginal likelihood and predictive distribution of a Student-t process, by integrating away an inverse Wishart process prior over the covariance kernel of a Gaussian process model. We show surprising equivalences between different hierarchical Gaussian process models leading to Student-t processes, and derive a new sampling scheme for the inverse Wishart process, which helps elucidate these equivalences. Overall, we show that a Student-t process can retain the attractive properties of a Gaussian process – a nonparametric representation, analytic marginal and predictive distributions, and easy model selection through covariance kernels – but has enhanced flexibility, and predictive covariances that, unlike a Gaussian process, explicitly depend on the values of training observations. We verify empirically that a Student-t process is especially useful in situations where there are changes in covariance structure, or in applications such as Bayesian optimization, where accurate predictive covariances are critical for good performance. These advantages come at no additional computational cost over Gaussian processes.

1 INTRODUCTION

Gaussian processes are rich distributions over functions, which provide a Bayesian nonparametric approach to regression. Owing to their interpretability, non-parametric flexibility, large support, consistency, simple exact learning and inference procedures, and impressive empirical performances [Rasmussen, 1996], Gaussian processes as kernel machines have steadily grown in popularity over the last decade.

At the heart of every Gaussian process (GP) is a parametrized covariance kernel, which determines the properties of likely functions under a GP. Typically simple parametric kernels, such as the Gaussian (squared exponential) kernel are used, and its parameters are determined through marginal likelihood maximization, having analytically integrated away the Gaussian process. However, a fully Bayesian nonparametric treatment of regression would place a nonparametric prior over the Gaussian process covariance kernel, to represent uncertainty over the kernel function, and to reflect the natural intuition that the kernel does not have a simple parametric form.

Likewise, given the success of Gaussian process kernel machines, it is also natural to consider more general families of elliptical processes [Fang et al., 1989], such as Student-t processes, where any collection of function values has a desired elliptical distribution, with a covariance matrix constructed using a kernel.

As we will show, the Student-t process can be derived by placing an inverse Wishart process prior on the kernel of a Gaussian process. Given their intuitive value, it is not surprising that various forms of Student-t processes have been used in different applications [Yu et al., 2007, Zhang and Yeung, 2010, Xu et al., 2011, Archambeau and Bach, 2010]. However, the connections between these models, and the theoretical properties of these models, remain largely unknown. Similarly, the practical utility of such models remains uncertain. For example, Rasmussen and Williams [2006] wonder whether "the Student-t process is perhaps not as exciting as one might have hoped".

In short, our paper answers in detail many of the "what, when and why?" questions one might have about Student-t processes (TPs), inverse Wishart processes, and elliptical processes in general. Specifically:

- We precisely define and motivate the inverse Wishart process [Dawid, 1981] as a prior over covariance matrices of arbitrary size.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

- We propose a Student-t process, which we derive from hierarchical Gaussian process models. We derive analytic forms for the marginal and predictive distributions of this process, and analytic derivatives of the marginal likelihood.

- We show that the Student-t process is the most general elliptically symmetric process with analytic marginal and predictive distributions.

- We derive a new way of sampling from the inverse Wishart process, which intuitively resolves the seemingly bizarre marginal equivalence between inverse Wishart and inverse Gamma priors for covariance kernels in hierarchical GP models.

- We show that the predictive covariances of a TP depend on the values of training observations, even though the predictive covariances of a GP do not.

- We show that, contrary to the Student-t process described in Rasmussen and Williams [2006], an analytic TP noise model can be used which separates signal and noise analytically.

- We demonstrate non-trivial differences in behaviour between the GP and TP on a variety of applications. We specifically find the TP more robust to change-points and model misspecification, to have notably improved predictive covariances, to have useful "tail-dependence" between distant function values (which is orthogonal to the choice of kernel), and to be particularly promising for Bayesian optimization, where predictive covariances are especially important.

We begin by introducing the inverse Wishart process in section 2. We then derive a Student-t process by using an inverse Wishart process over covariance kernels (section 3), and discuss the properties of this Student-t process in section 4. Finally, we demonstrate the Student-t process on regression and Bayesian optimization problems in section 5.

2 INVERSE WISHART PROCESS

In this section we argue that the inverse Wishart distribution is an attractive choice of prior for covariance matrices of arbitrary size. The Wishart distribution is a probability distribution over $\Pi(n)$, the set of real valued, $n \times n$, symmetric, positive definite matrices. Its density function is defined as follows.

Definition. A random $\Sigma \in \Pi(n)$ is *Wishart* distributed with parameters $\nu > n-1$, $K \in \Pi(n)$, and we write $\Sigma \sim \mathrm{W}_n(\nu, K)$ if its density is given by

$$p(\Sigma) = c_n(\nu, K)|\Sigma|^{(\nu-n-1)/2} \exp\Big(-\frac{1}{2}\mathrm{Tr}\big(K^{-1}\Sigma\big)\Big), \tag{1}$$

where $c_n(\nu, K) = \Big(|K|^{\nu/2}2^{\nu n/2}\Gamma_n(\nu/2)\Big)^{-1}$.

The Wishart distribution defined with this parameterization is consistent under marginalization. If $\Sigma \sim \mathrm{W}_n(\nu, K)$, then any $n_1 \times n_1$ principal submatrix Σ_{11} is $\mathrm{W}_{n_1}(\nu, K_{11})$ distributed. This property makes the Wishart distribution appear to be an attractive of prior over covariance matrices. Unfortunately the Wishart distribution suffers a flaw which makes it impractical for nonparametric Bayesian modelling.

