Journal of Machine Learning Research 20 (2019) 1-30 Submitted 5/18; Revised 4/19; Published 6/19

Maximum Likelihood for Gaussian Process Classification and Generalized Linear Mixed Models under Case-Control Sampling

Omer Weissbrod OWEISSBROD@HSPH.HARVARD.EDU
Epidemiology Department
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Shachar Kaufman SHACHARK@POST.TAU.AC.IL
Department of Statistics and Operations Research
School of Mathematical Sciences
Tel-Aviv University
Tel-Aviv 6139101, Israel

David Golan DAVID@VIZ.AI
Viz.ai
Tel-Aviv 6121001, Israel

Saharon Rosset SAHARON@POST.TAU.AC.IL
Department of Statistics and Operations Research
School of Mathematical Sciences
Tel-Aviv University
Tel-Aviv 6139101, Israel

Editor: Ryan Adams

Abstract

Modern data sets in various domains often include units that were sampled non-randomly from the population and have a latent correlation structure. Here we investigate a common form of this setting, where every unit is associated with a latent variable, all latent variables are correlated, and the probability of sampling a unit depends on its response. Such settings often arise in case-control studies, where the sampled units are correlated due to spatial proximity, family relations, or other sources of relatedness. Maximum likelihood estimation in such settings is challenging from both a computational and statistical perspective, necessitating approximations that take the sampling scheme into account. We propose a family of approximate likelihood approaches which combine composite likelihood and expectation propagation. We demonstrate the efficacy of our solutions via extensive simulations. We utilize them to investigate the genetic architecture of several complex disorders collected in case-control genetic association studies, where hundreds of thousands of genetic variants are measured for every individual, and the underlying disease liabilities of individuals are correlated due to genetic similarity. Our work is the first to provide a tractable likelihood-based solution for case-control data with complex dependency structures.

Keywords: Gaussian Processes, Expectation Propagation, Composite Likelihood, Selection Bias, Linear Mixed Models

1. Introduction

In the analysis of scientific data, a common phenomenon is the existence of complex dependencies between analyzed units. This is encountered in diverse fields such as epidemiology, econometrics, ecology, geostatistics, psychometrics and genetics, and can arise due to spatial correlations, temporal correlations, family relations, or other sources of heterogeneity (Pfeiffer, 2008; Rabe-Hesketh et al., 2005; Bolker et al., 2009; Rabe-Hesketh et al., 2004; Yang et al., 2014; Burton et al., 1999; Diggle et al., 1998). This idea is often captured through the use of Gaussian processes (GPs; Rasmussen and Williams 2006) or equivalently, through generalized linear mixed models (GLMMs; McCulloch et al. 2008) or latent Gaussian models (Fahrmeir and Tutz, 2001). Such models associate sampled units with latent variables, and express the dependencies through covariance matrices of latent variables.

A second important concept is that of *ascertainment*, where the probability of sampling a unit depends on its response. Ascertainment is especially common in case-control studies where a binary response variable has a rare outcome, such as a rare disease, leading to oversampling of disease cases relative to their population prevalence (Breslow, 1996).

In this paper we consider situations that contain both elements—a complex covariance structure and case-control sampling—and the statistical modeling solutions available for these situations. Our interest lies in an extreme form of this combination, involving:

(a) Unit-level ascertainment, where the sampling probability of a unit depends only on its response (Figure 1a). This stands in contrast to common study designs such as family studies (Neuhaus et al., 2002), ascertained longitudinal studies (Liang and Zeger, 1986) or clustered case-control studies (Neuhaus and Jewell, 1990). In such studies each cluster is either entirely selected or entirely omitted from the study, such that the dependency structure in the sample and in the population are the same.

(b) A full-rank covariance matrix, indicating that dependencies cannot be captured by a small number of variables (Figure 1b). Such settings are common in modern data sets due to either high-dimensional settings or to the use of kernels or basis expansions, which implicitly project a small number of features into a large (possibly infinite-dimensional) space (Diggle et al., 1998; Rasmussen and Williams, 2006).

(c) A non-sparse covariance matrix, indicating that all units are correlated (Figure 1c). Such settings exacerbate both the computational challenge, because the density does not factorize into multiplicative terms, and the statistical challenge, because classic statistical theory requires a large number of independent units.

Special cases of this combination have been addressed in the literature (Glidden and Liang, 2002; Epstein et al., 2002; Neuhaus et al., 2006, 2014), but to our knowledge, there is a limited set of available solutions for the general setting, which is indeed very challenging.

A major motivating application for our study is genome-wide association studies of diseases with a case-control sampling design (GWAS-CC) (Price et al., 2015; Visscher et al., 2017). In GWAS-CC, the genomes of individuals affected with a disease and of unaffected controls are collected in an effort to uncover the genetic mechanisms driving disease risk. Studies in this field have diverse goals, reflected in the diversity of the statistical inference tasks they seek to solve: testing for association between genetic variants and disease (Yang

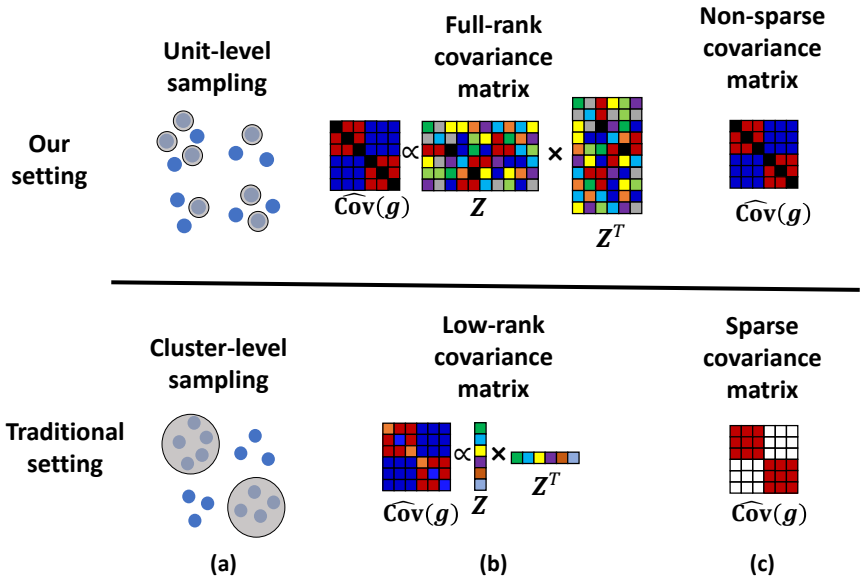

Figure 1: The properties that are unique to our setting of interest (top row) compared to more traditional settings (bottom row). (a) Unit-level sampling, where the decision whether to sample a unit depends only on their response, in contrast to studies that either sample or omit an entire cluster of units. (b) A full-rank covariance matrix of latent variables, indicating a complex dependency structure. (c) A non-sparse covariance matrix, indicating that the latent variables of every pair of units are correlated.

et al., 2014; Weissbrod et al., 2015), estimating disease heritability (Yang et al., 2010; Golan et al., 2014; Weissbrod et al., 2018), risk prediction (Zhou et al., 2013; Golan and Rosset, 2014; Weissbrod et al., 2016), and more.

GWAS-CC typically employ case-control designs, where patients are recruited in hospitals or clinics whereas healthy controls are recruited independently, owing to the small prevalence of complex genetic diseases—even common diseases like type 1 diabetes or schizophrenia typically have a population prevalence $\leq$1%. Furthermore, statistical models for such studies typically treat the effects of genetic variants on disease as random effects sampled from a distribution. This is because such studies include millions of variants, and the effects are typically very small (Yang et al., 2010; Golan et al., 2014; Golan and Rosset, 2014). Hence, GWAS-CC give rise to settings with case-control sampling and a full-rank and non-sparse covariance structure: every individual has a latent genetic effect (given by the inner product of her genotype and the random effects), and the genetic effects are correlated due to genetic similarity. Despite the extensive interest that GWAS-CC have attracted in recent years (Wellcome Trust Case Control Consortium et al., 2007; Ehret et al., 2011; Sawcer et al., 2011; Ripke et al., 2014; Okada et al., 2014), the statistical modeling problems this setting generates have been discussed in a limited manner, with application of heuristic methods that do not formally take the probabilistic structure of the problem into account

(Lee et al., 2011; Hayeck et al., 2015; Weissbrod et al., 2015; Chen et al., 2016; Jiang et al., 2016a).

Similar settings arise in other scientific domains, where case-control sampling and a full-rank, non-sparse covariance structure are observed. Prominent examples include disease mapping studies with a smoothing kernel (Diggle et al., 1998; Kelsall and Diggle, 1998; Held et al., 2005) and GP-based classification of data collected in case-control studies (Chu et al., 2010; Ziegler et al., 2014; Young et al., 2013). The analyses employed in these examples often ignore the effects of case-control sampling, a practice we would like to avoid and whose fundamental flaws we discuss and illustrate below.

The problem we consider poses substantial statistical and computational challenges. The main statistical framework for inference with binary responses and latent variables are GP classifiers, which are mathematically equivalent to latent Gaussian models, and recover GLMMs as a special case when using a linear kernel. Such models provide a likelihood-based solution but can pose significant computational difficulties. Modern approach to alleviate computational difficulties include (1) Pairwise likelihood (PL; Renard et al. 2004), which approximates the joint distribution of all variables as a product of marginal distributions of pairs of variables; (2) Expectation propagation (EP) (Minka, 2001; Seeger, 2005), which replaces multiplicative terms in the distribution with simpler terms from an exponential family distribution; (3) Variational approximation (Opper and Archambeau, 2009; Hensman et al., 2015), which approximates the distribution with the closest distribution from a more tractable class; (4) Markov chain Monte Carlo (MCMC) sampling combined with thermodynamic integration (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Gelman and Meng, 1998), and (5) Laplace approximation (Tierney and Kadane, 1986; Raudenbush et al., 2000) and its close variant, penalized quasi likelihood approximation (Breslow and Clayton, 1993; Wolfinger and O'connell, 1993), which approximate the distribution as a Gaussian distribution via a second-order Taylor expansion. Of these, PL and EP have proved to often outperform the alternatives (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Varin et al., 2011) and thus form the basis of our proposed approach.

The main approach for statistical inference in the presence of ascertainment is called *ascertained maximum likelihood* (AML). AML consists of defining a binary variable indicating inclusion in the study for every unit, and conditioning the analysis on these variables (Scott and Wild, 2001). However, none of the above GP approximation methods is suitable for AML inference. The only existing method that is both scalable and statistically sound is a method-of-moments approach called phenotype-correlation-genotype-correlation (PCGC; Golan et al. 2014). However, this approach is not likelihood-based and thus cannot naturally be used for model comparison, inference, prediction, and hypothesis testing.

Here we propose two approaches for approximate likelihood computation in our setting of interest. Our approaches combine (1) GP + AML + PL, which provides a tractable likelihood approximation but is very sensitive to model misspecification; and (2) GP + AML + modified EP, which is more computationally intensive, but is more robust and is closer to traditional maximum likelihood estimation. We evaluate the merits of our approaches on both synthetic and real data sets of genetic studies involving thousands of individuals and hundreds of thousands of features treated as having random effects.

2. Formal Problem Description

Here we provide a formal description of our problem and the statistical and computational challenges.

2.1. Gaussian Processes / Generalized Linear Mixed Models

We are presented with n units, with each unit i having c features $\boldsymbol{X}_i \in \mathbb{R}^c$ associated with fixed (non-random) effects $\boldsymbol{\beta} \in \mathbb{R}^c$, m additional features $\boldsymbol{Z}_i \in \mathbb{R}^m$ that are not associated with fixed effects, and an outcome variable y_i (here we consider binary outcomes, but the framework can also be applied to other types). We associate every unit i with a latent variable $g_i(\boldsymbol{Z}_i)$ that depends on $\boldsymbol{Z}_i$ such that $P\left(y_i = 1 \mid \boldsymbol{X}_i, g_i, \boldsymbol{\beta}\right) = h\left(\boldsymbol{X}_i^T\boldsymbol{\beta} + g_i\right)$, where $h\left(\cdot\right)$ is a likelihood function (which is closely related to an inverse link function in GLMM terminology), such as probit or logit.

