Flexible Modelling of Genetic Effects on Function-Valued Traits

Nicolo Fusi and Jennifer Listgarten

Microsoft Research, One Memorial Drive, Cambridge, MA, USA
{fusi,jennl}@microsoft.com

Abstract. Genome-wide association studies commonly examine one trait at a time. Occasionally they examine several related traits with the hopes of increasing power; in such a setting, the traits are not generally smoothly varying in any way such as time or space. However, for function-valued traits, the trait is often smoothly-varying along the axis of interest, such as space or time. For instance, in the case of longitudinal traits like growth curves, the axis of interest is time; for spatially-varying traits such as chromatin accessibility it would be position along the genome. Although there have been efforts to perform genome-wide association studies with such function-valued traits, the statistical approaches developed for this purpose often have limitations such as requiring the trait to behave linearly in time or space, or constraining the genetic effect itself to be constant or linear in time. Herein, we present a flexible model for this problem—the Partitioned Gaussian Process—which removes many such limitations and is especially effective as the number of time points increases. The theoretical basis of this model provides machinery for handling missing and unaligned function values such as would occur when not all individuals are measured at the same time points. Further, we make use of algebraic re-factorizations to substantially reduce the time complexity of our model beyond the naive implementation. Finally, we apply our approach and several others to synthetic data before closing with some directions for improved modelling and statistical testing.

Keywords: Genome-wide association study · Longitudinal traits · Time-series traits · Functional traits · Function-valued traits · Linear mixed models · Gaussian process regression · Radial basis function

1 Introduction

Genome-wide association studies commonly examine one trait at a time. Occasionally they examine several related traits with the hopes of increasing power; in such a setting, the traits are not generally smoothly varying in any way such as time or space. However, with the advent of wearables for health and the "quantified self" movement; the broad deployment of cheap sensors in domains such as agriculture and breeding; and the approaching ubiquity of electronic health records, we shall soon see the ubiquity of function-valued traits. Longitudinal

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 95–110, 2016.
DOI: 10.1007/978-3-319-31957-5_7

traits are one example of function-valued traits—traits which can be viewed as a smooth function of some variable. For example, that variable could be time in a clinical history corresponding to a longitudinal trait, or it could be position in the genome, corresponding to a spatial trait such as chromatin accessibility [1]. Such function-valued traits offer new opportunities to dissect genetics. However, maximally benefiting from such opportunities requires that the rich, smoothly-varying structure within these traits can be leveraged by the statistical model of choice. Rich trait structure arises from constraints in the physical world such as that time moves forward and is smoothly varying, or that the correlation between positions on the genome is slowly decreasing according to genetic distance on the chromosome. Modelling approaches in these settings should take into account such constraints while still allowing for flexibility in the shapes of the traits. Furthermore, it stands to reason that the genetic effect might alter the functional form of a trait, such as the shape of a growth curve, a pattern of weight gain, bone loss, or electrocardiogram signal. Thus, flexible modelling beyond linear genetic effects is also one of our goals. Figure 1 shows a set of simple canonical traits and genetic effects that we would like to be able to detect. These canonical traits will also serve as the basis of our synthetic experiments for comparing the behaviour of several modelling approaches. In these examples, by design, a genetic effect which is constant or linear in time will fail to properly model the data. Although these traits are rather idealized, they present a good starting point with which to examine the problem.

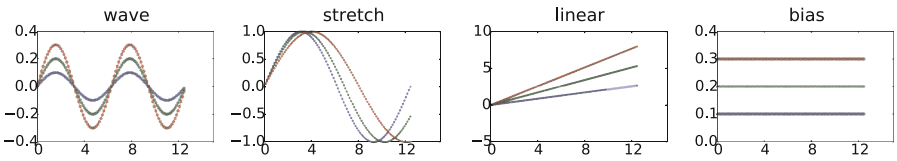


Fig. 1. Simulated traits with 100 time points taking on values uniformly spaced between 0 and 12. Each plot shows what the mean (noise-free) trait looks like for each of the SNP values 0 (blue), 1 (green) and 2 (red). The noise added (not shown) is *iid* with respect to both time and individuals. Note that we hear display the maximum genetic effect for each kind of trait for visual clarity.

The simplest problem one might tackle in our chosen setting is to find out which individual single-nucleotide polymorphisms (SNPs) are correlated to the trait of interest, a so-called marginal test. Those that are correlated are then assumed to have a reasonable probability of being causal for the trait, or of tagging a nearby SNP which is causal for the trait. While it is also of interest to test sets of SNPs jointly [2–4], we here focus on marginal SNP testing, leaving a generalization to set tests for future work. The solution to this marginal testing problem entails (1) proposing a statistical model of the data, and (2) obtaining some weight-of evidence of a genetic effect such as a p-value or Bayes factor. In

this work we focus primarily on the first task but discuss our future directions for the second task in concluding.

Numerous approaches for analyzing function-valued genetic associations have been proposed in recent years [1,4–15]. However, these do not necessarily make effective use of the rich trait structure to increase power because they often assume restrictive forms of the genetic effect or the trait itself. Also, in some cases the statistical efficiency does not scale well with the number of time points, which are expected to be quite numerous in the settings discussed earlier. Next we give a brief overview of some of these approaches and their weaknesses in tackling the kinds of problems we are interested in.