Suppose we wish to model a covariance matrix using $\nu^{-1}\Sigma$, so that its expected value $\mathbb{E}[\nu^{-1}\Sigma] = K$, and $\mathrm{var}[\nu^{-1}\Sigma_{ij}] = \nu^{-1}(K_{ij}^2 + K_{ii}K_{jj})$. Since we require $\nu > n-1$, we must let $\nu \to \infty$ to define a process which has positive semidefinite Wishart distributed marginals of arbitrary size. However, as $\nu \to \infty$, $\nu^{-1}\Sigma$ tends to the constant matrix K almost surely. Thus the requirement $\nu > n-1$ prohibits defining a useful process which has Wishart marginals of arbitrary size. Nevertheless, the *inverse Wishart* distribution does not suffer this problem. Dawid [1981] parametrized the inverse Wishart distribution as follows:

Definition. A random $\Sigma \in \Pi(n)$ is *inverse Wishart* distributed with parameters $\nu \in \mathbb{R}_+$, $K \in \Pi(n)$ and we write $\Sigma \sim \mathrm{IW}_n(\nu, K)$ if its density is given by

$$p(\Sigma) = c_n(\nu, K)|\Sigma|^{-(\nu+2n)/2} \exp\Big(-\frac{1}{2}\mathrm{Tr}\big(K\Sigma^{-1}\big)\Big), \tag{2}$$

with $c_n(\nu, K) = \dfrac{|K|^{(\nu+n-1)/2}}{2^{(\nu+n-1)n/2}\Gamma_n((\nu+n-1)/2)}$.

If $\Sigma \sim \mathrm{IW}_n(\nu, K)$, Σ has mean and covariance only when $\nu > 2$ and $\mathbb{E}[\Sigma] = (\nu-2)^{-1}K$. Both the Wishart and the inverse Wishart distributions place prior mass on every $\Sigma \in \Pi(n)$. Furthermore $\Sigma \sim \mathrm{W}_n(\nu, K)$ if and only if $\Sigma^{-1} \sim \mathrm{IW}_n(\nu - n + 1, K^{-1})$.

Dawid [1981] shows that the inverse Wishart distribution defined as above is consistent under marginalization. If $\Sigma \sim \mathrm{IW}_n(\nu, K)$, then any principal submatrix Σ_{11} will be $\mathrm{IW}_{n_1}(\nu, K_{11})$ distributed. Note the key difference in the parameterizations of both distributions: the parameter ν does not need to depend on the size of the matrix in the inverse Wishart distribution. These properties are desirable and motivate defining a process which has inverse Wishart marginals of arbitrary size. Let $\mathcal{X}$ be some input space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive definite kernel function.

Definition. σ is an *inverse Wishart process* on $\mathcal{X}$ with parameters $\nu \in \mathbb{R}_+$ and base kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if

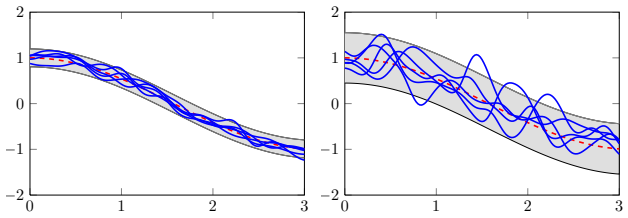

Figure 1: Five samples (blue solid) from $\mathcal{GP}(h, \kappa)$ (left) and $\mathcal{TP}(\nu, h, \kappa)$ (right), with $\nu = 5$, $h(x) = \cos(x)$ (red dashed) and $\kappa(x_i, x_j) = 0.01 \exp(-20(x_i - x_j)^2)$. The grey shaded area represents a 95% predictive interval under each model.

for any finite collection $x_1, ..., x_n \in \mathcal{X}$, $\sigma(x_1, ..., x_n) \sim \text{IW}_n(\nu, K)$ where $K \in \Pi(n)$ with $K_{ij} = k(x_i, x_j)$. We write $\sigma \sim \mathcal{IWP}(\nu, k)$.

In the next section we use the inverse Wishart process as a nonparametric prior over kernels in a hierarchical Gaussian process model.

3 DERIVING THE STUDENT-t PROCESS

Gaussian processes (GPs) are popular nonparametric Bayesian distributions over functions. A thorough guide to GPs has been provided by Rasmussen and Williams [2006]. GPs are characterized by a mean function and a kernel function. Practitioners tend to use parametric kernel functions and learn their hyperparameters using maximum likelihood or sampling based methods. We propose placing an inverse Wishart process prior on the kernel function, leading to a Student-t process.

For a base kernel k_θ parameterized by θ, and a continuous mean function $\phi : \mathcal{X} \to \mathbb{R}$, our generative approach is as follows

$$\begin{aligned} \sigma &\sim \mathcal{IWP}(\nu, k_\theta) \\ y|\sigma &\sim \mathcal{GP}(\phi, (\nu - 2)\sigma)\,. \end{aligned} \tag{3}$$

Since the inverse Wishart distribution is a conjugate prior for the covariance matrix of a Gaussian likelihood, we can analytically marginalize σ in the generative model of (3). For any collection of data $\boldsymbol{y} = (y_1, ..., y_n)^\top$ with $\boldsymbol{\phi} = (\phi(x_1), ..., \phi(x_n))^\top$, $\Sigma = \sigma(x_1, \ldots, x_n)$,

$$\begin{aligned} p(\boldsymbol{y}|\nu, K) &= \int p(\boldsymbol{y}|\Sigma) p(\Sigma|\nu, K) d\Sigma \\ &\propto \int \frac{\exp\left(-\frac{1}{2} Tr\left(\left(K + \frac{(\boldsymbol{y}-\boldsymbol{\phi})(\boldsymbol{y}-\boldsymbol{\phi})^\top}{\nu - 2} \right) \Sigma^{-1} \right) \right)}{|\Sigma|^{(\nu+2n+1)/2}} d\Sigma \\ &\propto \left(1 + \frac{1}{\nu - 2} (\boldsymbol{y} - \boldsymbol{\phi})^\top K^{-1} (\boldsymbol{y} - \boldsymbol{\phi}) \right)^{-(\nu+n)/2} \end{aligned} \tag{4}$$

Definition. $\boldsymbol{y} \in \mathbb{R}^n$ is *multivariate Student-t* distributed with parameters $\nu \in \mathbb{R}_+ \backslash [0, 2]$, $\boldsymbol{\phi} \in \mathbb{R}^n$ and $K \in \Pi(n)$ if it has density

$$\begin{aligned} p(\boldsymbol{y}) = &\frac{\Gamma(\frac{\nu+n}{2})}{((\nu-2)\pi)^{\frac{n}{2}} \Gamma(\frac{\nu}{2})} |K|^{-1/2} \\ &\times \left(1 + \frac{(\boldsymbol{y} - \boldsymbol{\phi})^\top K^{-1} (\boldsymbol{y} - \boldsymbol{\phi})}{\nu - 2} \right)^{-\frac{\nu+n}{2}} \end{aligned} \tag{5}$$

We write $\boldsymbol{y} \sim \text{MVT}_n(\nu, \boldsymbol{\phi}, K)$.