GPs impose a Gaussian process prior over the the functional form of the latent variables, $g \sim \mathcal{GP}\left(\mathbf{0}, k(\cdot,\cdot\,;\boldsymbol{\theta})\right)$, where $k(\cdot,\cdot\,;\boldsymbol{\theta})$ is a *kernel function* parameterized by $\boldsymbol{\theta}$. This indicates that for every finite set of units $i = 1 \ldots n$, the vector $\boldsymbol{g} = [g_1(\boldsymbol{Z}_1), \ldots, g_n(\boldsymbol{Z}_n),]^T$ follows a multivariate normal distribution:

$$\boldsymbol{g} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}(\boldsymbol{Z};\boldsymbol{\theta})\right),$$

where $\boldsymbol{K}$, which is non-sparse and full-rank in our setting, has entries $K_{ij} = k(\boldsymbol{Z}_i, \boldsymbol{Z}_j;\boldsymbol{\theta})$.

GLMMs are a class of models that are mathematically equivalent to GPs with linear kernels, i.e., $\boldsymbol{K}_{ij} = \theta\boldsymbol{Z}_i^T\boldsymbol{Z}_j$, where θ is a scalar hyperparameter called a *variance component*. GLMM literature typically uses the alternative notation $g_i = \boldsymbol{Z}_i^T\boldsymbol{b}$, where $\boldsymbol{b} \sim \mathcal{N}\left(\mathbf{0}, \sqrt{\theta}\boldsymbol{I}\right)$ is a vector of *random effects*. The equivalence can be extended to non-linear kernels by making use of Mercer's theorem (König, 2013), which allows expressing every kernel function as a (possibly infinite) linear combination of basis functions, $k(\boldsymbol{Z}_i, \boldsymbol{Z}_j;\boldsymbol{\theta}) = \sum_{r=1}^{\infty}\phi_r(\boldsymbol{Z}_i;\boldsymbol{\theta}) \cdot \phi_r(\boldsymbol{Z}_j;\boldsymbol{\theta})$. Hence, every GP can be written as a GLMM over the feature space $\{\phi_r(\boldsymbol{Z})\}$ with iid random effects $b_r \sim \mathcal{N}(0,1)$. GLMMs can also be defined with non-normal random effects, in which case the equivalence with GPs breaks down, but we do not consider such models here. We focus on linear kernels because we are interested in extremely high-dimensional settings, where non-linear kernels often overfit (Weissbrod et al., 2016). A graphical model of GPs and GLMMs is shown in Figure 2.

Our main aim in this work is estimating the kernel hyperparameters $\boldsymbol{\theta}$ (i.e., training the model). Given a vector of observed outcomes $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ and the matrices $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^T$, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n]^T$, the GP likelihood is given by:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) = P\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta}\right) = \int P\left(\boldsymbol{g}|\boldsymbol{Z}, \boldsymbol{\theta}\right)\prod_i P(y_i|\boldsymbol{X}_i, g_i, \boldsymbol{\beta})d\boldsymbol{g}. \tag{1}$$

As we do not impose a prior over the hyperparameters $\boldsymbol{\beta}, \boldsymbol{\theta}$, we can estimate them via type-II maximum likelihood, by finding the values that maximize Equation 1.

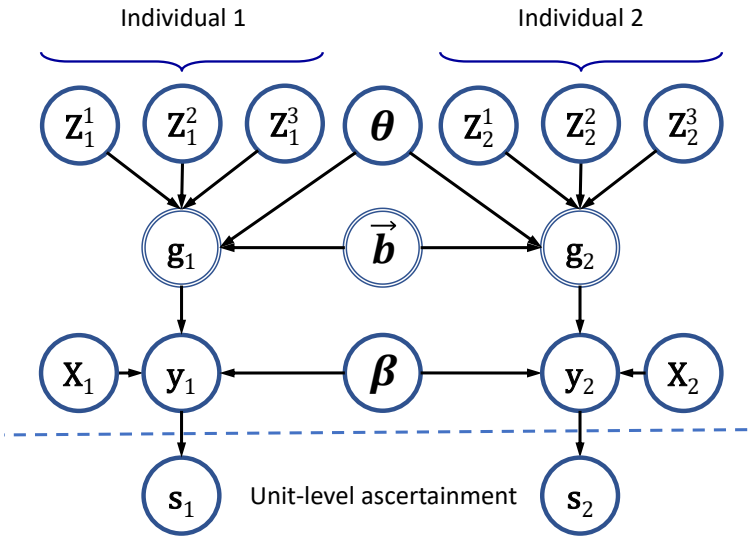

Figure 2: A directed graphical model for a GP with one feature X associated with a fixed effect β, three features Z^1, Z^2, Z^3 associated with an implicit vector of random effects $\vec{b}$ and with a hyperparameter θ, and two sampled units (indicated by subscript indices) with latent variables g_1, g_2 and observed responses y_1, y_2. Also shown is the extension to unit-level ascertainment, which consists of adding a sampling indicator s_i that depends on y_i and is equal to 1 for every sampled unit. Latent (non-observed) random variables are marked with a double-lined border.

2.2. The Implications of Ignoring Ascertainment in GPs

Up until now we implicitly assumed that n units were sampled completely at random from an underlying population. We now assume that the sample is ascertained, i.e., that the probability of sampling cases (units with $y_i = 1$) and controls ($y_i = 0$) is different.

We first demonstrate that using GPs while ignoring ascertainment leads to nonsensical conclusions which stand in contrast to fundamental motivations for GP use, like the central limit theorem. We focus on binary GPs, which can be formulated according to the liability threshold model (Dempster and Lerner, 1950). Under this model, every unit i has a latent liability $l_i = g_i + \epsilon_i$, where ϵ_i is an iid latent residual variable whose distribution depends on the likelihood function (e.g. normally distributed for probit, or logit distributed for logit), and unit i is a case (having $y_i = 1$) if and only if $l_i > t$ for some cutoff t. The *prevalence* K is the proportion of units in the population having $l_i > t$. We emphasize that a normally distributed ϵ_i is completely equivalent to a standard GP with a probit likelihood $h(\cdot)$.

It is common to use likelihood functions associated with a smooth and symmetrically distributed ϵ_i, such as logit or probit, which leads to a smooth and symmetric distribution of liabilities in the population. However, due to the ascertainment mechanism, the liabilities and latent variables g_i in an ascertained sample follow a non-symmetric and possibly discontinuous distribution (Figure 3a), and thus cannot be analyzed with standard likelihood functions. This problem motivates our proposed solutions for analysis of case-control

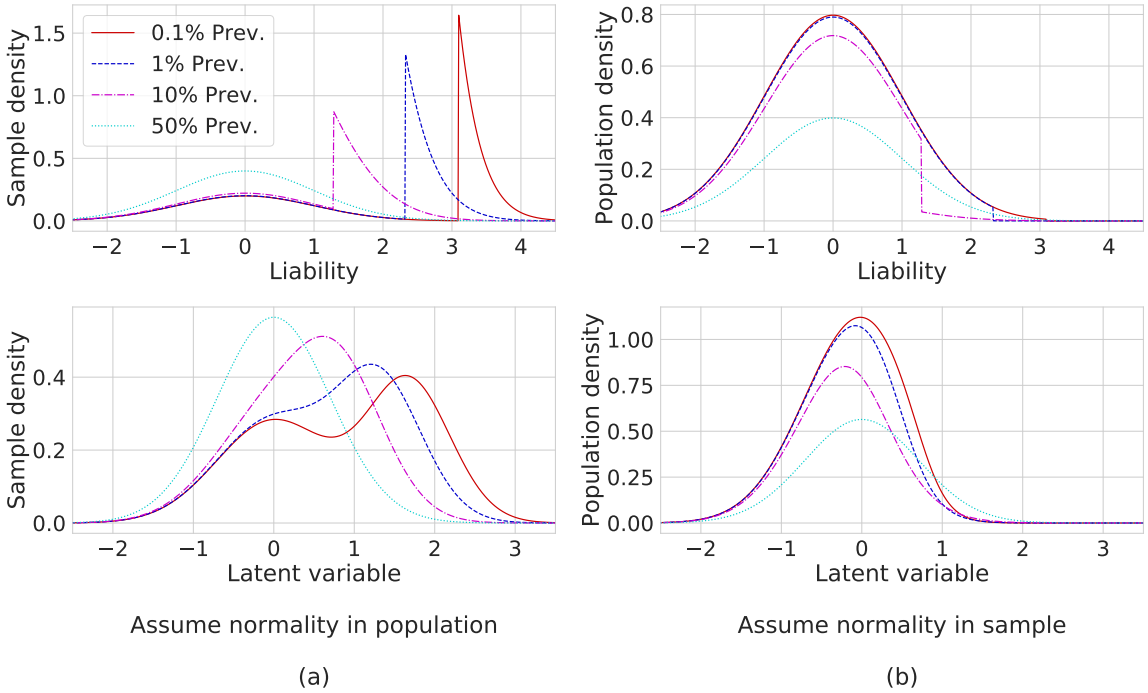

Figure 3: The implications of assuming normality of latent variables in the population from which units are sampled (panel a) or in a case-control study (panel b), for a GP with a probit likelihood and a sample consisting of 50% cases. The liability is given by $l_i = g_i + \epsilon_i$, $g_i, \epsilon_i \sim \mathcal{N}(0, \sqrt{0.5})$. Units with liabilities greater than their (1-prevalence) population quantile are cases. (a) When assuming normality in the population, latent variables and liabilities in a case-control study are not normally distributed (unless the cases prevalence is 50%, in which case there is no ascertainment). (b) When naively assuming normality of latent variables in a case-control study, the latent variables and the liabilities are *not* normally distributed in the population from which the data was sampled, in contradiction to the liability threshold model. Specifically, the liabilities distribution is discontinuous, and the latent variables distribution has a heavy left tail. All distributions were computed analytically by conditioning on the sampling indicators defined in Section 2.3.

studies. Many studies in practice ignore the complexities above, and instead use common likelihood functions such as a logit or a probit in case-control studies (e.g. Chen et al. 2016; Jiang et al. 2015; Kramer et al. 2017; Qi et al. 2017). However, this solution implies a non-symmetric and discontinuous distribution of latent variables in the population from which units are sampled, in stark contrast to the central limit theorem assumptions (Figure 3b). Thus, ignoring the ascertainment scheme in GPs may lead to nonsensical probabilistic settings under common assumptions.

The practical implications of modeling violations due to ascertainment have been investigated extensively in the statistical genetics literature. These include severe biases in estimation of quantities such as disease heritability (Golan et al., 2014; Weissbrod et al.,

2018), inaccurate risk prediction (Golan and Rosset, 2014) and loss of power in hypothesis testing (Weissbrod et al., 2015; Hayeck et al., 2015; Yang et al., 2014).

2.3. Modeling Ascertainment in GPs

Analysis of ascertained data is typically performed by (1) defining a binary sampling indicator s_i for every unit i such that s_i depends only on y_i (Figure 2); and (2) performing all statistical inference tasks conditional on $s_1 = 1, \ldots, s_n = 1$. This requires specifying two numbers: the sampling probabilities of cases $P(s_i = 1|y_i = 1)$ and of controls $P(s_i = 1|y_i = 0)$. There are two main approaches for estimating the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}$ of the distribution $P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z})$ with such indicators, differing with respect to how the sampling probabilities are determined.

Maximum profile likelihood estimates $\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}$ by jointly maximizing the so-called profile likelihood $P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z},s_1 = 1,\ldots,s_n = 1)$ over both $\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}$ and the nuisance hyperparameters $\boldsymbol{\theta}_{s|y}$ of the distributions $P(s_i = 1|y_i)$ of every possible value of y_i, i.e., $\hat{\boldsymbol{\theta}}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}$, $\hat{\boldsymbol{\theta}}_{s|y}$= $\text{argmax}_{\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}},\boldsymbol{\theta}_{s|y}} P\left(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{Z},s_1 = 1,\ldots,s_n = 1,\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}},\boldsymbol{\theta}_{s|y}\right)$ (Scott and Wild, 2001). The resulting estimator is maximally efficient as it attains the Cramér-Rao lower bound.