Sikorska *et al.* use an approximate linear mixed model that accounts for correlation in time and assumes that a trait evolves over time in a linear manner; they also assume that the SNP effect itself is additive. Musolf *et al.* first cluster the trait without accounting for genetics and then seek genetic effects on the cluster labels, thereby pre-supposing that all causal SNPs segregate the traits in a similar manner. Shim *et al.* first apply a wavelet-transform to the trait data, thereby transforming the traits to lie in a coordinate system based on (hierarchical) scales and locations; they then perform association testing in this new space. While this approach enables flexible functions of time to be modelled, the SNP effects are restricted to be linear because the wavelet transform itself is linear. Das *et al.* construct a different Legendre polynomial-based model to model the trait for each test SNP allele, learning each model in a largely independent manner. They then test whether the time-specific mean effects are different between the alleles, although it's not clear how they combine time points in their statistical testing framework. Also note that Das *et al.* remove SNPs with minor allele frequency (MAF) less than ten percent from their experiments since the MAF dictates the amount of data available to each allele-specific model. Finally, there has been some related work on detecting differential expression using Gaussian Process regression which shares many aspects of our approach, while differing in several respects including parameter sharing, independence among individuals, and substantial differences in time complexity in the case of aligned time points, partly owing to the use of a different noise model and inference algorithm [16].

In our work, we propose an extremely flexible approach for modelling function-valued traits with genetic effects. In particular, our approach, based on Gaussian Process (GP) regression with a Radial Basis Function (RBF) kernel [17] at its core, can in principle capture any smoothly-varying trait in time, where the smoothness is controlled by a "length scale" parameter. This length scale parameter is estimated using maximum likelihood, thereby effectively deducing the complexity of the trait functional form directly from the data. As for the genetic effect, similarly to Das *et al.*, our model has three components corresponding to three partitions of the data, yielding an extremely non-restrictive class of genetic effects since the GP for each allele can look completely different from the other alleles when no parameters are shared. In our experiments we assume that basic properties such as the noise level and length-scale are likely to be common to all alleles and hence tie these parameters together for more efficient statistical

estimation. However, the model need not be used in this manner. Furthermore, because the RBF kernel effectively integrates out the time points, the number of model parameters does not scale with the number of time points, but is instead fixed–a desirable property when many time points are observed. We call our model the *Partitioned GP* for partitioned Gaussian Process regression.

2 Partitioned Gaussian Processes

As already mentioned our model uses at its core GP regression [17], a class of models which encompasses linear mixed models, the more widely-used concept in genetics [18–21]. The GP regression literature contains results not typically found in the genetics community that we make use of including the use of RBF kernels and Kronecker-product-based refactorizations of matrix-variate normal probability distributions yielding computational efficiencies [22] in the case of aligned and non-missing time points. Also, although we have not yet implemented it, by virtue of using the GP machinery we can immediately access variational approximations to reduce computational time complexity [23,24] in the case of missing data or unaligned time points. We now formally introduce our null model, followed by an exposition of how to do efficient computations in it before introducing the alternative model and computation of p-values.

2.1 Null Model

Our null model, M_0, assumes that the SNP has no effect on the trait (and so does not enter the model), but does capture correlation in time by way of an RBF kernel. Let $\mathbf{Y}$ be the $N \times T$ matrix of traits for N individuals and T time points. Let $\mathbf{W}$ be the $NT \times 1$ times at which the traits were measured, and let $\text{vec}(\mathbf{Y})$ denote the unrolled version of $\mathbf{Y}$ into a vector of dimension $NT \times 1$,

$$\text{vec}(\mathbf{Y}) = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{NT} \end{pmatrix}$$

Then

$$M_0 : p(\text{vec}(\mathbf{Y})) = \mathcal{N}\left(\text{vec}(\mathbf{Y}) \,\middle|\, \mathbf{0}, \sigma_r^2 \mathbf{K}_{RBF}(\mathbf{W}, \mathbf{W}|l) + \sigma_e^2 \mathbf{I}_{NT}\right), \tag{1}$$

where $\mathcal{N}(a \,|\, b, \mathbf{C})$ is a Gaussian distribution in vector a with mean b and covariance $\mathbf{C}$; $\mathbf{I}_{NT}$ is the $NT \times NT$ identity matrix; σ_r^2 and σ_e^2 are scalar parameters which control the overall variance contributed by each kernel; $\mathbf{K}_{RBF}(l)$ is an $NT \times NT$ radial basis function kernel with length-scale parameter l and elements defined by $K_{RBF}(w_{ij}, w_{qp}|l) \equiv exp\left(-\frac{||w_{ij}-w_{qp}||}{2l^2}\right)$. The length-scale parameter determines the overall scale on which the trait varies within an individual. For very rapidly varying traits, it is small, and for slowly varying traits it is large.