We easily compute the mean and covariance of the MVT using the generative derivation: $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\mathbb{E}[\boldsymbol{y}|\Sigma]] = \boldsymbol{\phi}$ and $\text{cov}[\boldsymbol{y}] = \mathbb{E}[\mathbb{E}[(\boldsymbol{y}-\boldsymbol{\phi})(\boldsymbol{y}-\boldsymbol{\phi})^\top|\Sigma]] = \mathbb{E}[(\nu - 2)\Sigma] = K$. We prove the following Lemma in the supplementary material [Shah et al., 2014].

Lemma 1. *The multivariate Student-t is consistent under marginalization.*

We define a Student-t process as follows.

Definition. f is a *Student-t process* on $\mathcal{X}$ with parameters $\nu > 2$, mean function $\Psi : \mathcal{X} \to \mathbb{R}$, and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if any finite collection of function values have a joint multivariate Student-t distribution, i.e. $(f(x_1), ..., f(x_n))^\top \sim \text{MVT}_n(\nu, \boldsymbol{\phi}, K)$ where $K \in \Pi(n)$ with $K_{ij} = k(x_i, x_j)$ and $\phi \in \mathbb{R}^n$ with $\phi_i = \Phi(x_i)$. We write $f \sim \mathcal{TP}(\nu, \Phi, k)$.

4 TP PROPERTIES & RELATION TO OTHER PROCESSES

In this section we discuss the conditional distribution of the TP, the relationship between GPs and TPs, another covariance prior which leads to the same TP, elliptical processes, and a sampling scheme for the IWP which gives insight into this equivalence. Finally we consider modelling noisy functions with a TP.

4.1 Relation to Gaussian process

The Student-t process generalizes the Gaussian process. A GP can be seen as a limiting case of a TP as shown in Lemma 2, which is proven in the supplementary material.

Lemma 2. *Suppose $f \sim \mathcal{TP}(\nu, \Phi, k)$ and $g \sim \mathcal{GP}(\Phi, k)$. Then f tends to g in distribution as $\nu \to \infty$.*

The ν parameter controls how *heavy tailed* the process is. Smaller values of ν correspond to heavier tails. As ν gets larger, the tails converge to Gaussian tails. This is illustrated in prior sample draws shown in Figure 1. Notice that the samples from the TP tend to have more extreme behaviour than the GP.

ν also controls the nature of the dependence between variables which are jointly Student-t distributed, and

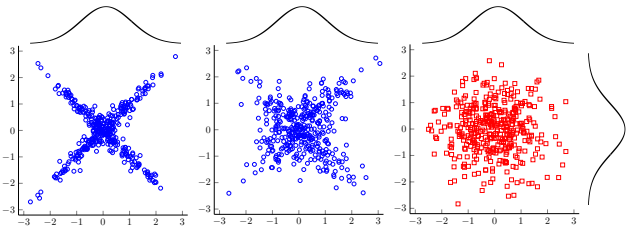

Figure 2: Uncorrelated bivariate samples from a Student-t copula with $\nu = 3$ (left), a Student-t copula with $\nu = 10$ (centre) and a Gaussian copula (right). All marginal distributions are $\mathrm{N}(0, 1)$ distributed.

not just their marginal distributions. In Figure 2 we show plots of samples which all have Gaussian marginals but different joint distributions. Notice how the tail dependency of these distributions is controlled by ν. For example, the dependencies between $y(x_p)$ and $y(x_q)$ are different depending on whether y is a TP or a GP, even if the TP and GP have the same kernel.

4.2 Conditional distribution

The conditional distribution for a multivariate Student-t has an analytic form which we state in Lemma 3 and prove in the supplementary material.

Lemma 3. *Suppose* $\boldsymbol{y} \sim \mathrm{MVT}_n(\nu, \boldsymbol{\phi}, K)$ *and let* $\boldsymbol{y_1}$ *and* $\boldsymbol{y_2}$ *represent the first* n_1 *and remaining* n_2 *entries of* $\boldsymbol{y}$ *respectively. Then*

$$\boldsymbol{y_2}|\boldsymbol{y_1} \sim \mathrm{MVT}_{n_2}\Big(\nu + n_1, \tilde{\boldsymbol{\phi}}_{\boldsymbol{2}}, \frac{\nu + \beta_1 - 2}{\nu + n_1 - 2} \times \tilde{K}_{22}\Big), \quad (6)$$

where $\tilde{\boldsymbol{\phi}}_{\boldsymbol{2}} = K_{21}K_{11}^{-1}(\boldsymbol{y_1} - \boldsymbol{\phi_1}) + \boldsymbol{\phi_2}$, $\beta_1 = (\boldsymbol{y_1} - \boldsymbol{\phi_1})^\top K_{11}^{-1}(\boldsymbol{y_1} - \boldsymbol{\phi_1})$ *and* $\tilde{K}_{22} = K_{22} - K_{21}K_{11}^{-1}K_{12}$. *Note that* $\mathbb{E}[\boldsymbol{y_2}|\boldsymbol{y_1}] = \tilde{\boldsymbol{\phi}}_{\boldsymbol{2}}$, $\mathrm{cov}[\boldsymbol{y_2}|\boldsymbol{y_1}] = \frac{\nu+\beta_1-2}{\nu+n_1-2} \times \tilde{K}_{22}$.