Ascertained Maximum Likelihood (AML) estimates $\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}$ using a pre-specified assignment $\boldsymbol{\theta}_{s|y} = \boldsymbol{\theta}^0_{s|y}$, i.e., $\hat{\boldsymbol{\theta}}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}} = \text{argmax}_{\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}}} P\left(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{Z},s_1 = 1,\ldots,s_n = 1,\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}},\boldsymbol{\theta}^0_{s|y}\right)$ (Scott and Wild, 1997). For a binary outcome with a population prevalence K and an in-sample prevalence P, any assignment $\boldsymbol{\theta}^0_{s|y}$ obeying the constraint $\frac{P(s_i=1|y_i=0)}{P(s_i=1|y_i=1)} = \frac{K(1-P)}{(1-K)P}$ guarantees consistent estimates (i.e., estimators converge to the true parameter values as sample size tends to infinity if the model is true), because it yields the observed case-control ratio in expectation. This approach is often termed pseudo likelihood or conditional likelihood (Manski, 1981; Hsieh et al., 1985), but as both terms have alternative meanings in GLMM literature, we use the term ascertained likelihood instead. AML is less statistically efficient than maximum profile likelihood, in the sense that the estimator has a larger variance. However, the loss of efficiency has been shown to be negligible in practice (Wild, 1991; Scott and Wild, 1997). AML has previously been used for family-based studies (Glidden and Liang, 2002; Epstein et al., 2002), but to our knowledge it has not been used under the combination of a dependency structure and unit-level sampling.

To combine GPs with the AML framework, we define the ascertained GP likelihood and apply Bayes' law as follows:

$$L^*(\boldsymbol{\beta},\boldsymbol{\theta}) = P(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{Z},\boldsymbol{s} = 1,\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{P(\boldsymbol{y} \mid \boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})}{P(\boldsymbol{s} = \boldsymbol{1} \mid \boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})} \prod_i P(s_i = 1|y_i), \quad (2)$$

where $\boldsymbol{s} = 1$ is a shorthand notation for $s_1 = 1,\ldots,s_n = 1$ and $\boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}} = \{\boldsymbol{\theta},\boldsymbol{\beta}\}$. The last term in the rhs of Equation 2 is considered known under AML and requires no special treatment. The numerator is the likelihood of a standard GP under no ascertainment, and the denominator is the likelihood of a GP in which the outcome is s_i instead of y_i. A naive approach is to approximate the numerator and denominator separately. However, obtaining an accurate estimate of the ratio is extremely challenging, because both the numerator and denominator are challenging to approximate, and any inaccuracy is compounded by the division. In our experience, this approach does not lead to reasonable estimators.

3. Approximate Inference in GPs under Ascertainment

We now propose two methods for approximate inference in GPs under ascertainment.

3.1. Ascertained Pairwise Likelihood

PL is a composite likelihood approximation, which approximates multivariate joint densities via products of marginal bivariate densities: (Varin et al., 2011; Renard et al., 2004):

$$P\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta}\right) \overset{\propto}{\approx} \prod_{i \neq j} P\left(y_i, y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}, \boldsymbol{\beta}\right), \tag{3}$$

where i,j iterate over all pairs of units, and $\overset{\propto}{\approx}$ indicates approximate proportionality with respect to the hyperparameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$. The maximum pairwise likelihood estimate is approximately the maximum likelihood estimate. PL is computationally efficient owing to its quadratic dependency on the sample size, and is statistically consistent under suitable regularity conditions (Varin et al., 2011).

Ascertained PL (APL) is an extension of PL that approximates the ascertained likelihood in Equation 2 by modifying Equation 3 to condition on $\boldsymbol{s} = 1$:

$$\begin{aligned} P\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}=\boldsymbol{1}\right) &\overset{\propto}{\approx} \prod_{i \neq j} P\left(y_i, y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j, s_i = 1, s_j = 1\right) \\ &= \prod_{i \neq j} \frac{P\left(y_i, y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j\right)}{P\left(s_i = s_j = 1 \mid \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j\right)} P\left(s_i = s_j = 1 \mid y_i, y_j\right), \end{aligned}$$

where $P\left(s_i = s_j = 1 \mid y_i, y_j\right) = P\left(s_i = 1 \mid y_i\right) P\left(s_j = 1 \mid y_j\right)$ are known constants which can be ignored, and we omitted the hyperparameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ for brevity. The terms in the numerator and the denominator can be separately evaluated as in standard PL, where we treat the denominator as a GP with a suitable likelihood function. Unlike Equation 2, the evaluation of the ratio is accurate since both the numerator and denominator can be computed exactly. In certain settings, PL evaluation can be substantially accelerated via a Taylor approximation around $\boldsymbol{Z}_i^T \boldsymbol{Z}_j = 0$, which enables factoring each bivariate distribution into a product of marginal distributions (Appendix A).

3.2. Ascertained Expectation Propagation

EP is a popular approach for approximating complex distributions by iteratively replacing every multiplicative term in the joint distribution of the observed and latent variables with a simpler term from an exponential family distribution (Minka, 2001; Rasmussen and Williams, 2006; Seeger, 2005). This joint distribution in GPs is given by $P\left(\boldsymbol{g}|\boldsymbol{Z}, \boldsymbol{\theta}\right) \prod_i P(y_i|\boldsymbol{X}_i, g_i, \boldsymbol{\beta})$. EP replaces every term in this product by an unnormalized Gaussian, $P\left(y_i|\boldsymbol{X}_i, g_i\right) \approx t_i\left(g_i\right) \triangleq r_i \mathcal{N}\left(g_i; \widetilde{\alpha}_i, \widetilde{\gamma}_i\right)$, where we omitted the hyperparameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ for brevity, and the site parameters r_i, $\widetilde{\alpha}_i$, $\widetilde{\gamma}_i$ implicitly depend on $\boldsymbol{X}_i$, y_i and $\boldsymbol{\beta}$. EP iteratively updates the terms $t_i(g_i)$, such that each term minimizes the generalized Kullback Leibler divergence (GKL) between the functions $q_{-i}(g_i)t_i(g_i)$ and $q_{-i}(g_i)P(y_i|\boldsymbol{X}_i, g_i)$ (i.e., the KL divergence between these functions after standardizing them to integrate to unity), where the cavity distribution $q_{-i}(g_i) \propto \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_{j \neq i} t_j(g_j) d\boldsymbol{g}_{j \neq i}$ represents the current approximation of $P\left(g_i|\boldsymbol{Z}, \boldsymbol{y}_{j \neq i}\right)$.

Given an EP approximation, the GP likelihood can be approximated as:

$$P\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{Z}\right) \approx \int P\left(\boldsymbol{g}|\boldsymbol{Z}\right) \prod_i t_i\left(g_i\right) d\boldsymbol{g}.$$

This expression can be evaluated analytically because it is an integral of a product of (unnormalized) Gaussian densities. EP has proven to consistently outperform alternative approximation methods for binary data (Nickisch and Rasmussen, 2008), and recent theoretical analysis has demonstrated that is is statistically consistent under certain modeling assumptions (Dehaene and Barthelmé, 2016, 2018).

Ascertained EP (AEP) is our proposed method to generalize standard EP to handle ascertainment. AEP approximates the ascertained likelihood in Equation 2 by replacing the standard EP step with a modified step that equates the functions $\int q_{-i}(g_i)t_i(g_i)dg_i$ and $\frac{\int q_{-i}(g_i)P(y_i,s_i|g_i,\boldsymbol{X}_i)dg_i}{\int q_{-i}(g_i)P(s_i|g_i,\boldsymbol{X}_i)dg_i}$. Unlike standard EP, we cannot minimize the GKL divergence beteween these functions, because this will lead to the same solution as standard EP, up to a scaling constant. Instead, AEP finds the unnormalized Gaussian $t_i(g_i)$ which makes these functions and their first two partial derivatives with respect to μ_{-i} (the mean of the Gaussian $q_{-i}(g_i)$) have the same value when evaluated at μ_{-i} (see Appendix B for details).

EP is a special case of AEP, because the proposed step objective coincides with the standard EP objective in the absence of ascertainment (i.e., when $P(s_i|y_i)$ is a constant regardless of y_i). To see this, observe that EP minimizes the GKL divergence between $\hat{m}(g_i) \triangleq q_{-i}(g_i)P(y_i|g_i, \boldsymbol{X}_i)$ and $\tilde{m}(g_i) \triangleq q_{-i}(g_i) \cdot t_i(g_i)$. Since $\tilde{m}(g_i)$ is an unnormalized Gaussian, EP minimizes the GKL divergence by equating its zeroth, first and second moments with those of $\hat{m}(g_i)$ (Rasmussen and Williams, 2006). Hence, standard EP requires computing the mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ of $\hat{m}(g_i)$. A straightforward but lengthy derivation shows that we can compute these quantities as follows:

$$\begin{aligned}
\hat{\mu}_i &= \frac{\partial}{\partial \mu_{-i}} \left[\log \int \hat{m}(g_i) dg_i\right] \sigma_{-i}^2 + \mu_{-i} \\
\hat{\sigma}_i^2 &= \frac{\partial^2}{\partial \left(\mu_{-i}\right)^2} \left[\log \int \hat{m}(g_i) dg_i\right] \left(\sigma_{-i}^2\right)^2 + \sigma_{-i}^2,
\end{aligned}$$

where μ_{-i}, σ_{-i}^2 are the mean and variance of the Gaussian $q_{-i}(g_i)$, and the derivatives are evaluated at the actual value of μ_{-i}. Hence, there is a one-to-one correspondence between the first two moments of $\hat{m}(g_i)$ and its first two partial derivatives with respect to μ_{-i} (when evaluated at μ_{-i}). Consequently, each step of standard EP can alternatively be described as imposing the constraint that the zeroth, first and second derivatives of the integrals of $\tilde{m}(g_i)$ and $\hat{m}(g_i)$ with respect to μ_{-i} are the same. This is the same constraint used in AEP. Hence, EP and AEP coincide in the absence of ascertainment, where $P(s_i = 1|y_i)$ is constant regardless of the value of y_i.

The sampling variance of the AEP maximum likelihood estimator can be estimated efficiently via jackknife sampling, by reusing the functions $t_i(g_i)$ (Opper and Winther, 2000; Qi et al., 2004; Vehtari et al., 2016). The evaluation of each jackknife sample requires inverting a matrix that is a submatrix of a matrix that was inverted in the original computation, with one row and one column removed. Such an inversion can be computed rapidly while retaining numerical stability, by combining a Cholesky decomposition with a series of Givens

rotations (Seeger, 2004). A formal analysis of AEP is difficult because there are relatively few theoretical guarantees for standard EP, which is a special case of AEP, except under relatively strong assumptions (Dehaene and Barthelmé, 2016, 2018). However, we sketch a heuristic argument supporting the objective function of AEP in Appendix B.

4. Results

We evaluated the performance of our methods using extensive simulations and real data analysis. We first describe our simulation studies, and then present the results obtained on real data.

4.1. Simulations Overview

We simulated data that closely mimics real GWAS-CC using the liability threshold model, where each individual has liability $l_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + g_i + \epsilon_i$, g_i is a GP latent variable, and individuals with l_i greater than some cutoff are cases. In most simulations we used a linear kernel with a single hyperparameter, $\boldsymbol{\theta} = \{\sigma_g^2\}$, though we also investigate radial basis function (RBF) kernels below. Our aim was estimating the hyperparameter σ_g^2, which we call a variance component. Importantly, $\sigma_g^2/\text{var}(l_i)$ is an estimator of genetic heritability, defined as the proportion of $\text{var}(l_i)$ explained by genetics.

