Light-speed Inference of Genetic Effects via Approximate Message Passing

Al Depope[1,*], Marco Mondelli[1,†,*], Matthew R. Robinson[1,†,*]

[1] Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria.

* corresponding author: al.depope@ist.ac.at, marco.mondelli@ist.ac.at, matthew.robinson@ist.ac.at

† indicates joint supervision

Abstract

Efficient utilization of large-scale biobank data is crucial for inferring the genetic basis of disease and predicting health outcomes from the DNA. Yet we lack accurate methods that scale to data where electronic health records are linked to whole genome sequence information. To address this issue, we develop a new algorithmic paradigm based on Approximate Message Passing (AMP), which is specifically tailored for genomic prediction and association testing. Our method yields comparable out-of-sample prediction accuracy to the state of the art on UK Biobank traits, whilst dramatically improving computational complexity, with a 10x-speed up in the run time. In addition, AMP theory provides a joint association testing framework, which outperforms the commonly used REGENIE method, in a third of the compute time. This first, truly large-scale application of the AMP framework lays the foundations for a far wider range of statistical analyses for hundreds of millions of variables measured on millions of people.

Introduction

Large-scale biobank data will link the whole genome sequence of hundreds of thousands of people to electronic health record outcomes [1]. These massive datasets have the potential to generate great biological insight, but they are currently being generated more abundantly than they can efficiently be analyzed. Generally, classical statistical theory is relied upon [2,3] for two main learning problems: *(i)* estimation and statistical testing of the relationship between genome and health outcome, and *(ii)* prediction of health outcomes from the genome. For both tasks, restricted maximum likelihood [4], ridge regression [5], variational inference [6], or Gibbs sampling [7] approaches are commonly used. However, existing algorithms are forced to restrict estimation to limited sets of genetic loci, or would require months to model the whole genome sequence data of ~ 140 million DNA variants alongside thousands of health outcomes.

A line of work at the interface of high dimensional statistics, information theory and statistical physics has focused on a set of prototypical problems in which both the number of measurements and the dimension of the quantity being estimated are very large. The fundamental limits of inference have been precisely characterized, and it has been shown that efficient algorithms, such as Approximate Message Passing (AMP), can meet such limits [8, 9]. Specifically, AMP represents a family of iterative algorithms that has been applied to a range of statistical estimation problems, including

linear regression [10, 11], generalised linear models [12–14], and low-rank matrix estimation [9, 15]. The popularity of AMP stems from its unique attractive features: *(i)* AMP can be tailored to take advantage of structural information available about the signal (i.e., a wide range of Bayesian priors can be used); *(ii)* under mild model assumptions, the AMP performance in the high-dimensional limit is precisely characterized by a succinct deterministic recursion called state evolution [11]; and *(iii)* using state evolution, it has been proved that AMP achieves Bayes-optimal performance in a number of settings [8, 9]. Applications of AMP include wireless communications [16], compressive imaging [17], MRI [18], and PCA in genetics [19]. Biobank-scale genome-wide association studies are much more extreme in their dimensionality than these aforementioned applications, and within these data regimes, AMP implementations have remained unexplored.

This paper introduces a novel family of AMP algorithms tailored for datasets that integrate electronic health records with genomic data for millions of individuals. The sheer scale of the data, as well as the fact that the typical assumptions required by the theory (i.i.d. Gaussian entries, or rotational invariance) may not hold, lead to a number of algorithmic issues. By solving them, we show that AMP provides excellent performance for both association testing and prediction in *light-speed time*: our running times for large-scale biobank data are of the order of a few hours, as opposed to the several days/weeks of current approaches.

Methods

Overview of the model

In a study involving N individuals and P single nucleotide polymorphisms (SNPs), the data is typically stored in the so-called genotype matrix whose entries are encoded to a set $\{0, 1, 2\}$ depending on homozygosity and allele frequencies of each particular SNP (column). Its normalized version $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ – with each column having zero mean and unit variance – is called the normalized genotype matrix.