As ν tends to infinity, this predictive distribution tends to a Gaussian process predictive distribution as we would expect given Lemma 2. Perhaps less intuitively, this predictive distribution also tends to a Gaussian process predictive as n_1 tends to infinity.

The predictive mean has the same form as for a Gaussian process, conditioned on having the same kernel k, with the same hyperparameters. The key difference is in the predictive covariance, which now explicitly depends on the training observations. Indeed, a somewhat disappointing feature of the Gaussian process is that for a given kernel, the predictive covariance of new samples does not depend on training observations. Importantly, since the marginal likelihood of the TP in (5) differs from the marginal likelihood of the GP, both the predictive mean and predictive covariance of a TP will differ from that of a GP, after learning kernel hyperparameters.

The scaling constant of the multivariate Student-t predictive covariance has an intuitive explanation. Note that β_1 is distributed as the sum of squares of n_1 independent $\mathrm{MVT}_1(\nu, 0, 1)$ distributions and hence $\mathbb{E}[\beta_1] = n_1$. If the observed value of β_1 is larger than n_1, the predictive covariance is scaled up and vice versa. The magnitude of scaling is controlled by ν.

4.3 Another Covariance Prior

Despite the apparent flexibility of the inverse Wishart distribution, we illustrate in Lemma 4 the surprising result that a multivariate Student-t distribution can be derived using a much simpler covariance prior which has been considered previously [Yu et al., 2007]. The proof can be found in the supplementary material.

Lemma 4. *Let* $K \in \Pi(n)$, $\boldsymbol{\phi} \in \mathbb{R}^n$, $\nu > 2$, $\rho > 0$ *and*

$$r^{-1} \sim \Gamma(\nu/2, \rho/2)$$
$$\boldsymbol{y}|r \sim \mathrm{N}_n(\boldsymbol{\phi}, r(\nu - 2)K/\rho), \quad (7)$$

then marginally $\boldsymbol{y} \sim \mathrm{MVT}_n(\nu, \boldsymbol{\phi}, K)$.

From (7), $r^{-1}|\boldsymbol{y} \sim \Gamma\big(\frac{\nu+n}{2}, \frac{\rho}{2}(1 + \frac{\beta}{\nu-2})\big)$ and hence $\mathbb{E}[(\nu - 2)r/\rho|\boldsymbol{y}] = \frac{\nu+\beta-2}{\nu+n-2}$. This is exactly the factor by which $\tilde{K}_{22}$ is scaled in the MVT conditional distribution in (6).

This result is surprising because we previously integrated over an infinite dimensional nonparametric object (the IWP) to derive the Student-t process, yet here we show that we can integrate over a single scale parameter (inverse Gamma) to arrive at the same marginal process. We provide some insight into why these distinct priors lead to the same marginal multivariate Student-t distribution in section 4.5.

4.4 Elliptical Processes

We now show that both Gaussian and Student-t processes are *elliptically symmetric*, and that the Student-t process is the more general elliptical process.

Definition. $\boldsymbol{y} \in \mathbb{R}^n$ is *elliptically symmetric* if and only if there exists $\boldsymbol{\mu} \in \mathbb{R}^n$, R a nonnegative random variable, Ω a $n \times d$ matrix with maximal rank d and $\boldsymbol{u}$ uniformly distributed on the unit sphere in $\mathbb{R}^d$ independent of R such that $\boldsymbol{y} \overset{\mathcal{D}}{=} \boldsymbol{\mu} + R\Omega\boldsymbol{u}$, where $\overset{\mathcal{D}}{=}$ denotes equality in distribution.

An overview of elliptically symmetric distributions and the following Lemma can be found in Fang et al. [1989].

Lemma 5. *Suppose* $R_1 \sim \chi^2(n)$ *and* $R_2 \sim \Gamma^{-1}(\nu/2, 1/2)$ *independently. If* $R = \sqrt{R_1}$, *then* $\boldsymbol{y}$ *is Gaussian distributed. If* $R = \sqrt{(\nu - 2)R_1R_2}$ *then* $\boldsymbol{y}$ *is MVT distributed.*

Elliptically symmetric distributions characterize a large class of distributions which are unimodal and where the likelihood of a point decreases in its distance from this mode. These properties are natural assumptions we often want to encode in our prior distribution, making elliptical distributions ideal for multivariate modelling tasks. The idea naturally extends to infinite dimensional objects.

Definition. Let $\mathcal{Y} = \{y_i\}$ be a countable family of random variables. It is an *elliptical process* if any finite subset of them are jointly elliptically symmetric.

Not all elliptical distributions have densities (e.g. Lévy, alpha-stable distributions). Even fewer elliptical processes have densities, and the set of those that do is characterized in Theorem 6 due to Kelker [1970].

Theorem 6. *Suppose $\mathcal{Y} = \{y_i\}$ is an elliptical process. Any finite collection $\boldsymbol{z} = \{z_1, ..., z_n\} \subset \mathcal{Y}$ has a density if and only if there exists a non-negative random variable r such that $\boldsymbol{z}|r \sim \mathrm{N}_n(\boldsymbol{\mu}, r\Omega\Omega^\top)$.*

A simple corollary of this theorem describes the only two cases where an elliptical process has an analytically representable density function (its proof is included in the supplementary material).

Corollary 7. *Suppose $\mathcal{Y} = \{y_i\}$ is an elliptical process. Any finite collection $\boldsymbol{z} = \{z_1, ..., z_n\} \subset \mathcal{Y}$ has an analytically representable density if and only if $\mathcal{Y}$ is either a Gaussian process or a Student-t process.*

Since the Student-t process generalizes the Gaussian process, it is the most general elliptical process which has an analytically representable density. The TP is thus an expressive tool for nonparametric Bayesian modelling.