We simulated genetic data based on single nucleotide polymorphisms (SNPs), which can be encoded as $0/1/2$, according to the number of minor alleles carried by an individual at a specific position in the genome. We first generated a minor allele frequency $f^j \sim \mathcal{U}(0.05, 0.5)$ for every SNP j, and then sampled a matrix $\boldsymbol{Z}$ of SNP, such that $Z_{ij} \sim \text{Bin}(2, f^j)$. Finally, we standardized each column in the matrix $\boldsymbol{Z}$ by subtracting the mean and dividing by the standard deviation corresponding to its allele frequency.

To simulate unit-level ascertainment, we (1) generated a population of 1,000,000 individuals, where for every individual i we generated a vector of standardized genotypes $\boldsymbol{Z}_i \in \mathbb{R}^m$ as described above, and a vector $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{I}) \in \mathbb{R}^c$ representing additional standardized risk factors such as sex or age; (2) generated vectors $\boldsymbol{b} \sim \mathcal{N}\left(0, \sqrt{\sigma_g^2/m}\boldsymbol{I}\right)$ of random effects, and $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \sqrt{\sigma_c^2/c}\boldsymbol{I}\right)$ of fixed effects; (3) assigned a latent variable $g_i = \boldsymbol{Z}_i^T\boldsymbol{b}$ and a liability $l_i = g_i + \boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i$ to every individual i, where $\epsilon_i \sim \mathcal{N}\left(0, \sqrt{1 - \sigma_c^2 - \sigma_g^2}\right)$ iid; (4) defined all individuals with l_i greater than the $1 - K$ quantile of the liability distribution as cases (i.e., $y_i = 1$), where K is the desired prevalence; and (5) selected a subset of $\frac{n}{2}$ cases and $\frac{n}{2}$ controls for the case-control study, where n is the desired study size. Unless stated otherwise, we used $m = 500$, $n = 500$, $K = 1\%$, $\sigma_g^2 = 0.25, \sigma_c^2 = 0.25$, and $c = 1$. We generated 100 datasets for each unique combination of evaluated settings.

We evaluated three likelihood-based methods: AEP, APL and plain EP, which does not account for ascertainment. We additionally evaluated the moment-based method PCGC, which is considered the state of the art approach for hyperparameter estimation in genetic case-control studies (Golan et al., 2014). We estimated hyperparameters in the likelihood-based methods via maximum likelihood, and in PCGC by finding the values that minimize the squared loss between the observed and expected values of $\text{cov}(y_i, y_j)$ across all pairs of individuals i, j. In all settings, we first estimated the fixed effects via a novel ascertainment-

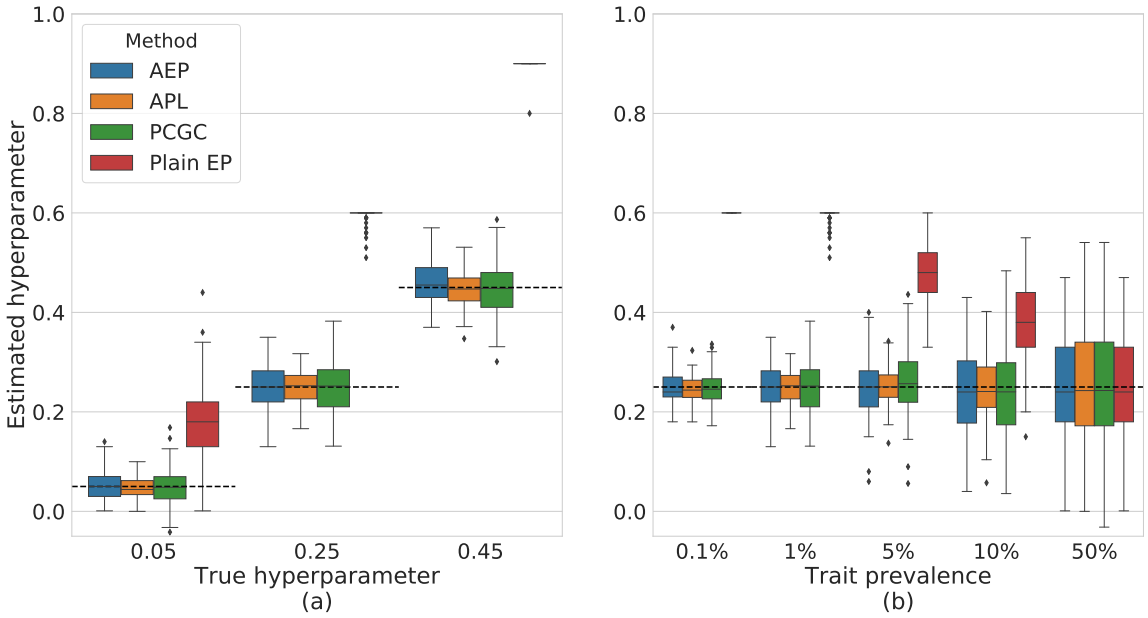

Figure 4: Evaluating hyperparameter estimation accuracy. Shown are box-plots depicting the estimates of each method across 100 different simulations, under data sets with an equal number of cases and controls, and a model with a single scale hyperparameter σ_g^2. The dashed horizontal lines represent the true underlying values of σ_g^2 used to generate the data. (a) AEP, APL and PCGC provide accurate estimates of σ_g^2 when the true trait prevalence (the prevalence of cases in the population) is 1%, for various values of σ_g^2, whereas plain EP is severely biased. (b) All methods except for plain EP accurately estimate σ_g^2 regardless of the underlying trait prevalence. Plain EP is accurate only when the prevalence is 50%, in which case there is no ascertainment.

aware generalized estimating equations (AGEE) approach that we developed, and then adjusted the affection cutoffs accordingly (Appendix C).

4.2. Simulation Studies: Estimating hyperparameters

Our first experiment evaluated variance component estimation accuracy. All methods except plain EP yielded empirically unbiased estimates, whereas plain EP was severely biased (Figure 4a). We also generated data under different prevalence values K and verified that all methods except plain EP remained accurate regardless of K, whereas plain EP was only accurate when $K = 0.5$, in which case there is no ascertainment (Figure 4b).

In the next experiment we examined sensitivity to sample size n and dimensionality m (corresponding to the number of rows and columns in the matrix $\boldsymbol{Z}$, respectively). We first verified that all methods became increasingly accurate with increasing sample size, but PCGC had a consistently larger sampling variance, because it uses a moment-based rather than a likelihood-based estimator (Figure 5a). We also observed that all methods became increasingly accurate with increasing dimensionality, but AEP was substantially

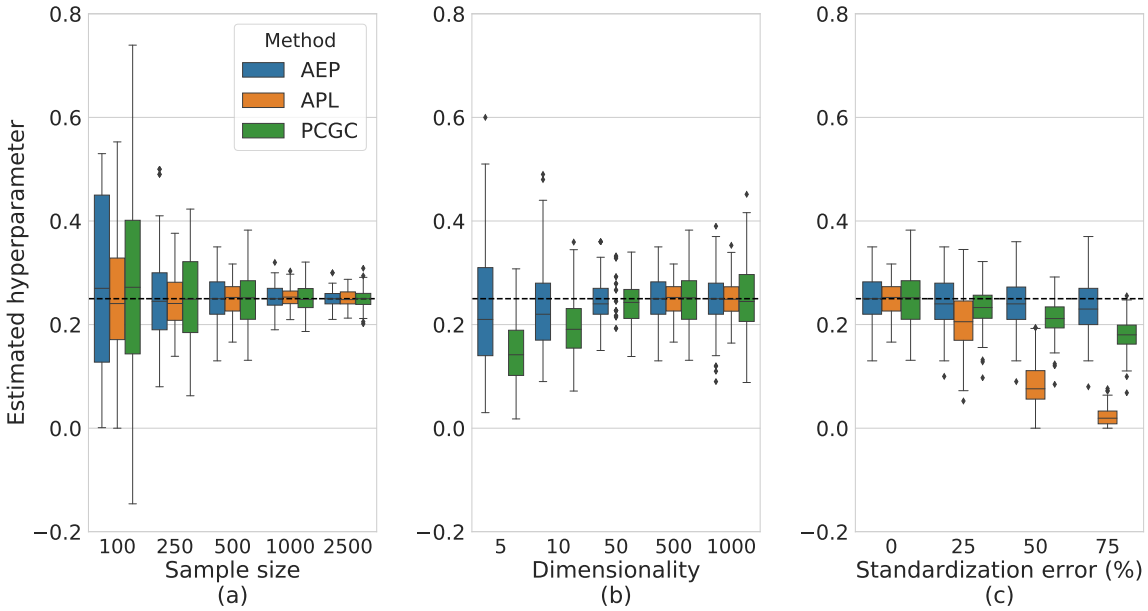

Figure 5: Investigating how hyperparameter estimation performance is affected by sample size, data dimensionality and modeling violations. (a) All the methods gain accuracy as the sample size increases. The sampling variance of PCGC is consistently larger, because it uses a moment-based rather than a likelihood-based estimator. (b) All the methods gain accuracy as dimensionality increases. AEP is substantially more accurate than the other methods in the presence of a small number of features, because the other two methods use a first-order Taylor expansion around $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{I}$, which is less accurate in the presence of a small number of features. APL estimates for numbers <50 are equal to 1.0, and are omitted for clarity. (c) AEP is robust to feature standardization misspecification (see main text), whereas PCGC is moderately sensitive and APL is highly sensitive.

more accurate when $m < 50$ (Figure 5b). This is because the other two methods use a first-order Taylor expansion around $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{I}$, which is less accurate when m is small.

We also examined robustness to modeling violations by introducing noise into the feature standardization procedure. We multiplied the estimated frequency of every binary variable j by $r_j \sim U\left(\frac{1}{1+e}, 1+e\right)$ before standardizing it, where $e \in [0,1]$ is the error magnitude, and used this value for estimation, but not for the true generative model. This noise model is motivated by GWAS, where genetic variants are often standardized according to (somewhat noisy) estimates of their population frequency rather than their sample frequency, to prevent bias due to ascertainment. AEP was highly robust to such modeling misspecification, whereas PCGC was moderately sensitive and APL was highly sensitive to such misspecification (Figure 5c). These experiments indicate that AEP is more reliable than the other methods under a wide variety of modeling assumptions, and is thus the method of choice for GP likelihood-based hyperparameter estimation in case-control studies.

Next, we examined estimation accuracy under non-linear kernels. We generated data with a scaled RBF kernel, $K_{ij} = \sigma_g^2 \exp\left(-\left\|\boldsymbol{Z}_i - \boldsymbol{Z}_j\right\|^2 / \left(2\gamma^2\right)\right)$, and estimated the hyper-parameters $\boldsymbol{\theta} = \left\{\sigma_g^2, \gamma\right\}$. Our generative model used $\sigma_g^2 = 0.25$, $\gamma = 0.5$, and the same values as in the linear kernel simulations for all other parameters, except for restricting to m =10 normally-distributed features. This is because RBF kernels tend to overfit under a large m, yielding K_{ij} that is very close to either 0 or σ_g^2 regardless of $\boldsymbol{Z}$.

A technical challenge of the RBF experiments is that our simulations first generate true latent variables g_i for a population of 1M units. This requires computing a $1M \times 1M$ RBF covariance matrix $\boldsymbol{K}$ and sampling $\boldsymbol{g} = [g_1, \ldots, g_n]^T$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$, which is computationally intractable under a non-linear kernel. Instead we (1) generated a base population of 10K feature vectors $\boldsymbol{Z}_i$ and a corresponding $10K \times 10K$ RBF kernel matrix $\boldsymbol{K}_{10K}$; (2) sampled 10K g_i values from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{10K})$; and (3) created a population of 1M units, such that each unit has a vector $\boldsymbol{Z}_i$ and a corresponding g_i value selected at random from the base population, along with uniquely generated values of $\boldsymbol{X}_i$ and ϵ_i. Afterwards we followed the same procedure as in the linear kernel simulations of (1) generating a liability $l_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + g_i + \epsilon_i$ for each unit; (2) determining the liability cutoff according to the desired prevalence; and (3) sampling $n/2$ cases and $n/2$ controls. We modified PCGC and APL to ignore pairs of units with identical features in these experiments.