Bayesian linear regression is the most widely used model in the life sciences, and it allows for information on the signal structure to be incorporated in the inference process. The inference problem consists of estimating the signal vector $\boldsymbol{\beta} \in \mathbb{R}^P$ from a vector of phenotype measurements $\boldsymbol{y} \in \mathbb{R}^N$ that satisfies the following linear relationship:

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \epsilon_i, \qquad \text{for } i \in \{1, \ldots, N\}.$$

Here, $\boldsymbol{x}_i$ is a row vector of the normalized genotype matrix $\boldsymbol{X}$, $\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle = \boldsymbol{x}_i^T \boldsymbol{\beta}$ denotes the Euclidean inner product, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is an unknown noise vector that is assumed to follow the multivariate normal distribution $\mathrm{MVN}(0, \gamma_\epsilon^{-1} \cdot \boldsymbol{I})$ with unknown noise precision parameter γ_ϵ^{-1}. To allow for a range of genetic effects, from heavy-tail distributions to sparse, we model the prior on the signal $\boldsymbol{\beta}$ to be of a spike-and-slab form. Namely, for every $i = 1, \ldots, P$, we model

$$\beta_i \sim (1 - \lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{i=1}^{L} \pi_i \cdot \mathcal{N}(\cdot, 0, \sigma_i^2),$$

Algorithm 1 EM-VAMP

Input: max number of iterations N_{it}
initial noisy estimate $\boldsymbol{r}_{10} \in \mathbb{R}^P$
initial signal precision estimate $\gamma_{10} > 0$
initial set of parameters defining prior distribution Θ_0
threshold for stopping criterion $\varepsilon > 0$
for $t = 0, 1, \ldots, N_{\text{it}}$ **do**
  Denoising step
  EM update of the prior distribution parameters Θ, called Θ_t
  $\hat{\boldsymbol{\beta}}_{1t} = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{r}_{1t} = \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1t}^{-1}\boldsymbol{I}), \gamma_{1t}, \Theta_t]$
  $\alpha_{1t} = \gamma_{1t} \cdot \langle \text{Var}[\boldsymbol{\beta}|\boldsymbol{r}_{1t} = \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1t}^{-1}\boldsymbol{I}), \gamma_{1t}, \Theta_t]\rangle$
  $\gamma_{2t} = \gamma_{1t} \cdot (1 - \alpha_{1t})/\alpha_{1t}$
  $\boldsymbol{r}_{2t} = (\hat{\boldsymbol{\beta}}_{1t} - \alpha_{1t}\boldsymbol{r}_{1t})/(1 - \alpha_{1t})$
  LMMSE step
  EM update of the estimate of γ_ϵ, called $\gamma_{\epsilon,t}$
  $\hat{\beta}_{2t} = (\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2t}\boldsymbol{I})^{-1}(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{y} + \gamma_{2t}\boldsymbol{r}_{2t})$
  $\alpha_{2t} = \gamma_{2t} \cdot \text{Tr}[(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_2\boldsymbol{I})^{-1}]/P$
  $\gamma_{1,t+1} = \gamma_{2t} \cdot (1 - \alpha_{2t})/\alpha_{2t}$
  $\boldsymbol{r}_{1,t+1} = (\hat{\boldsymbol{\beta}}_{2t} - \alpha_{2t}\boldsymbol{r}_{2t})/(1 - \alpha_{2t})$
  If $t \geq 1$ and $||\hat{\boldsymbol{\beta}}_{1t} - \hat{\boldsymbol{\beta}}_{1,t-1}||_2/||\hat{\boldsymbol{\beta}}_{1,t-1}||_2 < \varepsilon$, then **break**
end for
return $\hat{\boldsymbol{\beta}}_{1t}$

where $\lambda \in [0, 1]$ is the sparsity rate; $L \in \mathbb{N}$, $(\pi_i)_{i=1}^L$ and $(\sigma_i^2)_{i=1}^L$ denote, respectively, the fixed number of Gaussians mixtures, the mixture probabilities and the variances for the slab component. We learn these parameters from the data in an adaptive Bayes expectation-maximisation (EM) framework. This avoids expensive cross-validation and yields biologically informative inferences about signal sparsity (polygenicity) and phenotypic variance attributable to the genomic data. We highlight that the problem is high dimensional, in the sense that both P and N can be very large, with the former going up to roughly 140 million DNA pairs and the latter going up to half a million individuals.

Overview of the proposed gVAMP

The expectation-maximization vector AMP algorithm (EM-VAMP) introduced in [20, 21] is an iterative procedure consisting of two steps: *(i)* denoising, and *(ii)* linear minimum mean square error estimation (LMMSE). The denoising step accounts for the prior structure given a noisy estimate of the signal $\boldsymbol{\beta}$, while the LMMSE step utilizes phenotype values to further refine the estimate. The key feature of EM-VAMP is in the so called *Onsager correction*: this is added to ensure the asymptotic normality of the noise corrupting the estimates of $\boldsymbol{\beta}$ given by the algorithm at every iteration. This property then allows a precise performance analysis via *state evolution* and, consequently, the optimization of the denoisers. Specifically, by taking the conditional expectation with respect to the posterior marginal density given the estimates on the signal and the noise precision, the Bayes-optimal performance is provably achieved in some settings [20]. Algorithm 1 provides a succinct description of the method.