With analytic expressions for the predictive distributions, the same computational costs as a Gaussian process and increased flexibility, the Student-t process can be used as a drop-in replacement for a Gaussian process in many applications.

4.5 A New Way to Sample the IWP

We show that the density of an inverse Wishart distribution depends only on the eigenvalues of a positive definite matrix. To the best of our knowledge this change of variables has not been computed previously. This decomposition offers a novel way of sampling from an inverse Wishart distribution and insight into why the Student-t process can be derived using an inverse Gamma or an inverse Wishart process covariance prior.

Let $\Xi(n)$ be the set of all $n \times n$ orthogonal matrices. A matrix is orthogonal if it is square, real valued and its rows and columns are orthogonal unit vectors. Orthogonal matrices are compositions of rotations and reflections, which are volume preserving operations. Symmetric positive definite (SPD) matrices can be represented through a diagonal and an orthogonal matrix:

Theorem 8. *Let $\Sigma \in \Pi(n)$, the set of SPD, $n \times n$ matrices. Suppose $\{\lambda_1, ..., \lambda_n\}$ are the eigenvalues of Σ. There exists $Q \in \Xi(n)$ such that $\Sigma = Q\Lambda Q^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$.*

Now suppose $\Sigma \sim \mathrm{IW}_n(\nu, I)$. We compute the density of an IW using the representation in Theorem 8, being careful to include the Jacobian of the change of variable, $J(\Sigma; Q, \Lambda)$, given in Edelman and Rao [2005]. From (2) and using the facts that $Q^\top Q = I$ and $|AB| = |BA|$,

$$\begin{aligned} p(\Sigma)d\Sigma &= p(Q\Lambda Q^\top)|J(\Sigma; Q, \Lambda)|d\Lambda dQ \\ &\propto |Q\Lambda Q^\top|^{-(\nu+2n)/2} \exp\Big(-\frac{1}{2}\mathrm{Tr}\big((Q\Lambda Q^\top)^{-1}\big)\Big) \\ &\quad \times \Big|Q^\top \prod_{1 \le i < j \le n} |\lambda_i - \lambda_j|\Big| d\Lambda dQ \\ &\propto \prod_{i=1}^{n} \Big(\lambda_i^{-\frac{\nu+2n}{2}} e^{-\frac{1}{2\lambda_i}} \prod_{j \neq i} |\lambda_i - \lambda_j|^n d\lambda_i\Big)\, dQ \end{aligned} \tag{8}$$

(8) tells us that Q is uniformly distributed over $\Xi(n)$ (e.g. from a $\Upsilon_{n,n}$ distribution as described in Dawid [1977]) and that the λ_i are exchangeable, i.e., permuting the $\mathrm{diag}(\Lambda)$ does not affect its probability. We denote this exchangeable distribution $\Theta_n(\nu)$. We generate a draw from an inverse Wishart distribution by sampling $Q \sim \Upsilon_{n,n}$, $\Lambda \sim \Theta_n(\nu)$ and setting $\Sigma = Q\Lambda Q^\top$.

This result provides a geometric interpretation of what a sample from $\mathrm{IW}_n(\nu, I)$ looks like. We first uniformly at random pick an orthogonal set of basis vectors in $\mathbb{R}^n$ and then stretch these basis vectors using an exchangeable set of scalar random variables. An analogous interpretation holds for the Wishart distribution.

Recall from Lemma 5 that if $\boldsymbol{u}$ is uniformly distributed on the unit sphere in $\mathbb{R}^n$ and $R \sim \chi^2(n)$ independently, then $\sqrt{R}\boldsymbol{u} \sim \mathrm{N}_n(0, I)$. By (4) and Lemma 5, if we sample Q and Λ from the generative process above, then $\sqrt{(\nu-2)R}Q\Lambda^{1/2}\boldsymbol{u}$ is marginally a draw from $\mathrm{MVT}(\nu, 0, I)$. Since the diagonal elements of Λ are exchangeable, Q is orthogonal and sampled uniformly over $\Xi(n)$, and $\boldsymbol{u}$ is spherically symmetric, we must have that $Q\Lambda^{1/2}\boldsymbol{u} \overset{\mathcal{D}}{=} \sqrt{R'}\boldsymbol{u}$ for some positive scalar random variable R' by symmetry. By Lemma 5 we know $R' \sim \Gamma^{-1}(\nu/2, 1/2)$. In summary, the action of $Q\Lambda^{1/2}$ on $\boldsymbol{u}$ is equivalent in distribution to a rescaling by an inverse Gamma variate.

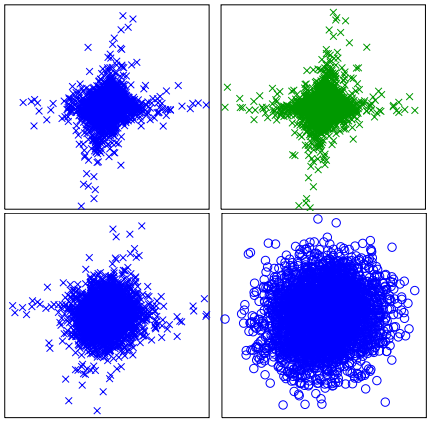

Figure 3: Scatter plots of points drawn from various 2-dim processes. Here $\nu = 2.1$ and $K_{ij} = 0.8\delta_{ij} + 0.2$. **Top-left**: $\mathrm{MVT}_2(\nu, 0, K) + \mathrm{MVT}_2(\nu, 0, 0.5I)$. **Top-right**: $\mathrm{MVT}_2(\nu, 0, K + 0.5I)$ (our model). **Bottom-left**: $\mathrm{MVT}_2(\nu, 0, K) + \mathrm{N}_2(0, 0.5I)$. **Bottom-right**: $\mathrm{N}_2(0, K + 0.5I)$.

4.6 Modelling Noisy Functions

It is common practice to assume that outputs are the sum of a latent Gaussian process and independent Gaussian noise. Such a model is analytically tractable, since Gaussian distributions are closed under addition. Unfortunately the Student-t distribution is not closed under addition.