The RBF kernel models the dependence in time while the identity kernel models the remaining environmental noise. Note that the RBF kernel here models not only correlation between time points within an individual but also equally across individuals. That is, we make the assumption that the trait at time point t is more correlated across individuals i and j than between time points t and $t + t_0$ for the same person (where t_0 is an offset in time). While at first this may seem a counterintuitive choice, it turns out that for the types of traits we are interested in, it is the correct thing to do. Namely, we are interested in settings in which the traits are the same across all individuals (or later for those with the same genetics), other than by virtue of noise. Examples of such traits are shown in Fig. 1. An example where this is might be a reasonable assumption would be growth curves where on average the curves look the same for a species, but with a particular mutation the curve suddenly changes trajectory. An example where this is an unreasonable assumption would be un-aligned electrocardiographic signals where no two people would in general look the same at time t unless their signals had been re-scaled and aligned. When the assumption of correlation in time between individuals is not believed to be reasonable, one can easily remove this restriction from the model, leaving time correlations only within an individual. In fact, as we explain in the next section, it is algebraically and computationally trivial to make such a change while retaining all efficient computations. However, by removing this assumption from the model one loses statistical power if the assumption is actually valid in the data. In fact, when conducting our synthetic experiments we found that removal of this assumption in the model substantially weakened the results (data not shown).

Note that for simplicity, we assume that covariates such as age and gender have been regressed out of the trait ahead of time, although these could easily be incorporated in to the model, by way of the Gaussian mean (*i.e.*fixed effects). All remaining expositions (other than for the pseudo-inputs and variational inference) can be readily extended to having covariates directly included with no change to the computational time complexity. We make a similar assumption about population structure and family relatedness, which can be regressed out using either principle components [25] or linear mixed models [21], although investigating the best way to do this for function-valued traits is an open area for investigation. Finally, in Eq. 1 we did not assume that traits for each person were measured at the same time points or that no trait values were missing. However, in the next section on efficient computations, we will need to make this assumption. In Sect. 2.4 we outline ways to relax this assumption.

Efficient Computation of the Likelihood. In order to obtain a p-value by way of statistical testing we need to estimate the maximum likelihood parameters of our model over and over, once per genetic marker. Computing the maximum likelihood over and over again for each hypothesis is a non-trivial goal in the sense that general kernel-based methods have time complexity which scales cubically in the dimension of the kernel (here NT), and space complexity which is quadratic in that dimension. However, in some cases, structure in the kernel

can be leveraged to gain substantial speed-ups (*e.g.* [21]). For Partitioned GPs such structure arises when there is no missing data and all traits are measured at the same time points for all individuals. In this case, the likelihood can be re-written with Kronecker products in the covariance term, yielding dramatically reduced time and space complexities. Later we discuss how to achieve speed-ups in the face of missing or unevenly-spaced time points using the Partitioned GP, which can require some approximations, whereas the present exposition requires no approximation.

The RBF kernel (dimension $NT \times NT$ in Eq. 1) is a specially structured kernel because of the repeating times across individuals. This structure means that we can re-write the Gaussian likelihood in Eq. 1 in matrix-variate form as follows [22],

$$M_0 : p(\mathbf{Y}) = \mathcal{N}\left(\mathbf{Y} \,\middle|\, \mathbf{0}, \sigma_r^2 \mathbf{K}_{RBF}(\mathbf{W}, \mathbf{W}|l) \otimes \mathbf{J}_N + \sigma_e^2 \mathbf{I}_{NT}\right), \tag{2}$$

where here we have overloaded $\mathbf{K}_{RBF}(\mathbf{W}, \mathbf{W}|l)$ to now indicate a $T \times T$ matrix, and where $\mathbf{J}_N$ is the square matrix of all ones of size N. The symbol $\otimes$ denotes the Kronecker product which produces a square matrix of dimensions $ab \times ab$ for $A \otimes B$ if A and B are square matrices of dimension a and b respectively. The computational time complexity of evaluating the likelihood in Eq. 1 is $O(N^3T^3)$ because one must compute the inverse and determinant of the covariance matrix of dimension $NT \times NT$. In contrast, using a spectral-decomposition-based re-factoring [22] of Eq. 2, the computational time complexity can be reduced to $O(T^3)$.[1] In particular, if one defines $\mathbf{U}_r \mathbf{S}_r \mathbf{U}_r^T$ as the spectral decomposition of the $T \times T$ matrix $\mathbf{K}_{RBF}(l)$, and $\mathbf{U}_j \mathbf{S}_j \mathbf{U}_j^T$ as the spectral decomposition of $\mathbf{J}_N$, then one can write the log likelihood of the null model as follows [22]:

$$\mathcal{L}_0 = -\frac{NT}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{S}_r \otimes \mathbf{S}_j| - \frac{1}{2}\mathrm{vec}(\mathbf{U}_r^T \mathbf{Y} \mathbf{U}_j)^T (\mathbf{S}_r \otimes \mathbf{S}_j)^{-1} \mathrm{vec}(\mathbf{U}_r^T \mathbf{Y} \mathbf{U}_j). \tag{3}$$

It is also easy to generalize this expression and its derivative when the mean of the Gaussian is non-zero; we do so to make one of the models we compare against (Furlotte *et al.*) significantly faster than in their original presentation (they could not do the same because they jointly model population structure) [5].

Note that the individuals are not identically and independently distributed (*iid*) in our null model because of the term $\mathbf{J}_N$. If we were to replace $\mathbf{J}_N$ with the identity matrix, then the individuals would be iid, which thus amounts to relaxing the assumption mentioned in the introduction wherein time points across individuals are correlated.