AEP was empirically unbiased in the RBF experiments, having an average estimation bias of -0.016 (stdev 0.068) for σ_g^2 and of 0.022 (stdev 0.076) for γ across 100 simulations. In contrast, APL, PCGC and plain EP were severely biased, with an average bias >0.2 in the estimation of σ_g^2 and >0.15 (in absolute value) in the estimation of γ. We verified that the bias was not due to the modified data generation scheme by repeating the same experiments with a linear kernel, wherein PCGC and APL were empirically unbiased. These results likely arise because PCGC and APL both use a first-order Taylor expansion around $K_{ij} = 0$, which may be less accurate in the presence of non-linear kernels. We conclude that PCGC and APL cannot be trivially modified to handle non-linear kernels, whereas AEP can be used in more general settings.

Finally, we examined the computational speed of the methods. PCGC and APL are very efficient compared to AEP, because they scale quadratically with the sample size whereas AEP scales cubically, like standard EP (Nickisch and Rasmussen, 2008). Nevertheless, AEP can perform maximum likelihood estimation in data sets with 3,000 units in less than two hours, and is thus applicable to solve reasonably sized real-world problems. AEP can potentially be scaled up using novel methods developed for GPs (see Discussion).

4.3. Real Data Analysis

We evaluated the accuracy of GP hyperparameter estimation in real data sets. To this end we estimated the heritability of seven complex disorders, having population prevalence between 0.1% and 6%, based on data sets with ∼3,700 individuals and ∼280,000 genetic variants from the Wellcome Trust 1 case-control consortium (Wellcome Trust Case Control Consortium et al., 2007). We performed stringent preprocessing to avoid confounding artifacts, as reported in our previous publication (Weissbrod et al., 2016). We used the same model as in the simulation studies. Specifically, we used a liability threshold model (i.e., a scaled probit likelihood) and a linear kernel, $K_{ij} = \theta \boldsymbol{Z}_i^T \boldsymbol{Z}_j$, where $\boldsymbol{Z}_i$,$\boldsymbol{Z}_j$ are 280,000-

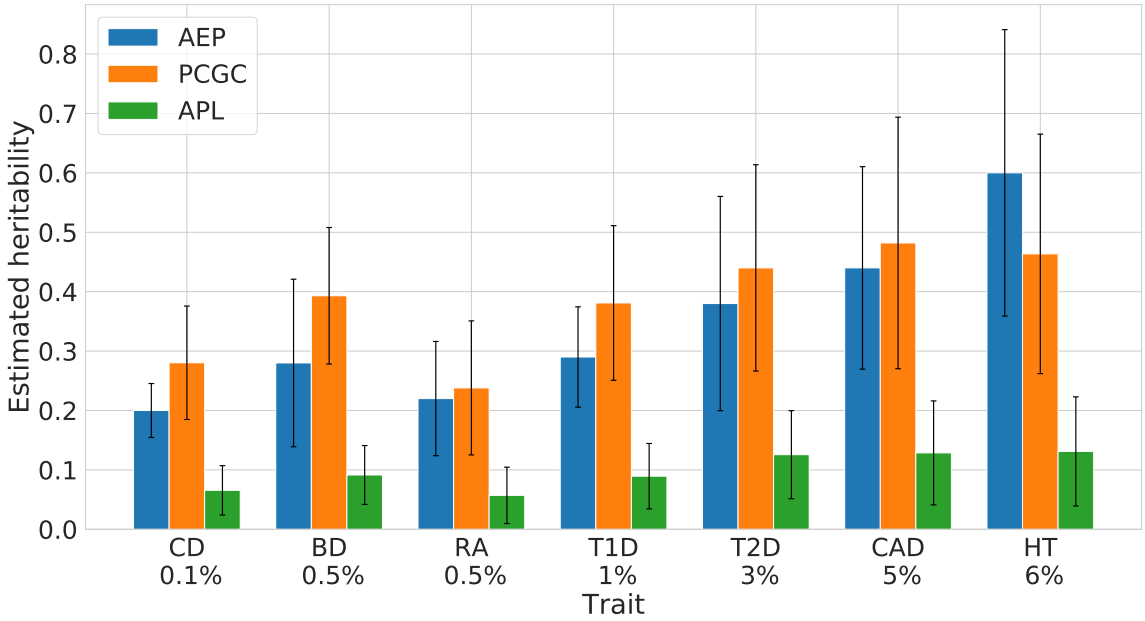

Figure 6: Shown are estimates of heritability (the proportion of liability variance explained by genetic factors) of seven complex disorders from (Wellcome Trust Case Control Consortium et al., 2007). The error bars are the standard deviation multiplied by 1.96, as estimated via jackknife. The disorders are Crohns disease (CD), rheumatoid arthritis (RA), bipolar disorder (BD), type 1 diabetes (T1D), type 2 diabetes (T2D), coronary artery disease (CAD) and hypertension (HT). The population prevalence of each trait is shown below its name. The estimates of AEP and PCGC are relatively concordant, whereas the APL estimates are significantly down-biased, in agreement with the modeling misspecification simulations.

dimensional vectors encoding the number of minor alleles in SNPs (standardized as in the simulation studies), and θ is a linear scaling factor. We associated sex with a fixed effect and estimated its effect via AGEE. Standard errors were computed via jackknife sampling.

The SNP heritability estimates of the investigated disorders lied in the range 20%-60% (Figure 6). There was a high degree of concordance between PCGC and AEP, whereas the APL estimates were substantially lower, consistent with the simulation studies. To further demonstrate the capabilities of AEP we estimated the posterior distribution of the GP latent variables g_i under both AEP and plain EP, using the standard EP approximation of the posterior distribution of latent variables (Rasmussen and Williams, 2006). Inferring the latent variables has a clinical utility as a measure of disease severity: individuals with larger values have a stronger genetic load of disease-inducing variants. The latent variables follow a clear bimodal distribution in all cases, but AEP provides a stronger separation between cases and controls (Figure 7). This result cannot be obtained with existing methods, because no existing likelihood-based GP approximation can model ascertainment.

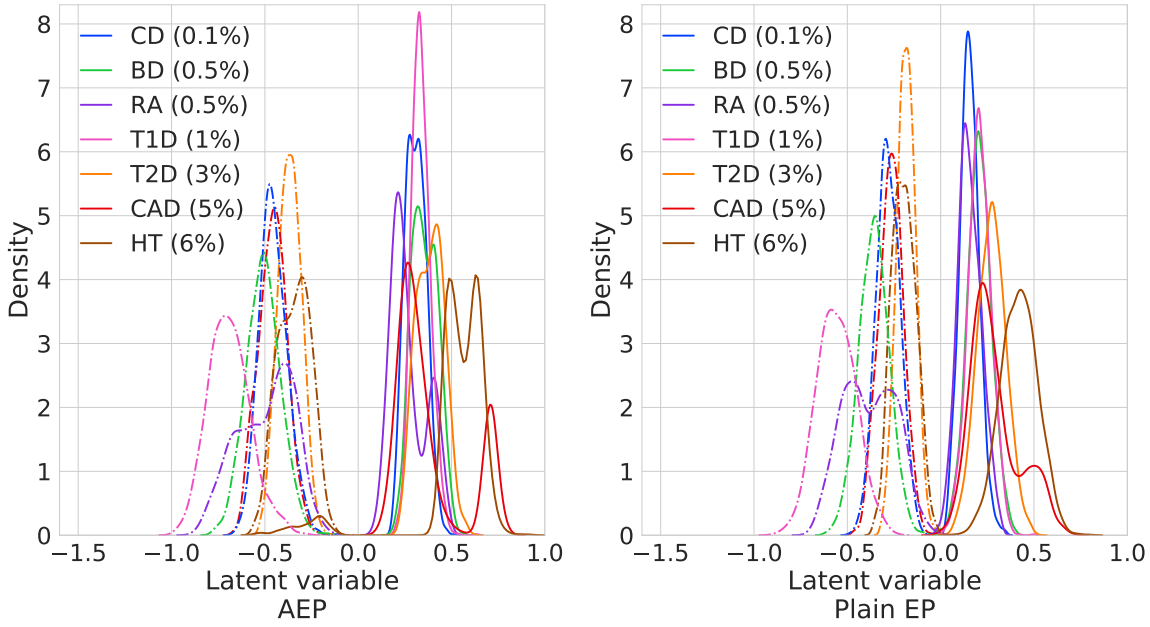

Figure 7: Inference of latent variables of complex genetic disorders. Shown is the distribution of the posterior mean of latent variables g_i provided by AEP (left) and plain EP (right), estimated via Gaussian kernel density estimation separately for controls (dashed lines) and cases (solid lines). The trait names are the same as in Figure 6, and prevalences are shown in parentheses. Individuals with larger posterior mean estimates carry a greater genetic load of disease-inducing variants. AEP provides a clearer separation between cases and controls by exploiting knowledge about the prevalence and sampling scheme. For both traits, the distribution variance increases with heritability and prevalence. Note that these distributions are not analogous to the ones in Figure 3 because they are based on posterior rather than marginal prior distributions.

5. Discussion

We presented several methods for inference of GP hyperparameters in settings with unit-level ascertainment, and a full-rank, non-sparse covariance structure. This was done by combining the ascertained likelihood framework with GPs and GLMMs, which form the statistical backbone of likelihood based analysis of non-iid data.

We proposed two approximate likelihood-based methods for the ascertained GP framework, AEP and APL, and empirically compared them with PCGC—the current state of the art method for estimating variance components in genetic case-control studies—which uses a moment-based rather than a likelihood-based estimator. APL is very computationally efficient but is sensitive to model misspecification. AEP, which is the most complex and empirically best approximation of maximum likelihood we proposed, is slower and is more technically complex than APL and PCGC, but is consistently more accurate than PCGC, and is less sensitive to modeling assumptions in our simulations. AEP additionally has

the advantage of providing a full probabilistic model with a well-defined likelihood, and it recovers standard EP as a special case under random ascertainment. On the other hand, PCGC has a principled underlying approximation, whereas APL and AEP are less well understood. Hence, the three methods are complementary in terms of their strengths and weaknesses, and we encourage future case-control studies to use multiple methods to gain a deeper understanding of high dimensional dependency structures.

The combination of unit-level ascertainment and a full-rank, non-sparse covariance structure is very common in statistical genetics (Golan et al., 2014), but is often encountered in other scientific domains, such as geostatistics and GP classification (Diggle et al., 1998; Chu et al., 2010; Ziegler et al., 2014; Young et al., 2013). Ascertained sampling is almost inevitable when studying rare phenomena, and the increasing dimensionality of studied data often necessitates the introduction of random rather than fixed effects, which in turn induce full-rank, non-sparse dependency structures. Additionally, it is often more convenient to perform dense sampling in a small number of clusters rather than collecting a large number of clusters (Bellamy et al., 2005; Zhang, 2004; Glidden and Vittinghoff, 2004), leading to non-sparse, full-rank dependency structures at the cluster level. Hence, we expect our work to be applicable in diverse scientific fields.

In this study we extended the well-known EP algorithm (Minka, 2001) to approximate GP likelihood. Another potential approach is MCMC sampling coupled with an integration scheme such as thermodynamic integration (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Gelman and Meng, 1998), but in our experience such approaches are too slow and complex for modern sized data sets. In recent years, Bayesian approaches have proven to be potential alternatives to likelihood based approaches in GPs (Ferkingstad and Rue, 2015). However, such approaches can be sensitive to the choice of prior distribution, and require prohibitively computationally expensive MCMC sampling. Several analytical approximations exist, but these are often inaccurate in the presence of binary data (Fong et al., 2010). The potential use of sampling-based approaches for inference in GPs under case-control ascertainment remains to be explored.