This problem was encountered by Rasmussen and Williams [2006], who went on to dismiss the multivariate Student-t process for practical purposes. Our approach is to incorporate the noise into the kernel function, for example, letting $k = k_\theta + \delta$, where k_θ is a parametrized kernel and δ is a diagonal kernel function. Such a model is not equivalent to adding independent noise, since the scaling parameter ν will have an effect on the squared-exponential kernel as well as the noise kernel. Zhang and Yeung [2010] propose a similar method for handling noise; however, they incorrectly assume that the latent function and noise are independent under this model. The noise will be uncorrelated with the latent function, but not independent.

As $\nu \to \infty$ this model tends to a GP with independent Gaussian noise. In Figure 3, we consider bivariate samples from a TP when ν is small and the signal to noise ratio is small. Here we see that the TP with noise incorporated into its kernel behaves similarly to a TP with independent Student-t noise.

There have been several attempts to make GP regression robust to heavy tailed noise that rely on approximate inference [Neal, 1997, Vanhatalo et al., 2009]. It is hence attractive that our proposed method can model heavy tailed noise whilst retaining an analytic inference scheme. This is a novel finding to the best of our knowledge.

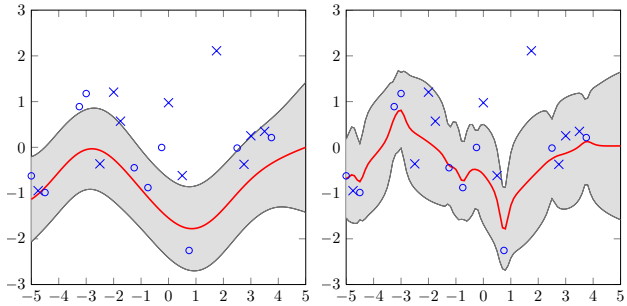

Figure 4: Posterior distributions of 1 sample from Synthetic Data B under GP prior (left) and TP prior (right). The solid line is the posterior mean, the shaded area represents a 95% predictive interval, circles are training points and crosses are test points.

5 APPLICATIONS

In this section we compare TPs to GPs for regression and Bayesian optimization.

5.1 Regression

Consider a set of observations $\{x_i, y_i\}_{i=1}^n$ for $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$. Analogous to Gaussian process regression, we assume the following generative model

$$\begin{aligned} f &\sim \mathcal{TP}(\nu, \Phi, k_\theta) \\ y_i &= f(x_i) \quad \text{for } i = 1, ..., n. \end{aligned} \tag{9}$$

In this work we consider parametric kernel functions. A key task when using such kernels is in learning the parameters of the chosen kernel, which are called the hyperparameters of the model. We include derivatives of the marginal log likelihood of the TP with respect to the hyperparameters in the supplementary material.

5.1.1 Experiments

We test the Student-t process as a regression model on a number of datasets. We sample hyperparameters using Hamiltonian Monte Carlo [Neal, 2011] and use a kernel function which is a sum of a squared exponential and a delta kernel function ($k_\theta = k_{\mathrm{SE}}$). The results for all of these experiments are summarized in Table 1.

Synthetic Data A. We sample 100 functions from a GP prior with Gaussian noise and fit both GPs and TPs to the data with the goal of predicting test points. For each function we train on 80 data points and test on 20. The TP, which generalizes the GP, has superior predictive uncertainty in this example.

Synthetic Data B. We construct data by drawing 100 functions from a GP with a squared exponential

Table 1: Predictive Mean Squared Errors (MSE) and Log Likelihoods (LL) of regression experiments. The TP consistently has the lowest MSE and highest LL.

Data set	Gaussian Process		Student-T Process	
	MSE	LL	MSE	LL
Synth A	2.24 ± 0.09	-1.66± 0.04	2.29 ± 0.08	**-1.00± 0.03**
Synth B	9.53 ± 0.03	-1.45± 0.02	**5.69 ± 0.03**	**-1.30± 0.02**
Snow	10.2 ± 0.08	4.00 ± 0.12	10.5 ± 0.07	**25.7 ± 0.18**
Spatial	6.89 ±0.04	4.34±0.22	**5.71 ±0.03**	**44.4±0.4**
Wine	4.84 ± 0.08	-1.4 ± 1	**4.20 ± 0.06**	**113 ± 2**

kernel and adding Student-t noise independently. The posterior distribution of one sample is shown in Figure 4. The predictive means are also not identical since the posterior distributions of the hyperparameters differ between the TP and the GP. Here the TP has a superior predictive mean, since after hyperparameter training it is better able to model Student-t noise, as well as better predictive uncertainty.

Whistler Snowfall Data[1]**.** Daily snowfall amounts in Whistler have been recorded for the years 2010 and 2011. This data exhibits clear changepoint type behaviour due to seasonality which the TP handles much better than the GP.

Spatial Interpolation Data[2]**.** This dataset contains rainfall measurements at 467 (100 observed and 367 to be estimated) locations in Switzerland on 8 May 1986.

Wine Data. This dataset due to Cortez et al. [2009] consists of 12 attributes of various red wines including acidity, density, pH and alcohol level. Each wine is given a corresponding quality score between 0 and 10. We choose a random subset of 400 wines: 360 for training and 40 for testing.

5.2 Bayesian Optimization

Machine learning algorithms often require tuning parameters, which control learning rates and abilities, via optimizing an objective function. One can model this objective function using a Gaussian process, under a powerful iterative optimization procedure known as Gaussian process Bayesian optimization [Brochu et al., 2010]. To pick where to query the objective function next, one can optimize the expected improvement (EI) over the running optimum, the probability of improving the current best or a GP upper confidence bound.

5.2.1 Method

In this paper we work with the EI criterion and for reasons described in Snoek et al. [2012] we use an ARD

[1]The snowfall dataset can be found at `http://www.climate.weatheroffice.ec.gc.ca`.