As described earlier, we have assumed that population structure and family structure have already been accounted for, but these could instead be incorporated in to the model by adding to $\mathbf{J}_N$ a genetic similarity matrix [21], incurring a time complexity of $O(N^3 + T^3)$ in the most general case.

[1] If $\mathbf{J}_N$ were an arbitrary matrix the time complexity would be $O(N^3 + T^3)$, but because the spectral decomposition of $\mathbf{J}_N$ can be computed once and cached, the complexity becomes $O(T^3)$. Moreover, because it is an all-ones matrix, its spectral decomposition can be computed more efficiently than in the general case.

For parameter estimation we use gradient descent to obtain the maximum likelihood solution in parameters $l, \sigma_r^2, \sigma_e^2$—all scalars. The reader is referred to Stegle *et al.* for the derivative expressions which have the same time complexity as Eq. 3 [22]. Because the log likelihood is not convex, we use multiple random re-starts, finding empirically that five restarts in our experiments yielded good results.

2.2 Alternative Model

Now that we have fully described the null model and how to efficiently compute its log likelihood, we generalize this model to an alternative model which handles a wide range of genetic effects. To do so, we create a separate GP for each partition of the data, where the partition is defined by the alleles of the test SNP (using whatever encoding of the data one desires, such as a $s = 0, 1, 2$ encoding of the number of mutant alleles across the two chromosomes),

$$M_A : p(\mathbf{Y}) = \sum_{s=1}^{S} \mathcal{N}\left(\mathbf{Y}_s \,\middle|\, \mathbf{0}, \sigma_{r_s}^2 \mathbf{K}_{RBF}(\mathbf{W}, \mathbf{W}|l) \otimes \mathbf{J}_{N_s} + \sigma_e^2 \mathbf{I}_{N_s T}\right), \qquad (4)$$

where S denotes the number of alleles in the SNP encoding, $\mathbf{Y}_s$ is the subset of trait data for which the individual has SNP value s, and where N_s is the number of such individuals. In principle, one could use a different length scale, l and variance parameters σ_e^2 for each partition s, but we have found that in our experiments, tying them together yielded good results and allowed us to test SNPs with much lower MAF owing to the data sharing offered by the shared parameters. While it may seem at first glance that this parameter tying might coerce the trait to look the same across SNP partitions, in fact, we are only coercing broad properties of the trait to be similar, such as the scale on which the signal changes, and only loosely at that. Because GP regression is a non-parametric model, the data itself plays a large role in defining the posterior distribution of functional forms; it is for this reason that our model is able to capture substantially different functional forms even with tied parameters.

The same efficient computations outlined earlier for the null model can just as well be applied to this alternative model, and so the time complexity of computing the alternative model likelihood has as an upper bound that of the null model, which happens only when all individuals are assigned to the same partition. Note too that the null model can computed just once and then cached across all SNPs tested for increased efficiency.

Beyond data sharing across partitions by virtue of shared parameters, the model has good statistical efficiency owing to the fact that GPs operate in the kernel space [17] where the number of parameters does not depend on the number of time points. All in all, we find in our experiments that as few as seven samples per partition appears to be sufficient, which with cohort sizes in the tens if not hundreds of thousands, imposes little restriction on the MAF.

2.3 Hypothesis Testing

Standard frequentist hypothesis testing uses a null model that is nested in the alternative model which then allows one to use a likelihood ratio or score test, for example. However, even when models are nested, these tests require that model assumptions are met, and typically that sample sizes are large enough for asymptotics to be valid. In cases where model or asymptotic assumptions are unmet, one can appeal to various forms of permutation testing to obtain calibrated p-values. Because our models are not nested, we cannot rely on standard theories to compute p-values, and could therefore turn to permutation testing. However, as it turns out, when we apply a standard χ^2 test to generate p-values for our Partitioned GP, we find that our type 1 error is controlled, albeit extremely conservatively even though the assumptions of this test are not here met (see Results). Furthermore, in the discussion, we outline a nested version of the Partitioned GP that we are currently working on.

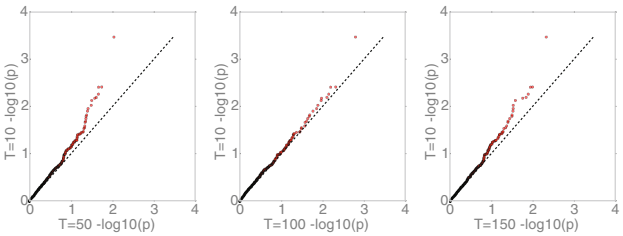


Fig. 2. Paired plot of the $-log$ p-values generated from the null distribution, for 10 time points versus each of 100 and 150 time points.

The precise way in which we apply a standard χ^2 test is that we compute the maximum likelihood of the data under the null and under the alternative models, $\mathcal{L}_A$ and $\mathcal{L}_0$, count the number of degrees of freedom different between them, d, and then apply the standard p-value computation. Our null model has no partitions and has three free scalar parameters: σ_r^2 and l, the overall-variance and length-scale for the time-based kernel, and σ_e^2 for the residual noise. Our alternative model shares all parameters across partitions except for the time-based kernel variances, $\sigma_{r^s}^2$ (one per SNP allele), leading to two more parameters than the null model. We count these two parameters as two extra degrees of freedom even though these parameters are constrained to be greater than zero and so are not truly full degrees of freedom—such miscounting can only lead to overly-conservative p-values in the case of properly nested models. Our test statistic is then twice the difference between the null and alternative maximum log likelihoods, $\Delta \equiv 2(\mathcal{L}_A - \mathcal{L}_0)$, from which we compute a p-value using a χ_d^2 test with $d = 2$ of freedom. While this p-value is uncalibrated, as we shall see in the Results section, it turns out to control type 1 error.