In recent years, genetic biobanks with hundreds of thousands of individuals have become available (Bycroft et al., 2018). AEP scales cubically with sample size and is thus not scalable to such datasets. GP approximation techniques from the machine learning community, such as mixture-of-experts models (Deisenroth and Ng, 2015), inducing points (Snelson and Ghahramani, 2006; Wilson and Nickisch, 2015; Gardner et al., 2018), random feature expansions (Rahimi and Recht, 2008; Le et al., 2013; Yang et al., 2015) and stochastic variational approximations (Hensman et al., 2013; Wilson et al., 2016; Cheng and Boots, 2017), or from the GWAS community (e.g. Loh et al. 2015) can potentially be used to scale up AEP to such large datasets.

A major challenge of non-sparse dependencies is that statistical theory is relatively undeveloped for this case. Specifically, assuming a study with r mutually independent clusters of m units, statistical theory is well developed for the asymptotic behavior $r/m \to \infty$, but is limited for $r/m \to 0$, which is our setting of interest (as $r=1$ when the covariance matrix is non-sparse). The statistical consistency of estimators in such cases has been established in limited settings, including GEEs (Xie and Yang, 2003), maximum penalized quasi likelihood (Bellamy et al., 2005), composite likelihood approximation (Heagerty and Lele, 1998), Laplace approximations (Shun and McCullagh, 1995), and specific geostatistical

models (Zhang and Zimmerman, 2005; Du et al., 2009). Several studies have established the statistical consistency of maximum likelihood estimators for linear mixed models (i.e., GPs with a linear kernel and a normal likelihood) in similar settings using random matrix theory (Bonnet et al., 2015; Jiang et al., 2016b; Dicker and Erdogdu, 2016), but to our knowledge such results have not been derived for GPs with non-normal likelihoods. We conclude that there is a major gap in statistical theory regarding $r/m \to 0$ asymptotics, representing questions of both theoretical and practical importance.

Several topics that remain unexplored in this work are GPs with more advanced kernels, outcome prediction and testing of fixed effects, for which several heuristic methods have been proposed in the statistical genetics literature (Hayeck et al., 2015; Weissbrod et al., 2015; Chen et al., 2016; Jiang et al., 2016a). Extending our approach to handle these topics is a potential avenue for future work.

Acknowledgments

This work was supported by grant 1804/16 from the Israel Science Foundation. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. We thank Malka Gorfine for fruitful discussions.

Appendix A

Here we describe a fast Taylor expansion-based approximation to APL estimation of GPs with a probit likelihood and a scaling parameter σ^2. Denote $\text{cov}(g_i, g_j) = \rho\sigma_g^2$, where g_i is the latent variable of unit i and ρ depends on $\boldsymbol{Z}_i$,$\boldsymbol{Z}_j$, and on all the other kernel hyperparameters. The joint likelihood of each pair of units can be written as:

$$P(y_i = a, y_j = b \,|\, \boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j, s_i = s_j = 1) = \frac{A_{ab}(\rho)}{B(\rho)} P(s_i = 1 \,|\, y_i) P(s_j = 1 \,|\, y_j), \quad (4)$$

where $A_{ab}(\rho) \triangleq P(y_i = a, y_j = b \,|\, \boldsymbol{X}_i, \boldsymbol{X}_j, \rho)$, $B(\rho) \triangleq P(s_i = s_j = 1 \,|\, \boldsymbol{X}_i, \boldsymbol{X}_j, \rho)$, and we omitted the hyperparameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ for brevity. Using the law of total probability, we can write: $B(\rho) = (s^1)^2 A_{11}(\rho) + s^1 s^0 \left(A_{10}(\rho) + A_{01}(\rho)\right) + (s^0)^2 A_{00}(\rho)$, where $s^t = P(s_i = 1 | y_i = t)$. Next, we explicitly evaluate these quantities at $\rho = 0$: $A_{ab}(0) = K_i^a (1 - K_i)^{1-a} K_j^b (1 - K_j)^{1-b}$, $B(0) = \left(s^0(1 - K_i) + s^1 K_i\right)\left(s^0(1 - K_j) + s^1 K_j\right)$, where $K_i = P(y_i = 1 | \boldsymbol{X}_i)$, and we omitted the dependence on $\boldsymbol{Z}_i$ because we assume that $g_i \sim \mathcal{N}(0, \sigma_g^2)$ marginally regardless of $\boldsymbol{Z}_i$. The above equations hold because y_i, y_j and s_i, s_j are independent given $\rho = 0$.

We next compute the partial derivatives of both expressions with respect to ρ at $\rho = 0$. Following (Golan et al., 2014), we have:

$$\begin{aligned}\frac{d}{d\rho}A_{ab}(\rho)|_{\rho=0} &= \phi(t_i)\phi(t_j)\sigma^2(-1)^{a\neq b} \\ \frac{d}{d\rho}B(\rho)|_{\rho=0} &= (s^1)^2\frac{d}{d\rho}A_{11}(\rho)|_{\rho=0} + 2s^1s^0\frac{d}{d\rho}A_{a\neq b}(\rho)|_{\rho=0} + (s^0)^2\frac{d}{d\rho}A_{00}(\rho)|_{\rho=0} \\ &= \phi(t_i)\phi(t_j)\sigma^2\left((s^1)^2 + (s^0)^2 - 2s^1s^0\right),\end{aligned}$$

where $\phi(\cdot)$ is the standard normal density, and $t_i = \Phi^{-1}(1-K) - \boldsymbol{X}_i^T\boldsymbol{\beta}$ is the liability cutoff for unit i, with $\Phi(\cdot)$ representing the standard normal cumulative density and K being the prevalence of cases in the population.

Finally, we plug in the above expressions into the Taylor expansion of Equation 4 at $\rho = 0$, which can be written as follows:

$$\frac{A_{ab}(\rho)}{B(\rho)}P(s_i \mid y_i)P(s_j \mid y_j) = \left(\frac{A'_{ab}(0)B(0) - B'(0)A_{ab}(0)}{B(0)^2}\rho + \mathcal{O}(\rho^2)\right)P(s_i \mid y_i)P(s_j \mid y_j).$$

Appendix B

Here we provide an informal analysis motivating the use of AEP. A formal analysis of AEP is difficult because there are relatively few theoretical guarantees for standard EP (Dehaene and Barthelmé, 2016, 2018). Instead, we state several assumptions and then provide an informal analysis under these assumptions. Throughout this Appendix, the notation $\boldsymbol{s}$ is a shorthand notation for $s_1 = \ldots = s_n = 1$, $\boldsymbol{u}_{-i}$ indicates the vector $\boldsymbol{u}$ with the i^{th} entry removed, and we omit the dependence on $\boldsymbol{\beta}, \boldsymbol{\theta}$ for brevity.

The parametric form of the AEP site approximation

We first demonstrate that in order for AEP to approximate the ascertained likelihood $P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s})$ (up to a scaling factor) at its fixed point, the site function $t_i(g_i)$ needs to approximately take the following parametric form:

$$t_i(g_i) \approx \frac{P(y_i, s_i|g_i, \boldsymbol{X}_i)}{P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i})}. \tag{5}$$

Note that this is different from the standard EP approximation $t_i(g_i) \approx P(y_i|X_i, g_i)$.

Our derivation requires an additional approximation:

Approximation 1

$$P(\boldsymbol{s}|\boldsymbol{X}, \boldsymbol{Z}) \approx \frac{1}{C(\boldsymbol{X}, \boldsymbol{Z})}\prod_i P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i}), \quad \text{for some proportionality factor } C(\boldsymbol{X}, \boldsymbol{Z}).$$

Approximation 1 is motivated by the theory of composite likelihood estimators (Varin et al., 2011). Specifically, the composite maximum likelihood estimator of $\boldsymbol{\beta}, \boldsymbol{\theta}$ is asymptotically normally distributed around their true values under suitable regularity conditions (Cox and Reid, 2004). Since both the full and the composite maximum likelihood estimators

are asymptotically normal with the same mean, the composite likelihood is approximately proportional to the full likelihood around this mean, with the approximation accuracy depending on the ratio between their variances. This ratio depends on the ratio between the diagonal entries of the Fisher and the Godambe information matrices (Varin et al., 2011). Note that if Approximation 1 holds, this implies that it also holds when replacing $\boldsymbol{s}$ with $\boldsymbol{s}_{-j}$ and omitting j from the product. This property will be used in the derivations below.

Under Equation 5 and Approximation 1, the AEP likelihood $\int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_i t_i(g_i) d\boldsymbol{g}$ is approximately proportional to the ascertained likelihood $P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s})$:

$$\begin{aligned}
\int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_i t_i(g_i) d\boldsymbol{g} &\overset{\text{Equation 5}}{\approx} \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_i \frac{P(y_i, s_i|g_i, \boldsymbol{X}_i)}{P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i})} d\boldsymbol{g} \\
&= \int \frac{P(\boldsymbol{g}|\boldsymbol{Z}) \prod_i P(s_i|g_i, \boldsymbol{X}_i)}{\prod_i P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i})} \prod_i P(y_i|s_i, g_i, \boldsymbol{X}_i) d\boldsymbol{g} \\
&\overset{\text{Approximation 1}}{\approx} C(\boldsymbol{X}, \boldsymbol{Z}) \int \frac{P(\boldsymbol{g}|\boldsymbol{Z}) P(\boldsymbol{s}|\boldsymbol{g}, \boldsymbol{X})}{P(\boldsymbol{s}|\boldsymbol{X}, \boldsymbol{Z})} P(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{X}) d\boldsymbol{g} \\
&= C(\boldsymbol{X}, \boldsymbol{Z}) \int P(\boldsymbol{g}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}) P(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{X}) d\boldsymbol{g} \\
&= C(\boldsymbol{X}, \boldsymbol{Z}) \cdot P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}).
\end{aligned}$$

The last two equalities use the fact that $\boldsymbol{y}$, $\boldsymbol{s}$ are conditionally independent of $\boldsymbol{Z}$ given $\boldsymbol{g}$. We conclude that if $t_i(g_i)$ approximately takes the form of Equation 5 then the hyperparameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ which maximize the AEP likelihood are approximately the maximum likelihood estimates.

Derivation of the AEP Step

Here we provide a heuristic motivation for the AEP step procedure. Recall from Section 3.2 that the AEP step consists of finding the unnormalized Gaussian $t_i(g_i)$ that optimizes the following approximation:

$$\int q_{-i}(g_i) t_i(g_i) dg_i \approx \frac{\int q_{-i}(g_i) P(y_i, s_i|g_i, \boldsymbol{X}_i) dg_i}{\int q_{-i}(g_i) P(s_i|g_i, \boldsymbol{X}_i) dg_i}, \tag{6}$$

where $q_{-i}(g_i) \propto \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_{j \neq i} t_j(g_j) d\boldsymbol{g}_{-i}$, and the optimization is performed by matching the zeroth, first, and second derivatives of both functions with respect to μ_{-i}.

We first write down the natural analogue of the standard EP step objective for AEP. According to Equation 5, this objective finds the unnormalized Gaussian $t_i(g_i)$ that optimizes the approximation:

$$\int q_{-i}(g_i) t_i(g_i) dg_i \approx \int q_{-i}(g_i) \frac{P(y_i, s_i|g_i, \boldsymbol{X}_i)}{P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i})} dg_i. \tag{7}$$

Unlike standard EP, we cannot minimize the GKL divergence between the functions in the integrals in Equation 7, because this will lead to the same solution $t_i(g_i)$ as in standard EP up to a scaling factor. To see this, note that the function in the rhs of Equation 7 can be

written as $q_{-i}(g_i)P(y_i|g_i, \boldsymbol{X}_i)W$, where $W = \frac{P(s_i|y_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})}$ is constant with respect to g_i. Hence, minimizing the GKL divergence will lead to the same approximation as in standard EP, up to the scaling factor W.