Previous work on AMP has mostly considered relatively low-scale setups [16–19]. Here, we

investigate practices both from the AMP and the genome-wide association study (GWAS) literature in a way that is tailored for inference of genetic effects. Our efforts tackle three different aspects: *(i)* exploring genotype matrix preprocessing procedures that give synergistic effects with the AMP framework; *(ii)* developing techniques that prevent the algorithm from diverging, while keeping it fast and suitable for large-scale applications; and *(iii)* optimizing the code using tools, such as bitwise operations and lookup-tables, that reduce the final memory and run time requirements, to accommodate truly large-scale genomic datasets.

The blind application of the EM-VAMP algorithm to the UK Biobank dataset typically leads to diverging estimates of the signal. However, different approaches are known in the AMP community to mitigate the fact that the channel matrix might not conform to the right-rotationally invariance property[1] and to scenarios where noise precisions are misestimated: row or/and column normalization, damping (i.e., doing convex combinations of new and previous estimates) [22], and variance auto-tuning [21]. In particular, to prevent EM-VAMP from diverging and ensure it follows its state evolution, we empirically observed that the combination of the following techniques is crucial. First, we perform *damping* in the space of denoised signals, i.e., for $t \geq 1$ and fixed $\rho \in (0,1)$, we replace line 5 of Algorithm 1 with $\hat{\boldsymbol{\beta}}_{1t} = \rho \cdot \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{r}_{1t}, \Theta_{1t}] + (1-\rho) \cdot \hat{\boldsymbol{\beta}}_{1,t-1}$. We also perform damping of the Onsager correction, i.e., we replace line 6 of Algorithm 1 with $\alpha_{1t} = \rho \cdot \gamma_{1t} \langle \mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{r}_{1t}, \gamma_{1t}, \Theta_{1t}] \rangle + (1-\rho) \cdot \alpha_{1,t-1}$. Second, we perform *auto-tuning* of γ_{1t} via the approach from [21] before step 1. Third, we *filter* the design matrix for first-degree relatives to reduce the correlation between rows. As the design matrix contains DNA, this standard practice in genomics involves simply randomly selecting one individual (row) from any pair correlated ≥ 0.125.

We find that reducing the maximum correlation between SNPs (columns) in the design matrix often helps convergence and performance. While this limits signal detection to one SNP from a set with correlation ≥ 0.95, this is acceptable as it is anyway infeasible to distinguish among these in the data. Additionally, initializing the sparsity parameter to $50,000$ included signals, setting the mixture variance and probabilities to follow a geometric series, and setting the damping factor to 0.3 performs well in the UK Biobank, with removed first-degree relatives and correlations between markers of up to 0.95. For the rest, we successfully run with damping of 0.1.

Our open-source **gVAMP** software, implemented in C++, incorporates parallelization using the OpenMP and MPI libraries. Moreover, data streaming is employed using a lookup table, enabling byte-by-byte processing of the genotype matrix, see Figure 1.

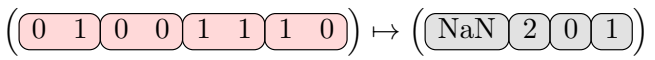

Figure 1. The entries of the unnormalized genotype matrix are encoded with values in a set $\{0, 1, 2\}$. Hence, the lookup table enables streaming in the data in bytes, where every byte (8 bits) encodes the information of 4 individuals.

Given the right computer architecture, this approach not only accelerates computation by using SIMD instructions and handling four consecutive entries of the design matrix simultaneously, but also eliminates the need for extra memory.

[1] A matrix is right-orthogonally invariant if the matrix containing its right singular vectors is distributed according to the Haar measure, i.e., it is uniform in the group of orthogonal matrices.

Results

Prediction: Polygenic risk scores

We show that our gVAMP algorithm achieves comparable out-of-sample prediction accuracy R^2 as an individual-level Gibbs sampler [7], optimised for genomic data with the same prior, across different sets of genomic markers (Figure 2) and phenotypes (Table 1). To ensure mixing and stabilization of out-of-sample prediction in the Gibbs sampler approach, typically at least $2,500$ iterations are performed [7]. In contrast, gVAMP reaches its maximal out-of-sample prediction usually within 15 iterations. Using 60 CPUs with 8GB of memory, gVAMP required less than 15 hours of run time for 2.17 million SNPs and $400,000$ individuals, while the Gibbs sampler took around 7 days. This leads to a speed-up of more than 10 times.