Table 1. Control of type 1 error at significance thresholds α for traits with 10 time points using 390,272 tests. Fraction of p-values below that threshold, with absolute numbers in parentheses.

Model	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$
Partitioned GP	$1.1 \times 10^{-3}(434)$	$6.7 \times 10^{-5}(26)$	0.0(0)	0.0(0)
Inverse K score	$9.1 \times 10^{-3}(3568)$	$8.8 \times 10^{-4}(342)$	$5.6 \times 10^{-5}(22)$	$1.0 \times 10^{-5}(4)$
Inverse linreg	$9.8 \times 10^{-3}(3828)$	$9.4 \times 10^{-4}(366)$	$6.1 \times 10^{-5}(24)$	$1.0 \times 10^{-5}(4)$
Furlotte et al.	$9.2 \times 10^{-3}(3589)$	$9.3 \times 10^{-4}(362)$	$6.1 \times 10^{-5}(24)$	$1.5 \times 10^{-5}(6)$

2.4 Handling Traits with Missing Data or Which are Unevenly-Sampled Across Individual

In a model with a vector Gaussian likelihood, such as Eq. 1, missing trait data can readily be handled by simply removing any rows with missing data, because this procedure is equivalent to marginalization in a Gaussian [17]. In such a manner, if using Eq. 1, one could take T to be the number of uniquely observed time points across all individuals, even if many individuals were missing many of these time points. This procedure could also capture the case where different individuals were measured at different time points. However, in the Kronecker version of the likelihood written for computational efficiency gains (Eq. 2), one can no longer perform this arbitrary marginalization by simply removing an element of the phenotype vector, because with the Kronecker-factorized covariance matrix one would have to either remove all individuals missing a time point, or all time points missing an individual. Therefore, if one wants both computational efficiency and a means to readily marginalize over missing data, one must appeal to alternative formulations and/or approximations. The approach we propose is keep the Gaussian likelihood in vector form, as in Eq. 1, but to augment the model with latent *inducing inputs* [23,24], which are points in time (or space, depending on the type of trait) that are included in the model. Inducing inputs can be thought of as pseudo-observations in time (or space) that are included in the RBF kernel inputs; when conditioned on, these pseudo-observations make any observed data conditionally independent of each other. This has the effect of reducing the time complexity from $O((NT)^3)$ in Eq. 1 to $O(NTQ^2)$ for Q inducing inputs. In such a variational approach, only the number of pseudo-observations need be specified, not the locations, as these are learned as part of the parameter estimation procedure. Also note that if one uses as many pseudo-observations as there are uniquely observed time points, then the algorithm is exact. As a consequence, one could use this approach as an alternative to the efficient Kronecker product approach we described. We have not yet performed experiments with this approach, but these methods are well-studied and their application should be rather direct.

3 Results

As discussed in the introduction, many models have been developed to perform genome-wide association studies with function-valued traits. However, these models tend to have constraints on the type of genetic or time effect that can be recovered (*e.g.*, only constant or linear effect in time, or only linear in the SNP), or are limited to relatively few time points because the number of parameters scales with the number of time points. For our experiments we have chosen a set of baseline models to test particular hypotheses about what kinds of models work and where they fail, in the settings we care about—in particular, exploring what happens when there are a large number of time points such as would be collected be wearables and other sensors. The models we compare and their short-hand notation are:

1. *Partitioned GP*: As described above, using the (exact) Kronecker product implementation.
2. *Furlotte et al.*: A linear mixed model where correlation in time is modelled using an auto-correlation kernel (here we use an RBF as we do with our Partitioned GP), and where in the alternative model, the SNP is a fixed effect, shifting the trait at all time points by the same amount [5]. A standard LRT test is used for the one-degree-of-freedom test. Note that we here do not use the population structure kernel used in [5] as our experiments are not affected by such factors.
3. *Inverse linreg*: To examine how models for which the number of parameters increases with the number of time points, we use inverse linear regression model wherein the SNP is modelled as the dependent variable and each trait in time is an independent variable. Testing is done with a χ^2 test with T degrees of freedom (total number of time points, assumed to be the same for all individuals). Note that in place of inverse linear regression, we could have used inverse multinomial/"soft-max" regression. However, because preliminary results suggested the results were similar, we chose to experiment with only the linear model.
4. *Inverse K score*: This model can be viewed as a Bayesian equivalent to *Inverse linreg* where the time-effects are integrated out, yielding a linear mixed model. In this way, the model does not dependend on the number of time points. We then apply a score test to obtain a p-value (*e.g.* [2]).