Instead of minimizing the GKL divergence, we will approximate the rhs of Equation 7 as follows:

$$\int q_{-i}(g_i)\frac{P(y_i, s_i|g_i, \boldsymbol{X}_i)}{P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i})}dg_i \approx \frac{\int q_{-i}(g_i)P(y_i, s_i|g_i, \boldsymbol{X}_i)dg_i}{\int q_{-i}(g_i)P(s_i|g_i, \boldsymbol{X}_i)dg_i}. \tag{8}$$

The AEP step in Equation 6 is obtained by equating the lhs of Equation 7 with the rhs of Equation 8.

It remain to derive Equation 8. Our derivation uses the following assumption:

Assumption 1 *Weak dependence between $\boldsymbol{y}_{-i}$ and s_i conditional on $\boldsymbol{X}$ and on $\boldsymbol{Z}$:*

$$P(\boldsymbol{y}_{-i}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}) \approx P(\boldsymbol{y}_{-i}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s}_{-i}).$$

The derivation additionally uses the following two approximations, which we derive below by using Equation 5, Approximation 1 and Assumption 1:

Approximation 2 *At the fixed point we have: $q_{-i}(g_i) \approx P(g_i, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})$.*

Approximation 3 $\int q_{-i}(g_i)\frac{P(y_i,s_i|g_i,\boldsymbol{X}_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})}dg_i \approx P(y_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i}, s_i)$.

We complete the derivation of Equation 8 by first using Approximation 3 to approximate the rhs of Equation 7 and the lhs of Equation 8 as $P(y_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i}, s_i)$, and then using Approximation 2 and the graphical model structure (Figure 2) to approximate $P(y_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i}, s_i)$ via $q_{-i}(g_i)$ as follows:

$$\begin{aligned} P(y_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i}, s_i) &= \frac{P(y_i, s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})}{P(s_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})} \\ &= \frac{\int P(g_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})P(y_i, s_i|\boldsymbol{X}_i, g_i)dg_i}{\int P(g_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})P(s_i|\boldsymbol{X}_i, g_i)dg_i} \\ &\approx \frac{\int q_{-i}(g_i)P(y_i, s_i|\boldsymbol{X}_i, g_i)dg_i}{\int q_{-i}(g_i)P(s_i|\boldsymbol{X}_i, g_i)dg_i}. \end{aligned}$$

This completes the derivation.

Derivation of Approximations 2–3

We now provide heuristic derivations of Approximations 2–3.

Approximation 2 $q_{-i}(g_i) \approx P(g_i, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}_{-i})$.

Our derivation consists of two stages. First, we define the unnormalized cavity distribution $q^*_{-i}(g_i) \triangleq \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_{j\neq i} t_j(g_j) d\boldsymbol{g}_{-i}$, and show that $q^*_{-i}(g_i) \approx C(\boldsymbol{X},\boldsymbol{Z}) \cdot P(g_i, \boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})$:

$$
\begin{aligned}
q^*_{-i}(g_i) &\triangleq \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_{j\neq i} t_j(g_j) d\boldsymbol{g}_{-i} \\
&\overset{\text{Equation 5}}{\approx} \int P(\boldsymbol{g}|\boldsymbol{Z}) \prod_{j\neq i} \frac{P(y_j, s_j|g_j, \boldsymbol{X}_j)}{P(s_j|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-j})} d\boldsymbol{g}_{-i} \\
&\overset{\text{Approximation 1}}{\approx} C(\boldsymbol{X},\boldsymbol{Z}) \int \frac{P(\boldsymbol{g}|\boldsymbol{Z})}{P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})} \prod_{j\neq i} P(y_j, s_j|g_i, \boldsymbol{X}_j) d\boldsymbol{g}_{-i} \\
&= C(\boldsymbol{X},\boldsymbol{Z}) \int \frac{P(\boldsymbol{g}_{-i}|\boldsymbol{Z}) P(g_i|\boldsymbol{g}_{-i},\boldsymbol{Z})}{P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})} P(\boldsymbol{y}_{-i}, \boldsymbol{s}_{-i}|\boldsymbol{g}_{-i}, \boldsymbol{X}_{-i}) d\boldsymbol{g}_{-i} \\
&\overset{\text{rearrangement}}{=} C(\boldsymbol{X},\boldsymbol{Z}) \int \frac{P(\boldsymbol{g}_{-i}|\boldsymbol{Z}) P(\boldsymbol{s}_{-i}|\boldsymbol{g}_{-i}, \boldsymbol{X}_{-i})}{P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})} P(\boldsymbol{y}_{-i}, |\boldsymbol{g}_{-i}, \boldsymbol{s}_{-i}, \boldsymbol{X}_{-i}) P(g_i|\boldsymbol{g}_{-i},\boldsymbol{Z}) d\boldsymbol{g}_{-i} \\
&\overset{\text{Bayes rule}}{=} C(\boldsymbol{X},\boldsymbol{Z}) \int P(\boldsymbol{g}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i}) P(\boldsymbol{y}_{-i}, |\boldsymbol{g}_{-i}, \boldsymbol{s}_{-i}, \boldsymbol{X}_{-i}) P(g_i|\boldsymbol{g}_{-i},\boldsymbol{Z}) d\boldsymbol{g}_{-i} \\
&\overset{\text{graphical model}}{=} C(\boldsymbol{X},\boldsymbol{Z}) \int P(\boldsymbol{g}_{-i}, \boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i}) P(g_i|\boldsymbol{g}_{-i},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i}) d\boldsymbol{g}_{-i} \\
&= C(\boldsymbol{X},\boldsymbol{Z}) \int P(\boldsymbol{g}, \boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i}) d\boldsymbol{g}_{-i} \\
&= C(\boldsymbol{X},\boldsymbol{Z}) \cdot P(g_i, \boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})
\end{aligned}
$$

Next, we note that since $q_{-i}(g_i) \triangleq \frac{q^*_{-i}(g_i)}{\int q^*_{-i}(g'_i) dg'_i}$ is a normalized distribution over g_i, we have $q_{-i}(g_i) \approx P(g_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},\boldsymbol{s}_{-i})$. Finally, we note that g_i is conditionally independent of $\boldsymbol{s}_{-i}$ given $\boldsymbol{y}_{-i}$ due to the graphical model structure, yielding $q_{-i}(g_i) \approx P(g_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i})$.

Approximation 3 $\int q_{-i}(g_i) \frac{P(y_i,s_i|g_i,\boldsymbol{X}_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} dg_i \approx P(y_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},s_i)$.

First, we invoke Approximation 2 and the graphical model structure to obtain the following approximation:

$$
\begin{aligned}
\int q_{-i}(g_i) \frac{P(y_i,s_i|g_i,\boldsymbol{X}_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} dg_i &\overset{\text{Approximation 2}}{\approx} \int P(g_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i}) \frac{P(y_i,s_i|g_i,\boldsymbol{X}_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} dg_i \\
&= \int \frac{P(g_i|\boldsymbol{X},\boldsymbol{Z}) P(\boldsymbol{y}_{-i}|g_i,\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z})} \frac{P(y_i,s_i|g_i,\boldsymbol{X}_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} dg_i \\
&= \int \frac{P(g_i|\boldsymbol{X},\boldsymbol{Z}) P(\boldsymbol{y}_{-i}|g_i,\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z})} \frac{P(y_i|g_i,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i}) P(s_i|y_i)}{P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} dg_i \\
&= \frac{\int P(g_i|\boldsymbol{X},\boldsymbol{Z}) P(\boldsymbol{y}|g_i,\boldsymbol{X},\boldsymbol{Z}) dg_i P(s_i|y_i)}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z}) P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} \\
&= \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}) P(s_i|y_i)}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z}) P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})}.
\end{aligned} \tag{9}
$$

Next, we multiply the rhs of Equation 9 by $\frac{P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})}$ and invoke Assumption 1:

$$\begin{aligned}
\frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z})P(s_i|y_i)}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z})P(s_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})}\frac{P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})} &= \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z})}\frac{P(\boldsymbol{s}|\boldsymbol{y})P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{s}|\boldsymbol{X},\boldsymbol{Z})P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})} \\
&= \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z})\frac{P(\boldsymbol{s}|\boldsymbol{y})}{P(\boldsymbol{s}|\boldsymbol{X},\boldsymbol{Z})}}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z})\frac{P(\boldsymbol{s}_{-i}|\boldsymbol{y}_{-i})}{P(\boldsymbol{s}_{-i}|\boldsymbol{X},\boldsymbol{Z})}} \\
&= \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} \\
&= \frac{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s})P(y_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},\boldsymbol{s})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} \\
&\overset{\text{Assumption 1}}{\approx} \frac{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})P(y_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},\boldsymbol{s})}{P(\boldsymbol{y}_{-i}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{s}_{-i})} \\
&= P(y_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},\boldsymbol{s}) \\
&= P(y_i|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}_{-i},s_i)
\end{aligned}$$

Appendix C

Here we describe our novel development of ascertained generalized estimating equations (AGEE). GEEs are extensions of generalized linear models that can estimate fixed effects while accounting for dependencies without requiring a probabilistic model (Liang and Zeger, 1993). GEEs require a correct specification of the mean of the outcome conditional on the features, $\mu_i = E\left[y_i|\boldsymbol{X}_i,\boldsymbol{\beta}\right]$, and a (possibly misspecified) working covariance matrix of the outcomes, denoted as $\boldsymbol{\Omega}\left(\theta^{\Omega}\right)$ and parameterized by θ^{Ω}. Given these, $\boldsymbol{\beta}$ is estimated by solving the estimating equation $\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}}\boldsymbol{\Omega}\left(\theta^{\Omega}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\left(\boldsymbol{\beta}\right)\right)=0$. GEEs yield consistent estimates of $\boldsymbol{\beta}$ and its sampling variance even if the covariance structure is misspecified and is non-sparse (Xie and Yang, 2003).

GEEs can naturally be adapted to case-control settings by using the ascertained conditional mean function $E\left[y_i|\boldsymbol{X}_i,s_i=1,\boldsymbol{\beta}\right]=P\left(y_i=1\,|\,\boldsymbol{X}_i,s_i=1,\boldsymbol{\beta}\right)$. We now show how the GEE fixed effect estimates can be plugged into GPs. In the general case it is not possible to reconcile fixed effect estimates of GEEs and GPs, because GEEs assume that the conditional mean of the outcome is affected only by the fixed effects, whereas GPs assume that it is affected by both the fixed effects and the GP latent variable. Fortunately, the probit likelihood provides a convenient way to reconcile the two approaches. Denote $\boldsymbol{\beta}_{\text{GEE}}$ and $\boldsymbol{\beta}_{\text{GP}}$ as the vectors of fixed effects used by GEE and GP, respectively. When using a probit likelihood, the GEE conditional mean is given by $\Phi(\boldsymbol{X}_i^T\boldsymbol{\beta}_{\text{GEE}})$, where $\Phi(\cdot)$ is the standard normal cumulative density. In contrast, the GP conditional mean is given by $\Phi\left(\frac{\boldsymbol{X}_i^T\boldsymbol{\beta}_{\text{GP}}}{(\text{var}(g_i)+1)^{1/2}}\right)$. If $\text{var}(g_i)$ is constant for every unit i (which corresponds to a constant value on the diagonal of the covariance matrix of $\boldsymbol{g}$), the two approaches can be reconciled by defining $\boldsymbol{\beta}_{\text{GP}}=\boldsymbol{\beta}_{\text{GEE}}\left(\text{var}(g_i)+1\right)^{1/2}$. In practice, the diagonal of the covariance matrix of $\boldsymbol{g}$ is often exactly or almost exactly constant, which enables exploiting the above relation. Therefore, we can use the GEE estimates in a GP by setting $\boldsymbol{\beta}_{\text{GP}}=\boldsymbol{\beta}_{\text{GEE}}\left(\text{var}(g_i)+1\right)^{1/2}$.

Our implementation of AGEE closely followed that of (Jiang et al., 2016a), with a suitable modification of the conditional mean to encode ascertainment, as described above.