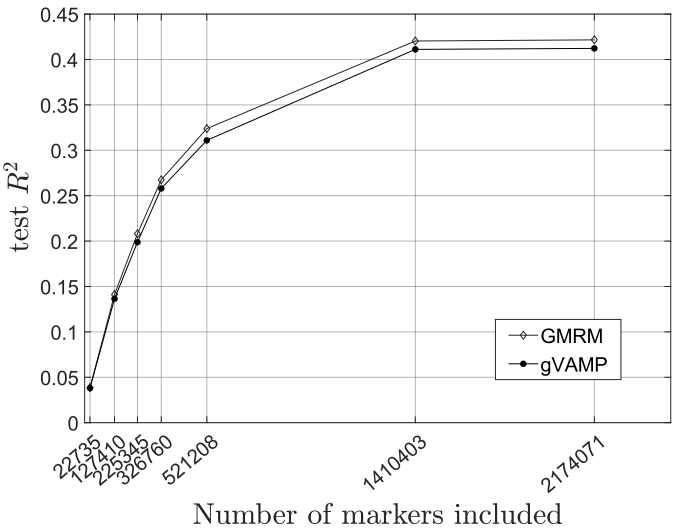

Figure 2. Coefficient of determination (R^2) values obtained from a hold-out dataset consisting of $15,000$ samples using varying numbers of SNPs included in the inference process for human height measured on $399,055$ individuals from the UK Biobank. gVAMP is halted after 10 iterations, while the state-of-the-art Gibbs sampler approach (GMRM) is run for 400 burn-in iterations and then until the accuracy stabilises [7].

Association testing: False positive rate and statistical power

In current approaches, frequentist association testing involves approximations of Henderson's mixed-linear model equation [4], represented as follows: $\boldsymbol{y} = \sum_{j \in S} \boldsymbol{x}_j \beta_j + \hat{\boldsymbol{g}}_{\notin S} + \boldsymbol{\epsilon}$,[2] where the set S contains either DNA variants on a focal chromosome (leave-one-chromosome-out approach, see [5]) or at a single loci (leave-one-out approach). The regression coefficient at each position is re-estimated using least squares and p-values are obtained for the test that an effect is equal to zero. However, these re-estimations do not take into account linkage disequilibrium between SNPs on the same chromosome to correctly weight their contributions to the phenotypic variance.

Here, we present an alternative approach based on the full vector of estimated regression coefficients. Our method relies on the properties of the EM-VAMP estimator, whose noise is

[2] For a set $S \subseteq \{1, \ldots, P\}$, $\hat{\boldsymbol{g}}_{\notin S}$ denotes a genetic prediction from all loci not in S.

Table 1. Out-of-sample R^2 on 6 different traits from the UK Biobank dataset such as cholesterol (CHOL), HDL cholesterol (HDL), standing height (HT), mean corpuscular haemoglobin (MCH), red blood cell count (RBC) and systolic blood pressure (SBP). Both gVAMP and GMRM were run using a genotype matrix containing $2,174,071$ SNPs. GMRM was run for $2,000$ iterations with 200 burn-in, while gVAMP was run for 35 iterations and maximum test R^2 over iterations is reported. The size of test set was $15,000$. We used a linear model and initialized gVAMP with noise precision equal to 2 and sparsity rate corresponding to $50,000$ causal markers. Table below indicates the test R^2 in all described scenarios and the number of individuals present in the study.

Trait code	gVAMP	GMRM	#samples
CHOL	0.139831	0.143448	395,025
HDL	0.198088	0.212956	360,286
HT	0.416071	0.421616	399,055
MCH	0.211646	0.220054	402,201
RBC	0.188189	0.198147	402,204
SBP	0.0612896	0.0723477	377,347