We systematically explore each of these approaches on simulated phenotypic data where we know the ground truth, examining type 1 error control, power, and ability to rank hypotheses regardless of calibration. We based our simulated data on the actual SNPs in the CARDIA data set (dbGaP phs000285.v3.p2) which, after filtering out individuals missing more than 10 % of their SNPs, any SNPs missing more than 2 % of individuals, or with MAF less than 5 % left 1,441 individuals with 540,038 SNPs. The only covariate we use is an off-set, which we regress on as a pre-processing step before applying the models.

To simulate time-varying traits, we used a set of canonical functions that were representative of the types of signal we were interested in exploring. In

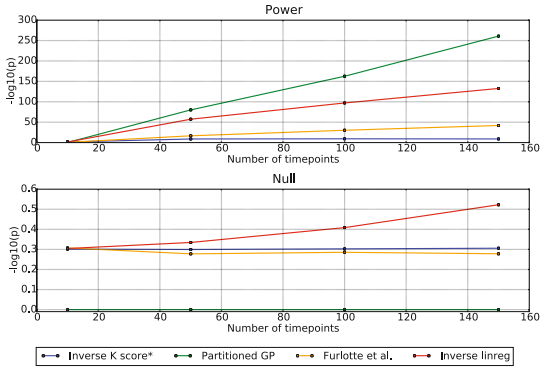


Fig. 3. Power curves as a function of time for all methods. On the vertical axis is the median $-log$ p-value for each method over eight thousand SNP tests. The top plot is for tests with SNP effect, and the lower plot is for those with no SNP effect. *As noted in the main text, the numerical routine used to get p-values for Inverse K score does not yield numeric values less than around 10^{-8}, thereby likely making this method appear worse than it might be; however, we get a better sense of its behaviour in Fig. 5.

particular, we used a *wave*, *linear*, *bias*, and a *stretch* as shown in Fig. 1. For null data, we generated noisy versions of these, where the noise was iid in time and individual. For non-null data we modified the noise-free trait in a smoothly varying way as a function of genotype before adding iid noise. For the wave (a *sin* wave), the amplitude increased as a linear function of the SNP; for the linear (a straight line), the slope changed as a linear function of the SNP; for the bias, the horizontal intercept changed as a linear function of the SNP; for stretch (a *sin* wave), the frequency changed as a linear function of the SNP. We varied both the SNP effect intensities and the amount of noise. One can summarize the strength of the SNP effect at each time point by the fraction of variance explained by the genetic signal at each time point (*i.e.*, the variance of the noiseless trait divided by total variance, all at a given time point) as shown in Fig. 4. Because we were interested specifically in seeing which models could handle many time points, we conducted experiments with 10, 50, 100, and 150 time points.

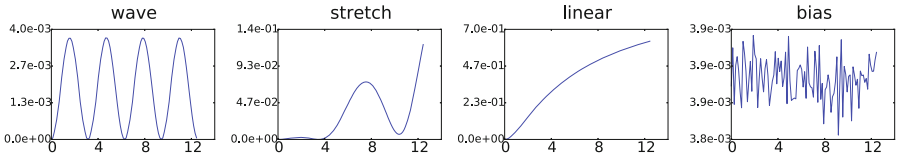


Fig. 4. Average fraction of variance accounted for by genetics at each time point in each canonical function over the range of settings used, for the traits with 100 time points.

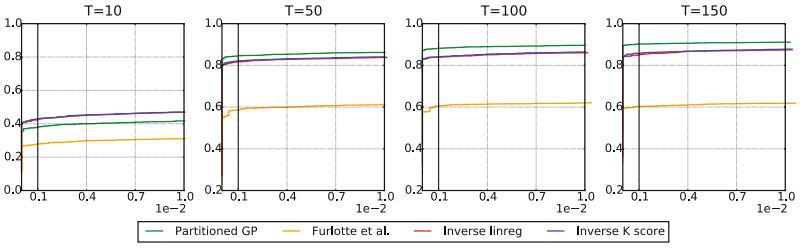
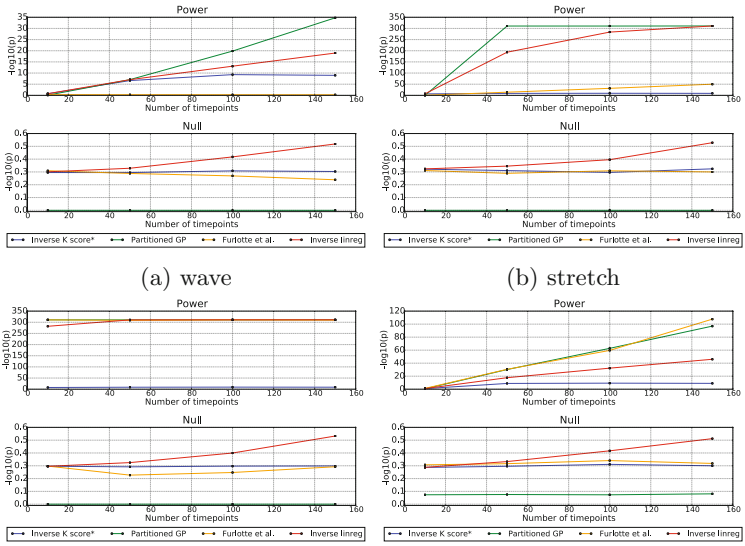


Fig. 5. ROC curves for the simulated data with T equal to 10, 50, 100 and 150 time points, for small False Positive Rates (less than 0.01). The vertical axis shows the False Positive Rate, and the horizontal axis, the True Positive Rate.