[2]The spatial interpolation data can be found at `http://www.ai_geostats.org` under SIC97.

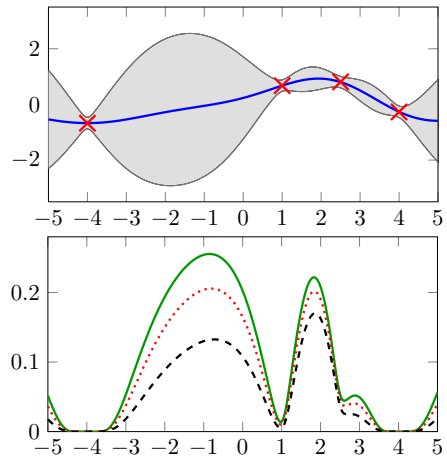

Figure 5: Posterior distribution of a function to maximize under a GP prior (top) and acquisition functions (bottom). The solid green line is the acquisition function for a GP, the dotted red and dashed black lines are for TP priors with $\nu = 15$ and $\nu = 5$ respectively. All other hyperparameters are kept the same.

Matérn 5/2 kernel defined as

$$k_{M52}(\boldsymbol{x}, \boldsymbol{x}') = \theta_0\Big(1 + \sqrt{5r^2_{\boldsymbol{x},\boldsymbol{x}'}}\Big)\exp\Big(-\sqrt{5r^2_{\boldsymbol{x},\boldsymbol{x}'}}\Big) \tag{10}$$

where $r^2(\boldsymbol{x}, \boldsymbol{x}') = \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\theta_d^2}$.

We assume that the function we wish to optimize over is $f : \mathbb{R}^D \to \mathbb{R}$ and is drawn from a multivariate Student-t process with scale parameter $\nu > 2$, constant mean μ and kernel function a linear sum of a ARD Matérn 5/2 kernel and a delta function kernel.

Our goal is to find where f attains its minimum. Let $X_N = \{\boldsymbol{x}_n, f_n\}_{n=1}^N$ be our current set of N observations and $f_{\text{best}} = \min\{f_1, ..., f_N\}$. To compress notation we let $\boldsymbol{\theta}$ represent the parameters θ, ν, μ. Let the acquisition function $a_{\text{EI}}(\boldsymbol{x}; X_N, \boldsymbol{\theta})$ denote the expected improvement over the current best value from choosing to sample at point $\boldsymbol{x}$ given current observations X_N and hyperparameters $\boldsymbol{\theta}$. Note that the distribution of $f(\boldsymbol{x})|X_N, \boldsymbol{\theta}$ is $\text{MVT}_1(\nu + N, \tilde{\mu}(\boldsymbol{x}; X_n), \tilde{\tau}(\boldsymbol{x}; X_n, \nu)^2)$, where the form of $\tilde{\mu}$ and $\tilde{\tau}$ are derived in (6). Let $\tilde{\gamma} = \frac{f_{\text{best}} - \tilde{\mu}}{\tilde{\tau}}$. Then

$$\begin{aligned} a_{\text{EI}}(\boldsymbol{x}; X_N, \boldsymbol{\theta}) &= \mathbb{E}\big[\max\big(f_{\text{best}} - f(\boldsymbol{x}), 0\big)|X_N, \boldsymbol{\theta}\big] \\ &= \int_{-\infty}^{f_{\text{best}}} dy (f_{\text{best}} - y)\frac{1}{\tilde{\tau}}\lambda_{\nu+N}\Big(\frac{y - \tilde{\mu}}{\tilde{\tau}}\Big) \\ &= \tilde{\gamma}\tilde{\tau}\Lambda_{\nu+N}(\tilde{\gamma}) + \tilde{\tau}\Big(1 + \frac{\tilde{\gamma}^2 - 1}{\nu + N - 1}\Big)\lambda_{\nu+N}(\tilde{\gamma}), \end{aligned} \tag{11}$$

where λ_ν and Λ_ν are the density and distribution functions of a $\text{MVT}_1(\nu, 0, 1)$ distribution respectively.

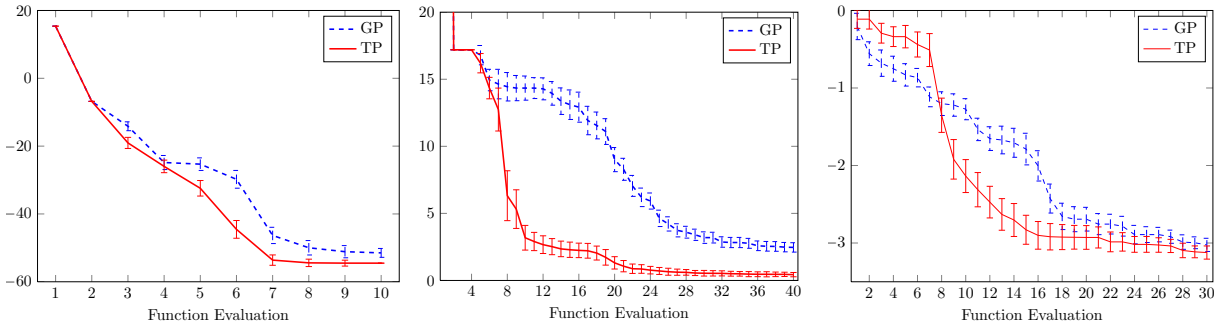

Figure 6: Function evaluations for the synthetic function (left), Branin-Hoo function (centre) and the Hartmann function (right). Evaluations under a Student-t process prior (solid line) and a Gaussian process prior (dashed line) are shown. Error bars represent the standard deviation of 50 runs. In each panel we are minimizing an objective function. The vertical axis represents the running minimum function value.