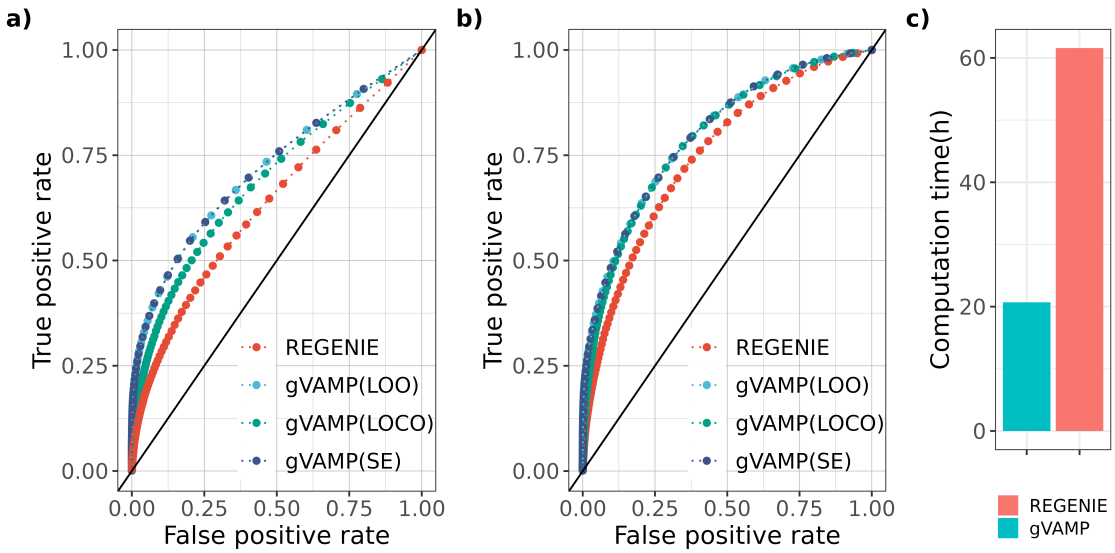

Figure 3. ROC curve comparison between REGENIE and gVAMP. For calculating p-values, REGENIE uses a leave-one-chromosome out (LOCO) approach, gVAMP offers LOCO, leave-one-out (LOO), or a state evolution approach. The left plot considers *exact* causal variant discovery comparison, while in the central one we group genetic markers into correlated sets, using an R^2 threshold of 0.1 and treating each of the $290,000$ sets as either causal or non-causal. The right plot presents a comparison of computation time between gVAMP (35 iterations) and REGENIE (step 1 is performed on a subset of $887,000$ genetic markers). The dataset contains $400,000$ individuals and 1.4 million markers, whose effects are simulated from a spike-and-slab model, where $50,000$ markers exhibit an effect with trait heritability $h^2 = 0.5$.

asymptotically Gaussian due to the Onsager correction [20], i.e., $\boldsymbol{r}_{1t} \approx \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1t}^{-1}\boldsymbol{I})$, where $\boldsymbol{\beta}$ is the ground-truth value of the underlying genetic effects vector.[3] Hence, given a marker index i, a double-sided p-value for the hypothesis test H0 : $\beta_i = 0$ can be calculated as $\Phi(-|r_{1t,i}| \cdot \gamma_{1t}^{1/2})$, where Φ is the CDF of a standard Gaussian.

[3]Specifically, one can show that $\frac{1}{N}\|\boldsymbol{r}_{1t} - \boldsymbol{\beta} - \mathcal{N}(0, \gamma_{1t}^{-1}\boldsymbol{I})\| \to 0$, as $N, P \to \infty$, with the ratio N/P being fixed.

Our approach is the *first* to generate genetic predictors and association test statistics in a single step, without additional computations. In a simulation, we find that the p-value testing approach based on state evolution, as well as the leave-one-out approach based on gVAMP genetic predictors, give better association testing performance compared to the state-of-the-art REGENIE software [5], in only a third of the compute time, see Figure 3.

Conclusion

A line of work spanning over a decade has shown that the AMP framework is highly promising due to its appealing theoretical properties. Our paper demonstrates that this promise can be effectively translated into practical performance for the large-scale application of genome-wide association studies. Our gVAMP approach overcomes current limitations in prediction and association testing for individual-level high-dimensional biobank data, both in terms of computational speed and quality of association testing. We highlight that gVAMP performs *both prediction and association testing at the same time*, unlike state-of-the-art methods which specialize in either of the two tasks.

We are now developing this approach for wide-scale application. Potential future directions include multiple annotation groups, genotype-environment testing, and multiple observable outcomes [23]. AMP facilitates going beyond linear models [24], which would allow for binary outcome probit link function models, as well as time-to-event models. These developments will maximise the potential of large-scale datasets towards the end goal of precision medicine.

Data availability

This project uses the UK Biobank data under project number 35520. UK Biobank genotypic and phenotypic data is available through a formal request at (`http://www.ukbiobank.ac.uk`).