Our first goal was to establish whether our Partitioned GP controls type 1 error so that we could use its p-values at face value for power comparisons, even if they are not calibrated. First we used 8,000 tests at each of 10, 50, 100 and 150 time points, finding that the smallest number of time points (10) was always the least conservative (Fig. 2). Therefore, we ran much larger scale simulations of null-only data for 10 time points, obtaining 390,272 test statistics. With just under half a million tests, we had resolution to check for control of type 1 error up to a significance level of $\alpha = 10^{-5}$. As can be seen in Table 1, all methods control the type 1 error up to $\alpha = 10^{-5}$. Note that our method controls the type 1 error extremely conservatively, which could potentially hurt our method in a power comparison. However, as we see next, our method is still the most powerful overall in our experiments.

Having established that our method controls type 1 error, we next set out to see if it had more power to detect associations than the other methods. Figure 3 shows the median test statistic for both our null (lower plot) and non-null (upper plot) experiments, and demonstrates that our methods has maximum power for the traits and methods chosen. Because our type 1 error control experiments only went to $\alpha = 10^{-5}$, we chose to include the lower plot (Null). This null plot shows that while the inverse kernel score remains calibrated, the inverse linear regression becomes substantially inflated, failing to control the type 1 error. Our method, is extremely conservative in controlling the type 1 error, yet maintains maximal power. We also break down these plots by trait type in Fig. 6. Here we see that Furlotte *et al.*, despite only modelling a mean shift in the trait, is able to capture stretch, though not wave, for which the mean between alleles is identical. For stretch and wave, the Partitioned GP is the clear winner, while for linear, all method work equally well, and for stretch, Furlotte *et al.* and the Partitioned GP have the most power.

Note that the inverse kernel score test appears to have terrible power. However, this plot is perhaps misleading in the sense that this method uses a numerical routine (Davies method) which has limited precision, yielding many zeros for tiny p-values (usually those smaller than 10^{-8}). The only way to handle this was either to keep these at zero, which would give that method an unfair advantage, or to replace all zero p-values with 10^{-8}, which is what we chose to do, thereby



(a) wave (b) stretch

(c) linear (d) bias

Fig. 6. Power curves as in Fig. 3, but separated by trait types shown in Fig. 1. *Again, as noted in the main text, the numerical routine used to get p-values for Inverse K score does not yield numeric values less than around 10^{-8}, thereby likely making this method appear worse than it might be; however, we get a better sense of its behaviour in Fig. 5.

showing the model in a worse light with respect to power than we believe it may have if there were a way to compute p-values with more precision. As a consequence, we next investigated the ability of each model to discriminate true nulls from alternatives by using a Receiver Operating Characteristics (ROC) curve—a metric which does not depend in any way on calibration and may be less sensitive to p-value resolution.

Figure 5 shows the ROCs for each method, where we now see that the inverse kernel score test performs extremely well, though not as well as the Partitioned GP. Note that inverse linear regression, though showing inflated test statistics in the lower panel of Fig. 3, here demonstrates that it maintains the ability to properly rank the hypotheses from most to least significant, though again, not as well as the Partitioned GP. Note that the performance of Furlotte *et al.* is not terribly surprising since it is only able to capture shifts in the mean of the functional trait, whereas our simulation scheme is deliberately testing richer SNP effects.

4 Discussion

We have introduced a new method for performing GWAS on function-valued traits. Our model is extremely flexible in its capacity to handle a wide range of

functional forms. This flexibility is achieved by using a non-parametric statistical model based on RBF Gaussian processes. Computations in this model are efficient when time points are aligned and traits are not missing, scaling only cubically with the number of time points as opposed to cubic in the number of time points times individuals, as would be the case in a naive computation. We have also outlined how to do efficient computations even in the presence of missing trait data or unaligned samples. In a comparison against three other models on synthetic data, each with different characteristics and ways of handling the problem, we achieved maximal power, and maximal ability to discriminate null versus alternative tests as judged by an ROC curve. Our model is especially good at handling traits with many time points.

One downside of the model as presented is that the null model is not nested inside the alternative model, making computation of calibrated p-values without permutations most likely impossible. We were able to bypass this issue by demonstrating empirically that naive application of a likelihood ratio test controls the type 1 error, yielding extremely conservative p-values. However, we are currently investigating a version of the Partitioned GP model which has its null model nested in the alternative model and is therefore likely to yield calibrated p-values and therefore potentially a larger power gain. In this model, the partitions of the alternative model are all placed within a single Gaussian, with correlation parameters for each pair of alleles dictating how similar the GP for each allele should be. When these parameters are equal to one, we obtain the present alternative model. When these parameters are zero, we obtain the null model, thereby making it nested inside of the alternative. Other directions of interest are to extend this type of modelling approach to testing sets of SNPs rather than only single SNPs, and to incorporate model-based warping of the phenotype so as to coerce the data to better adhere to the Gaussian residual assumption [26].

Acknowledgments. We thanks to Leigh Johnston, Ciprian Crainiceanu, Bobby Kleinberg and Praneeth Netrapalli for discussion; the anonymous reviewers for helpful feedback, and Carl Kadie for use of his HPC cluster code. Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226.