The parameters $\boldsymbol{\theta}$ are all sampled from the posterior using slice sampling, similar to the method used in Snoek et al. [2012]. Suppose we have H sets of posterior samples $\{\boldsymbol{\theta}_h\}_{h=1}^{H}$. We set

$$\tilde{a}_{\mathrm{EI}}(\boldsymbol{x}; X_N) = \frac{1}{H}\sum_{h=1}^{H} a_{\mathrm{EI}}(\boldsymbol{x}; X_N, \boldsymbol{\theta}_h) \qquad (12)$$

as our approximate marginalized acquisition function. The choice of the net place to sample is $\boldsymbol{x}_{\mathrm{next}} = \mathrm{argmax}_{\boldsymbol{x}\in\mathbb{R}^D} \tilde{a}_{\mathrm{EI}}(\boldsymbol{x}; X_N)$, which we find by using gradient descent based methods starting from a dense set of points in the input space.

To get more intuition on how ν changes the behaviour of the acquisition function, we study an example in Figure 5. Here we fix all hyperparameters other than ν and plot the acquisition functions varying ν. In this example, it is clear that in certain scenarios the TP prior and GP prior will lead to very different proposals given the same information.

5.2.2 Experiments

We compare a TP prior with a Matérn plus a delta function kernel to a GP prior with the same kernel, for Bayesian optimization. To integrate away uncertainty we slice sample the hyperparameters [Neal, 2003]. We consider 3 functions: a 1-dim synthetic sinusoidal, the 2-dim Branin-Hoo function and a 6-dim Hartmann function. All the results are shown in Figure 6.

Sinusoidal synthetic function In this experiment we aimed to find the minimum of $f(x) = -(x-1)^2\sin(3x + 5x^{-1} + 1)$ in the interval $[5, 10]$. The function has 2 local minima in this interval. TP optimization clearly outperforms GP optimization in this problem; the TP was able to come to within 0.1% of the minimum in 8.1 ± 0.4 iterations whilst the GP took 10.7 ± 0.6 iterations.

Branin-Hoo function This function is a popular benchmark for optimization methods [Jones, 2001] and is defined on the set $\{(x_1, x_2) : 0 \le x_1 \le 15, -5 \le x_2 \le 15\}$. We initialized the runs with 4 initial observations, one for each corner of the input square.

Hartmann function This is a function with 6 local minima in $[0, 1]^6$ [Picheny et al., 2013]. The runs are initialised with 6 observations at corners of the unit cube in $\mathbb{R}^6$. Notice that the TP tends to behave more like a step function whereas the Gaussian process' rate of improvement is somewhat more constant. The reason for this behaviour is that the TP tends to more thoroughly explore any modes which it has found, before moving away from these modes. This phenomenon seems more prevalant in higher dimensions.

6 CONCLUSIONS

We have shown that the inverse Wishart process (IWP) is an appropriate prior over covariance matrices of arbitrary size. We used an IWP prior over a GP kernel and showed that marginalizing over the IWP results in a Student-t process (TP). The TP has consistent marginals, closed form conditionals and contains the Gaussian process as a special case. We also proved that the TP is the only elliptical process other than the GP which has an analytically representable density function. The TP prior was applied in regression and Bayesian optimization tasks, showing improved performance over GPs with no additional computational costs.

The take home message for practitioners should be that the TP has many if not all of the benefits of GPs, but with increased modelling flexibility at no extra cost. Our work suggests that it could be useful to replace GPs with TPs in almost any application. The added flexibility of the TP is orthogonal to the choice of kernel, and could complement recent expressive closed form kernels [Wilson and Adams, 2013, Wilson et al., 2013] in future work.

References

C. Archambeau and F. Bach. Multiple Gaussian Process Models. *Advances in Neural Information Processing Systems*, 2010.

E. Brochu, M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Applications to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv*, 2010. URL `http://arxiv.org/abs/1012.2599`.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling Wine Preferences by Data Mining from Physiochemical Properties. *Decision Support Systems, Elsevier*, 47(4):547–553, 2009.

A. P. Dawid. Spherical Matrix Distributions and a Multivariate Model. *J. R. Statistical Society B*, 39: 254–261, 1977.

A. P. Dawid. Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application. *Biometrika*, 68:265–274, 1981.

A. Edelman and N. Raj Rao. Random Matrix Theory. *Acta Numerica*, 1:1–65, 2005.

K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman & Hall, 1989.

D. R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

D. Kelker. Distribution Theory of Spherical Distributions and a Location-Scale Parameter. *Sankhya, Ser. A,*, 32:419–430, 1970.

R. M. Neal. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. *Technical Report No. 9702, Dept of Statistics, University of Toronto*, 1997.

R. M. Neal. Slice Sampling. *Annals of Statistics*, 31(3): 705–767, 2003.

R. M. Neal. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC, 2011.

V. Picheny, T. Wagner, and D. Ginsbourger. A Benchmark of Kriging-Based Infill Criteria for Noisy Optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

C. E. Rasmussen. *Evaluation of Gaussian Processes and other Methods for Non-Linear Regression*. PhD thesis, Graduate Department of Computer Science, University of Toronto, 1996.

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

A. Shah, A. G. Wilson, and Z. Ghahramani. Student-t Processes as Alternatives to Gaussian Processes Supplementary Material. 2014. URL `http://mlg.eng.cam.ac.uk/amar/tpsupp.pdf`.

J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 2012.

J. Vanhatalo, P. Jylanki, and A. Vehtari. Gaussian Process Regression with Student-t Likelihood. *Advances in Neural Information Processing Systems*, pages 1910–1918, 2009.

A. G. Wilson and R. P. Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham. GPatt: Fast multidimensional pattern extrapolation with Gaussian processes. *arXiv preprint arXiv:1310.5288*, 2013. URL `http://arxiv.org/abs/1310.5288`.

Z. Xu, F. Yan, and Y. Qi. Sparse Matrix-Variate t Process Blockmodel. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.

S. Yu, V. Tresp, and K. Yu. Robust Multi-Task Learning with t-Processes. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Y. Zhang and D. Y. Yeung. Multi-Task Learning using Generalized t Process. *Proceedings of the 13th Conference on Artificial Intelligence and Statistics*, 2010.