Code availability

The gVAMP code is fully open source `https://github.com/medical-genomics-group/gVAMP`. The scripts used are available at `https://github.com/medical-genomics-group/gVAMP`. R version 4.2.1 is available at `https://www.r-project.org/`. Plink version 1.9 is available at `https://www.cog-genomics.org/plink/1.9/`. Regenie is here `https://github.com/rgcgithub/regenie`.

Acknowledgements

This project was funded by a Lopez-Loreta Prize to MM, by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. We would like to thank Philip Schniter, Matthew Stephens and Pragya Sur for providing valuable suggestions and comments on an early version of the work. We would like to acknowledge the participants and investigators of the UK Biobank study. High-performance computing was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Author contributions

MM and MRR conceived the study. AD, MM and MRR designed the study. AD derived the model and the algorithm, with input from MM and MRR. AD wrote the software and conducted the analysis. AD, MM and MRR wrote the paper. All authors approved the final manuscript prior to submission.

Ethical approval declaration

This project uses UK Biobank data under project 35520. UK Biobank genotypic and phenotypic data is available through a formal request at `http://www.ukbiobank.ac.uk`. The UK Biobank has ethics approval from the North West Multi-centre Research Ethics Committee (MREC). Methods were carried out in accordance with the relevant guidelines and regulations, with informed consent obtained from all participants.

References

1. C. L. B. Zawatsky et al., "Returning actionable genomic results in a research biobank: Analytic validity, clinical implementation, and resource utilization," *The American Journal of Human Genetics*, vol. 108, no. 12, pp. 2224–2237, 2021.

2. F. Aguet et al., "Molecular quantitative trait loci," *Nature Reviews Methods Primers*, vol. 3, no. 1, pp. 4, 2023.

3. L. J. Beesley et al., "The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities," *Statistics in Medicine*, vol. 39, no. 6, pp. 773–800, 2020.

4. C. R. Henderson, "Estimation of variance and covariance components," *Biometrics*, vol. 9, no. 2, pp. 226–252, 1953.

5. J. Mbatchou et al., "Computationally efficient whole-genome regression for quantitative and binary traits," *Nature Genetics*, vol. 53, no. 7, pp. 1097–1103, 2021.

6. P.-R. Loh et al., "Efficient bayesian mixed-model analysis increases association power in large cohorts," *Nature Genetics*, vol. 47, no. 3, pp. 284–290, 2015.

7. E. Orliac, D. Trejo Banos, S. Ojavee, K. Läll, R. Mägi, P. Visscher, and M. Robinson, "Maximizing GWAS discovery and genomic prediction accuracy in biobank data," *Proceedings of the National Academy of Sciences*, vol. 119, no. 31, 2022.

8. J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5451–5460, 2019.

9. A. Montanari and R. Venkataramanan, "Estimation of low-rank matrices via approximate message passing," *Annals of Statistics*, vol. 45, no. 1, pp. 321–345, 2021.

10. D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 18914–18919, 2009.

11. M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Information Theory*, vol. 57, pp. 764–785, 2011.

12. S. Rangan, "Generalized Approximate Message Passing for Estimation with Random Linear Mixing," in *IEEE International Symposium on Information Theory*, 2011.

13. P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14516–14525, 2019.

14. R. Venkataramanan, K. Kögler, and M. Mondelli, "Estimation in rotationally invariant generalized linear models via approximate message passing," in *International Conference on Machine Learning*, 2022, pp. 22120–22144.

15. J. Barbier, F. Camilli, M. Mondelli, and M. Sáenz, "Fundamental limits in structured principal component analysis and how to reach them," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, 2023.

16. C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *IEEE International Symposium on Information Theory*, 2015, pp. 1227–1231.

17. C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Trans. Information Theory*, vol. 62, no. 9, pp. 5117–5144, 2016.

18. E. M. Eksioglu and A. K. Tanc, "Denoising AMP for MRI reconstruction: BM3D-AMP-MRI," *SIAM Journal on Imaging Sciences*, vol. 11, no. 3, pp. 2090–2109, 2018.

19. X. Zhong, C. Su, and Z. Fan, "Empirical bayes pca in high dimensions," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 3, pp. 853–878, 2022.

20. S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

21. A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems," in *Advances in Neural Information Processing Systems*, 2017.

22. J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *International Conference on Acoustics, Speech and Signal Processing*, 2015.

23. P. Pandit, M. Sahraee Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher, "Matrix inference and estimation in multi-layer models," in *Advances in Neural Information Processing Systems*, 2020.

24. P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *Asilomar conference on signals, systems and computers*, 2016.