References

1. Shim, H., Stephens, M.: Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. Ann. Appl. Stat. **9**(2), 665–686 (2015)
2. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., Lin, X.: Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. **86**(6), 929–942 (2010)
3. Listgarten, J., Lippert, C., Kang, E.Y., Xiang, J., Kadie, C.M., Heckerman, D.: A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics **29**(12), 1526–1533 (2013)

4. He, Z., Zhang, M., Lee, S., Smith, J.A., Guo, X., Palmas, W., Kardia, S.L.R., Diez Roux, A.V., Mukherjee, B.: Set-based tests for genetic association in longitudinal studies. Biometrics **71**(3), 606–615 (2015)
5. Furlotte, N.A., Eskin, E., Eyheramendy, S.: Genome-wide association mapping with longitudinal data. Genet. Epidemiol. **36**(5), 463–471 (2012)
6. Smith, E.N., Chen, W., Kähönen, M., Kettunen, J., Lehtimäki, T., Peltonen, L., Raitakari, O.T., Salem, R.M., Schork, N.J., Shaw, M., Srinivasan, S.R., Topol, E.J., Viikari, J.S., Berenson, G.S., Murray, S.S.: Longitudinal genome-wide association of cardiovascular disease riskfactors in the Bogalusa heart study. PLoS Genet. **6**(9), e1001094 (2010)
7. Jaffa, M., Gebregziabher, M., Jaffa, A.A.: Analysis of multivariate longitudinal kidney function outcomes using generalized linear mixed models. J. Transl. Med. **13**(1), 192 (2015)
8. Das, K., Li, J., Wang, Z., Tong, C., Guifang, F., Li, Y., Meng, X., Ahn, K., Mauger, D., Li, R., Rongling, W.: A dynamic model for genome-wide association studies. Hum. Genet. **129**(6), 629–639 (2011)
9. Sikorska, K., Montazeri, N.M., Uitterlinden, A., Rivadeneira, F., Eilers, P.H.C., Lesaffre, E.: GWAS with longitudinal phenotypes: performance of approximate procedures. Eur. J. Hum. Genet. **23**, 1384–1391 (2015)
10. Ding, L., Kurowski, B.G., He, H., Alexander, E.S., Mersha, T.B., Fardo, D.W., Zhang, X., Pilipenko, V.V., Kottyan, L., Martin, L.J.: Modeling of multivariate longitudinal phenotypes in family geneticstudies with Bayesian multiplicity adjustment. BMC proceedings **8**(Suppl 1), S69 (2014)
11. Musolf, A., Nato, A.Q., Londono, D., Zhou, L., Matise, T.C., Gordon, D.: Mapping genes with longitudinal phenotypes via Bayesian posterior probabilities. BMC Proc. **8**(Suppl 1), S81 (2014)
12. Wang, T.: Linear mixed effects model for a longitudinal genome wideassociation study of lipid measures in type 1 diabetes linear mixed effectsmodel for a longitudinal genome wide association study of lipid measures in type 1 diabetes. Master's thesis, McMaster University (2012)
13. Zhang, H.: Multivariate adaptive splines for analysis of longitudinal data. J. Comput. Graph. Stat. **6**, 74–91 (1997)
14. Kendziorski, C.M., Cowley, A.W., Greene, A.S., Salgado, H.C., Jacob, H.J., Tonellato, P.J.: Mapping baroreceptor function to genome: a mathematical modeling approach. Genetics **160**(4), 1687–1695 (2002)
15. Chung, W., Zou, F.: Mixed-effects models for GAW18 longitudinal blood pressure data. BMC Proc. **8**(Suppl 1), S87 (2014)
16. Stegle, O., Denby, K.J., Cooke, E.J., Wild, D.L., Ghahramani, Z., Borgwardt, K.M.: A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. J. Comput. Biol. J. Comput. Mol. Cell Biol. **17**(3), 355–367 (2010)
17. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge (2005)
18. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S.: A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**, 203–208 (2006)
19. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E.: Efficient control of population structure in model organism association mapping. Genetics **178**(3), 1709–1723 (2008)

20. Listgarten, J., Kadie, C., Schadt, E.E., Heckerman, D.: Correction for hidden confounders in the genetic analysis of gene expression. Proc. Nat. Acad. Sci. **107**(38), 16465–16470 (2010)
21. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D.: FaST linear mixed models for genome-wide association studies. Nat. Methods **8**(10), 833–835 (2011)
22. Stegle, O., Lippert, C., Mooij, J.M., Lawrence, N.D., Borgwardt, K.M.: Efficient inference in matrix-variate gaussian models with \iid observation noise. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24, pp. 630–638. Curran Associates Inc. (2011)
23. Candela, J.Q., Rasmussen, C.E.: A unifying view of sparse approximate gaussian process regression. J. Mach. Learn. Res. **6**, 1939–1959 (2005)
24. Titsias, M.K.: Variational learning of inducing variables in sparse Gaussian processes. Artif. Intell. Stat. **12**, 567–574 (2009)
25. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**(8), 904–909 (2006)
26. Fusi, N., Lippert, C., Lawrence, N.D., Stegle, O.: Warped linear mixed models for the genetic analysis of transformed phenotypes. Nature Communications, 5:4890 (2014)