References

Scarlett L. Bellamy, Yi Li, Xihong Lin, and Louise M. Ryan. Quantifying PQL bias in estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials. *Stat. Sin.*, 15(4):1015–1032, 2005.

Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24(3):127–35, 2009.

Anna Bonnet, Elisabeth Gassiat, and Cline Lvy-Leduc. Heritability estimation in high dimensional sparse linear mixed models. *Electron. J. Stat.*, 9(2):2099–2129, 2015.

Norman E. Breslow. Statistics in epidemiology: the case-control study. *J. Am. Stat. Assoc.*, 91(433):14–28, 1996.

Norman E. Breslow and David G. Clayton. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, 88(421):9–25, 1993.

Paul R. Burton, Katrina J. Tiller, Lyle C. Gurrin, William OCM Cookson, A. William Musk, and Lyle J. Palmer. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genet. Epidemiol.*, 17(2):118–140, 1999.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared OConnell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

Han Chen, Chaolong Wang, Matthew P. Conomos, Adrienne M. Stilp, Zilin Li, Tamar Sofer, Adam A. Szpiro, Wei Chen, John M. Brehm, Juan C. Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.*, 98(4):653–666, 2016.

Ching-An Cheng and Byron Boots. Variational inference for Gaussian process models with linear complexity. In *Advances in Neural Information Processing Systems*, pages 5184–5194, 2017.

Carlton Chu, Peter Bandettini, John Ashburner, Andre Marquand, and Stefan Kloeppel. Classification of neurodegenerative diseases using Gaussian process classification with automatic feature determination. In *Workshop on brain decoding: Pattern recognition challenges in neuroimaging (WBD), 2010*, pages 17–20. IEEE, 2010.

David R. Cox and Nancy Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

Guillaume Dehaene and Simon Barthelmé. Bounding errors of Expectation-Propagation. In *Advances in Neural Information Processing Systems*, pages 244–252, 2016.

Guillaume Dehaene and Simon Barthelmé. Expectation Propagation in the large data limit. *J. R. Stat. Soc. B*, 80(1):199–217, 2018.

Marc Deisenroth and Jun Wei Ng. Distributed Gaussian processes. In *International Conference on Machine Learning*, volume 37, pages 1481–1490, 07–09 Jul 2015.

Everett R. Dempster and I. Michael Lerner. Heritability of threshold characters. *Genetics*, 35(2):212–36, 1950.

Lee H. Dicker and Murat A. Erdogdu. Maximum likelihood for variance estimation in high-dimensional linear models. In *International Conference on Artificial Intelligence and Statistics*, pages 159–167, 2016.

Peter J. Diggle, Jonathan A. Tawn, and Rana Moyeed. Model-based geostatistics. *J. R. Stat. Soc. C*, 47(3):299–350, 1998.

Juan Du, Hao Zhang, and Vidyadhar Mandrekar. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Stat.*, 37(6A):3330–3361, 2009.

Georg B. Ehret, Patricia B. Munroe, Kenneth M. Rice, Murielle Bochud, Andrew D. Johnson, Daniel I. Chasman, Albert V. Smith, Martin D. Tobin, Germaine C. Verwoert, Shih-Jen Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

Michael P. Epstein, Xihong Lin, and Michael Boehnke. Ascertainment-adjusted parameter estimates revisited. *Am. J. Hum. Genet.*, 70(4):886–95, 2002.

Ludwig Fahrmeir and Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer New York, Berlin, 2nd edition, 2001. ISBN 978-0-387-95187-4.

Egil Ferkingstad and Håvard Rue. Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electron. J. Stat.*, 9(2):2706–2731, 2015.

Youyi Fong, Håvard Rue, and Jon Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412, 2010.

Jacob Gardner, Geoff Pleiss, Ruihan Wu, Kilian Weinberger, and Andrew Wilson. Product kernel interpolation for scalable Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1407–1416, 2018.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13(2):163–185, 1998.

David V. Glidden and Kung-Yee Liang. Ascertainment adjustment in complex diseases. *Genet. Epidemiol.*, 23(3):201–208, 2002.

David V. Glidden and Eric Vittinghoff. Modelling clustered survival data from multicentre clinical trials. *Stat. Med.*, 23(3):369–88, 2004.

David Golan and Saharon Rosset. Effective genetic-risk prediction using mixed models. *Am. J. Hum. Genet.*, 95(4):383–93, 2014.

David Golan, Eric Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA*, 111(49):E5272–81, 2014.

Tristan J. Hayeck, Noah A. Zaitlen, Po-Ru Loh, Bjarni Vilhjalmsson, Samuela Pollack, Alexander Gusev, Jian Yang, Guo-Bo Chen, Michael E. Goddard, Peter M. Visscher, et al. Mixed model with correction for case-control ascertainment increases association power. *Am. J. Hum. Genet.*, 96(5):720–730, 2015.

Patrick J Heagerty and Subhash R Lele. A composite likelihood approach to binary spatial data. *J. Am. Stat. Assoc.*, 93(443):1099–1111, 1998.

Leonhard Held, Isabel Natrio, Sarah Elaine Fenton, Håvard Rue, and Nikolaus Becker. Towards joint disease mapping. *Stat. Methods Med. Res.*, 14(1):61–82, 2005.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. MCMC for variationally sparse Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1648–1656, 2015.

David A Hsieh, Charles F Manski, and Daniel McFadden. Estimation of response probabilities from augmented retrospective observations. *J. Am. Stat. Assoc.*, 80(391):651–662, 1985.

Duo Jiang, Joelle Mbatchou, and Mary Sara McPeek. Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. *Hum. Hered.*, 80(4):187–195, 2015.

Duo Jiang, Sheng Zhong, and Mary Sara McPeek. Retrospective binary-trait association test elucidates genetic architecture of Crohn disease. *Am. J. Hum. Genet.*, 98(2):243–255, 2016a.

Jiming Jiang, Cong Li, Debashis Paul, Can Yang, and Hongyu Zhao. On high-dimensional misspecified mixed model analysis in genome-wide association study. *Ann. Stat.*, 44(5): 2127–2160, 2016b.

Julia E Kelsall and Peter J. Diggle. Spatial variation in risk of disease: a nonparametric binary regression approach. *J. Royal Stat. Soc. C*, 47(4):559–573, 1998.

H. König. *Eigenvalue Distribution of Compact Operators*. Operator Theory: Advances and Applications. Birkhäuser Basel, 2013. ISBN 9783034862783.

Holly J Kramer, Adrienne M. Stilp, Cathy C. Laurie, Alex P. Reiner, James Lash, Martha L. Daviglus, Sylvia E. Rosas, Ana C. Ricardo, Bamidele O. Tayo, Michael F. Flessner, et al. African ancestry–specific alleles and kidney disease risk in Hispanics/Latinos. *J. Am. Soc. Nephrol.*, 28(3):915–922, 2017.

Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary Gaussian process classification. *J. Mach. Learn. Res.*, 6:1679–1704, 2005.

Quoc Le, Tamas Sarlos, and Alexander Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *International Conference on Machine Learning*, pages 244–252, 2013.

Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, 88(3):294–305, 2011.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Kung-Yee Liang and Scott L Zeger. Regression analysis for correlated data. *Annu. Rev. Public Health*, 14(1):43–68, 1993.

Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.*, 47(12):1385–1392, 2015.

Charles F. Manski. *Alternative estimators and sample designs for discrete choice analysis*. The MIT Press, 1981.

Charles E McCulloch, Shayle R Searle, and John M Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics, 2nd edition, 2008. ISBN 0-470-01181-5.

Thomas P Minka. Expectation Propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, pages 362–369, 2001. ISBN 1-55860-800-1.

J. M. Neuhaus, Alastair J. Scott, Chris J. Wild, Y. Jiang, C. E. McCulloch, and R. Boylan. Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics*, 70(1):44–52, 2014.

John M. Neuhaus and Nicholas P. Jewell. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, 46(4):977–990, 1990.

John M. Neuhaus, Alastair H. Scott, and Chris J. Wild. The analysis of retrospective family studies. *Biometrika*, 89(1):23–37, 2002.

John M. Neuhaus, Alastair J. Scott, and Chris J. Wild. Family-specific approaches to the analysis of case-control family data. *Biometrics*, 62(2):488–94, 2006.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary Gaussian process classification. *J. Mach. Learn. Res.*, 9:2035–2078, 2008.

Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Comput*, 21(3):786–92, 2009.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Comput*, 12(11):2655–2684, 2000.

Dirk Pfeiffer. *Spatial Analysis in Epidemiology*. Oxford University Press, 2008.

Alkes L. Price, Chris C. A. Spencer, and Peter Donnelly. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci*, 282(1821), 2015.

Qibin Qi, Adrienne M Stilp, Tamar Sofer, Jee-Young Moon, Bertha Hidalgo, Adam A Szpiro, Tao Wang, Maggie CY Ng, Xiuqing Guo, Yii-Der Ida Chen, et al. Genetics of type 2 diabetes in US Hispanic/Latino individuals: results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Diabetes*, 66(5):1419–1425, 2017.

Yuan Alan Qi, Thomas P Minka, Rosalind W Picard, and Zoubin Ghahramani. Predictive automatic relevance determination by Expectation Propagation. In *International Conference on Machine Learning*, page 85, 2004.

Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190, 2004.

Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econom.*, 128(2):301–323, 2005.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ISBN 978-0-262-18253-9.

Stephen W. Raudenbush, Meng-Li Yang, and Matheos Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comp. Graph. Stat.*, 9(1):141–157, 2000.

Didier Renard, Geert Molenberghs, and Helena Geys. A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Stat. Data Anal.*, 44(4):649–667, 2004.

Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris CA Spencer, Nikolaos A Patsopoulos, Loukas Moutsianas, Alexander Dilthey, Zhan Su, Colin Freeman, Sarah E Hunt, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–9, 2011.

Alastair J Scott and Chris J Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

Alastair J. Scott and Chris J. Wild. Maximum likelihood for generalised case-control studies. *J. Stat. Plan. Inference*, 96(1):3–27, 2001.

Matthias Seeger. Low rank updates for the cholesky decomposition. Technical report, 2004.

Matthias Seeger. Expectation propagation for exponential families. Technical report, 2005.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *J. R. Stat. Soc. B*, 57(4):749–760, 1995.

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

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.*, 81(393):82–86, 1986.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Stat. Sin.*, 21(1):5–42, 2011.

Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *J. Mach. Learn. Res.*, 17(1):3581–3618, 2016.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, 101(1):5–22, 2017.

Omer Weissbrod, Christoph Lippert, Dan Geiger, and David Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nat. Methods*, 12(4):332–4, 2015.

Omer Weissbrod, Dan Geiger, and Saharon Rosset. Multikernel linear mixed models for complex phenotype prediction. *Genome Res.*, 26(7):969–79, 2016.

Omer Weissbrod, Jonathan Flint, and Saharon Rosset. Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.*, 103(1):89–99, 2018.

Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Chris J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78 (4):705–717, 1991.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, pages 1775–1784, 2015.

Andrew G Wilson, Zhiting Hu, Ruslan R Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems*, pages 2586–2594, 2016.

Russ Wolfinger and Michael O'connell. Generalized linear mixed models: a pseudo-likelihood approach. *J. Stat. Comput. Simul.*, 48(3-4):233–243, 1993.

Minge Xie and Yaning Yang. Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Stat.*, 31(1):310–347, 2003.

Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G Martin, Grant W. Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, 2010.

Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, 46(2):100–106, 2014.

Zichao Yang, Andrew Wilson, Alex Smola, and Le Song. A la Carte – learning fast kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 1098–1106, 2015.

Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, the Alzheimer's Disease Neuroimaging Initiative, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neuroimage Clin.*, 2:735–745, 2013.

Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Stat. Assoc.*, 99(465):250–261, 2004.

Hao Zhang and Dale L. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936, 2005.

Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, 9(2):e1003264, 2013.

Gabriel Ziegler, Gerard R. Ridgway, Robert Dahnke, Christian Gaser, the Alzheimer's Disease Neuroimaging Initiative , et al. Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. *Neuroimage*, 97:333–348, 2014.